Organizational Data Visualization

Communicating clearly, concisely, and convincingly

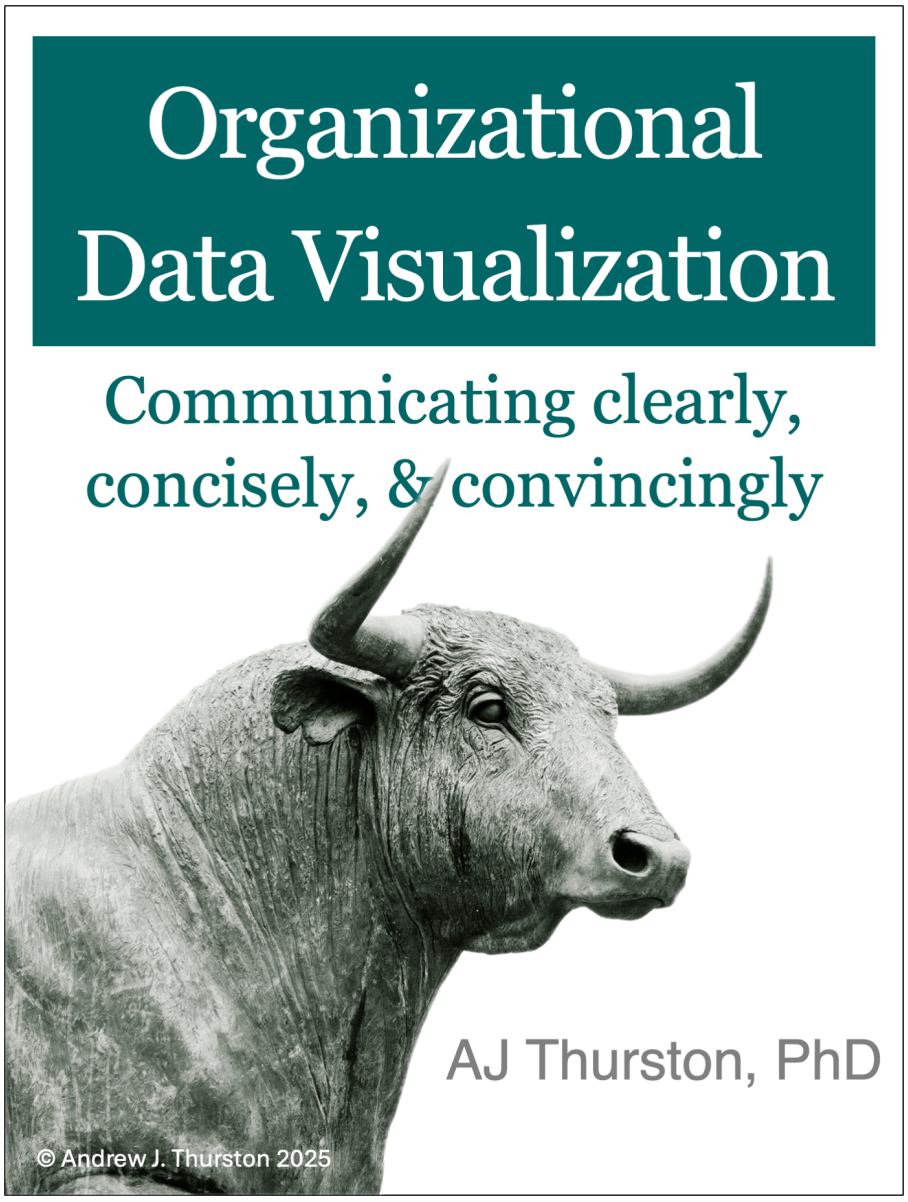
AJ Thurston

2025-01-01

Table of contents

Preface

Welcome to Dr. AJ Thurston's course on Organizational Data Visualization (ODV) as a part of the George Mason University Industrial and Organizational Psychology program. This course is focused on students in our Masters of Professional Studies (MPS) in I-O program, specifically, and is part of the newly established interdisciplinary certificate in data science. ODV is designed to train students to communicating quantitative information clearly, concisely, and convincingly to organizational stakeholders. Visual communication is increasingly important in the workplace, making ODV an increasingly valuable skill for both organizations and individuals.


Organizational
Data Visualization
Communicating clearly,
concisely, & convincingly
AJ Thurston, PhD
© Andrew J. Thurston 2025

Prerequisites

This course requires a graduate level understanding of psychological statistics and a basic understanding of working with R and the {ggplot2} R package in particular. Specifically, students should have already completed PSYC 601: Applied Data Analysis in Psychology 1 or expressed written approval from the instructor in advance of the course.

Students are encouraged to have already installed R, R Studio, Positron, Quarto, and Github. This course from Microsoft is recommended to build basic familiarity with R, the {tidyverse} and {ggplot2} R packages:

- Microsoft: Explore and analyze data with R (1 hour)

Students are encouraged to maintain their code on Github and to familiarize themselves with version control (e.g., git).

Software

Download and install the following software:

R	• Open source, statistical computing & graphics environment • Large package (tool) library • The engine of our data visualization
R Studio	• Integrated Development Environment (IDE) for writing and running R code
Positron (Beta)	• IDE for R code but flexible for other languages as well • Beta software with full release Fall 2025

Quarto	• Notes code an output in one R Markdown (.qmd) document • Uses Markdown language to create files exportable as HTML, PDF & Word docs
Github Desktop	• Popular git-based version control system • Desktop version uses a graphic user interface (GUI) to simplify version control

Textbook

The following textbook is required for the course:

Yau, N. (2024). *Visualize This: The flowingdata guide to design, visualization, and statistics second edition.* John Wiley and Sons.

What is Org Dataviz?

Data visualization is the communication of numeric information using graphic displays a skill, much like written & oral communication skills. Data visualization can also be art, designed to evoke a feeling or thought. Data visualization typically has two main functions: exploratory visualizations which reveal new patterns & relationships; explanatory visualizations evaluate hypotheses or aid decision-making. Not all data visualization is ODV.

Organizational data visualization is the production of visual displays of quantitative data to support organizational decision-making which aims to describe, explain, predict, and change human behavior at work. ODV is not just data visualization as in related technical fields like biology or operational business components like logistics since these focus on non-human behavior (e.g., animals in biology) or behaviors of non-human systems (e.g., supply chain bottlenecks in logistics). ODV is distinct from data art (e.g., reddit.com/r/dataisbeautiful) in that the primary purpose of data art is to entertain or for purely aesthetic purposes ODV is ruthlessly useful and simple and should make decisions obvious to stakeholders.

1 Introduction

1.1 A multidisplinary approach

The SIOP Education and Training (E&T) guidelines stipulate students of I-O psychology should be proficient in written and oral communication skills. Visual communication is becoming increasingly important in the workplace, and many important decisions rest on one simple visual tool for communicing quantiative information. Data visualization skills can make you a more competitive job applicant.

ODV requires a multidisplinary approach. First, borrowing from computer science, this course uses the R programming language as it's focused on statistics and visualization. You should already have familiarity with this software and this course aims to enhance it to allow your work to be reproducible. Second, ODV borrows from graphic design. Design thinking allows us to empathize with our audience to ensure they understand the data presented to then which can aid their decision-making. This course will require you to capitilize on your knowledge of I-O psychology and psychological statistics and enhance them with programming and design skills.

1.2 Data, Information, Knowledge, and Wisdom

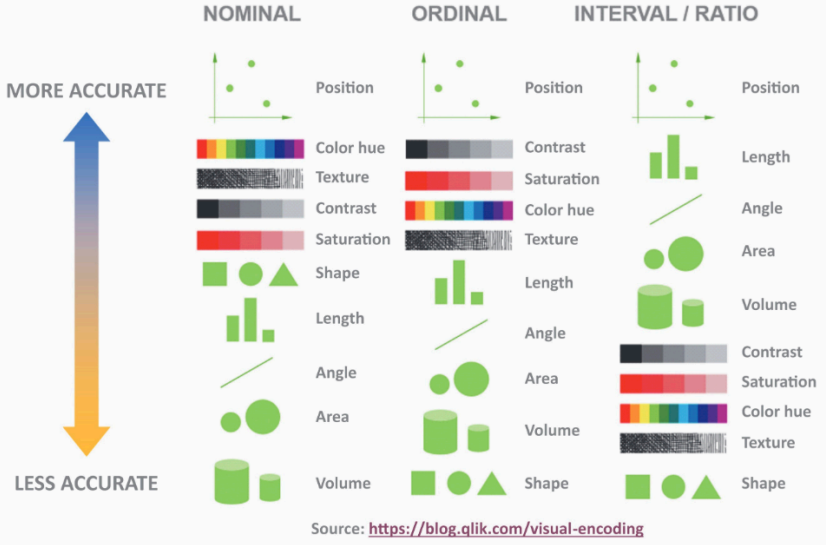
- The cognitive psychology of visual information processing
- Visual communication best practices

- Review of data types (e.g., nominal, ordinal, interval, ratio data)
- Data, information, knowledge, and wisdom pyramid
- The facts DON'T speak for themselves
- Keep it simple, stupid
- AGILE: Think like a software developer
  - BHAG: Big Hairy Audacious Goals
  - KPIs: Key Performance Indicator
  - OKRs: Objective Key Results
- The goal of this course is to give you a data visualization portfolio to make you competitive
- Tell 'em what you're going to tell them. Tell them. Then tell 'em what you just told them.
- TED: The beauty of data visualization - David McCandless (18 minutes)
  - "Let the dataset change your mindset"
- Canva: How design thinking is used to solve problemsArchive

1.3 Visual Hierarchy

```
-   Visual hierarchy refers to the arrangement of design elements in a way th
-   Key principles include:
    -   **Size**: Larger elements draw more attention.
    -   **Contrast**: Differences in color or tone highlight important areas
    -   **Alignment**: Consistent positioning helps organize information.
    -   **Proximity**: Elements placed closely together are perceived as rela
    -   **Whitespace**: Areas left empty create focus by reducing visual clut
-   This design approach helps prioritize content, making it easier for users
-   Resources:
    -   [Canva: Visual Hierarchy](https://www.canva.com/learn/visual-hierarch
```

Figure 3: Visual Features by Data Type, Ranked by Interpretative Accuracy



- Visual Encoding (Sinar, 2018)

1.4 Chartjunk

- Chartjunk refers to unnecessary or distracting visual elements in charts and graphs that do not contribute to the understanding of the data. These elements, such as excessive grid lines, unnecessary text, or 3D effects, clutter visual representations and obscure the data's message. While some argue that chartjunk can make charts more memorable, others believe it detracts from clarity and precision.
- Darkhorse analytics: data looks better naked, data minimalism: https://x.com/dhanalytics/status/1519650227329798144
- Although design is an important part of data visualization! It can go too far the other way as well.

- R is often critiqued for the necessity of supplementation (e.g., annotation) but this is not a technical limitation; rather, R users tend to underinvest in design at higher rates compared to Illustrator or D3.js users: https://medium.com/nightingale/you-can-design-a-good-chart-with-r-5d00ed7dd18e

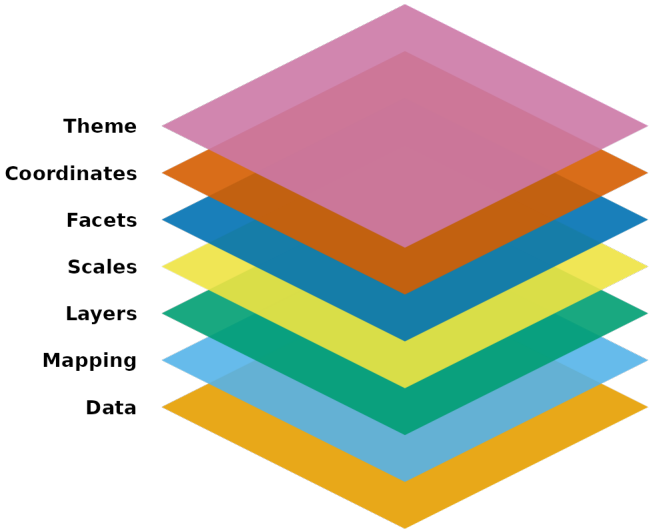
1.5 Audience

- A really simply difference is that in technical audiences I use words like error to mean uncertainty, for non-technical audiences I do not.
- Understanding the audience is key to data visualization.
- The data don't speak for themselves and advocates for the data often don't represent it well.
- The choir does need to be preached to.
- Many decision-makers try to outsource their decision-making to data.
- Decision-makers & advocates have more in common with each other than either of them have in common with the data.
- Subjective data is just as if not more important than the objective data, but objective data can conceal subjectivity as well; for example, even something as objective as sales must account for how difficult sales are in one region over another.
- Facts tell and stories sell, people prefer anecdotes.
- Even though knowledge is half the battle, the other half is yet won.

1.6 Tools

- Orientation to R, Rstudio, Posistron IDE, Quarto, and ggplot2
- Using Git and version control and GitHub
- R is just a very sophisticated calculator and it will do precisely what you tell it to do

- Data storytelling: https://m.youtube.com/watch?v=IfKlOC5HSHI
- How to create a Quarto document and export to PDF
    - Quarto execution opts: https://quarto.org/docs/computations/execution-options.html
    - Reprex package R https://reprex.tidyverse.org/articles/learn-reprex.html
    - Why individual level data are important: https://x.com/sc_cath/status/1799797169764844000
- R :: packages
    - I will always use the curly bracket for {packages}
- R :: ggplot2
    - The R programming language uses {ggplot2} to make great data visualizations.
    - https://ggplot2.tidyverse.org/articles/ggplot2.html
    - https://ggplot2.tidyverse.org/reference/index.html
    - Grammar of graphics
    - Uses layers to build



1.7 Tables

- Tables are not data visualization
- https://jthomasmock.github.io/gtExtras/articles/plotting-with-gtExtras.html
- Interactive tables in Quarto: https://holtzy.github.io/quarto-tricks/interactive-table/
- Reactable (Quarto integration limited): https://glin.github.io/reactable/articles/exam
- DT is a great package, does work with Quarto
- {gt} and {gtExtra} is for pretty tables (example in ajthurston.com/knolling)

1.8 Color

1.8.1 Color Perception

- Color perception: black face vs white face illusion: https://m.youtube.com/watch?v=zXU0O0OqkmQ
- Color is an important part of data visualization because it helps to highlight the most important differences or trends, making it easy to see. It's one of the fundamental ways to highlight data.
- Color is also an important part of organizational or individual branding. Thinking about the ubiquity of Coca-Cola red or Pepsi blue.
- Color perception from evolutionary psychology foundation and then cultural enforcement which drives differences

1.8.2 Color Theory

- Color Wheel: The primary colors are red, green, and blue because that's what our eyes can see. Red, blue, yellow is additive color theory from paint pigment days and not really representative of how color works.
- Color Harmony
  - Color harmony involves choosing colors that work well together and are pleasing to the eye. Common color harmony schemes include:
  - Complementary Colors: Colors opposite each other on the color wheel (e.g., red and green) that create high contrast and vibrant looks.
  - Analogous Colors: Colors next to each other on the color wheel (e.g., blue, blue-green, and green) that create serene and comfortable designs.

  - Triadic Colors: Three colors evenly spaced on the color wheel (e.g., red, yellow, and blue) that offer balanced and vibrant combinations.
  - Split-Complementary Colors: A base color and the two colors adjacent to its complementary color, offering high contrast but less tension than complementary colors.

- Color Context: The perception of a color can change depending on its context. For instance, a color may appear different against various background colors. Understanding how colors interact and influence each other helps in creating effective visualizations.

1.8.3 Color Psychology

(I need to look into the research on this more before I teach it) Colors can evoke emotions and associations: - Red: Energy, urgency, passion. - Blue: Trust, calm, professionalism. - Green: Growth, health, tranquility. - Yellow: Happiness, attention, caution. - Purple: Luxury, creativity, wisdom. - Orange: Enthusiasm, warmth, caution. - Color Associations - Red - Negative Values: Red is often used to indicate negative values, losses, or decreases. - Errors and Alerts: Red can signify errors, warnings, or urgent issues that need immediate attention. (there's research on this, the red pen effect) - Critical States: In status indicators, red often represents critical conditions or failures. - Green - Positive Values: Green is commonly associated with positive values, gains, or increases. - Success and Approval: Green can indicate success, completion, or approval. - Safe and Normal States: In status indicators, green represents safe, normal, or optimal conditions. - Yellow - Caution: Yellow often signifies caution, warnings, or conditions that require attention but are not yet critical. - Intermediate Values: Yellow can be used for intermediate values between positive (green) and negative (red) extremes. - Blue - Neutral: Blue is often used as a neutral color to represent general information, averages, or non-critical data. - Information and Support: Blue can indicate informa-

tional messages or support functions. - Orange - Warning: Orange is used to indicate warnings or issues that are of moderate concern. - Pending Actions: It can also represent pending actions or items that are in progress. - Color Usage: Using color effectively in data visualization helps in communicating information clearly: - Highlighting Important Data: Use bold or contrasting colors to emphasize key data points. - Grouping Data: Use similar colors to group related data, making patterns and relationships easier to identify. - Sequential Data: Use gradient scales (light to dark) to represent sequential data, indicating progression or intensity. - Diverging Data: Use two contrasting colors to show deviation from a central point, often used in heat maps or to show differences.

2 Visualizing Time

2.1 Categorical Events

2.1.1 Before and After

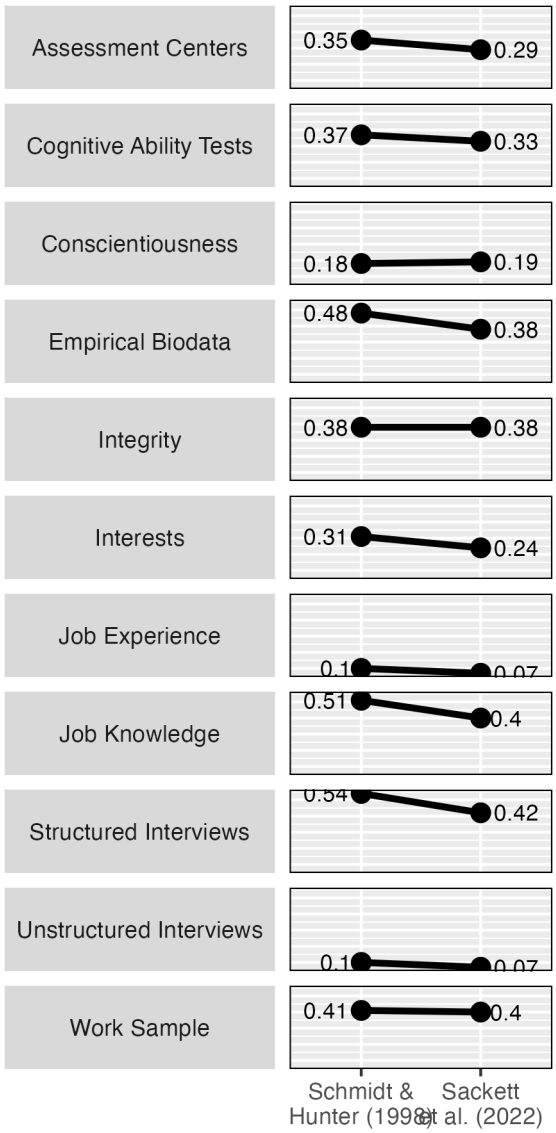
```
library(ggplot2)

s2022 <- read.csv(here::here("data/sackett2022.csv"), header = T)

ggplot(s2022, aes(x = Time_Point, y = Correlation, group = Predictor)) +
  geom_line() +
  geom_point() +
  facet_wrap(~ Predictor, scales = "fixed", ncol = 1, strip.position = "left") +
  labs(title = NULL,
       x = NULL,
       y = "Correlation") +
  theme_minimal() +
    theme(
    strip.text.y.left = element_text(angle = 0), # Forces horizontal text
    strip.placement = "outside",                 # Moves strips outside the plot
    legend.position = "none"
  )
```

```
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyr)

# Load your CSV file
s2022 <- read.csv(here::here("data/sackett2022.csv"), header = TRUE)

# Ensure the Time_Point factor levels are correctly ordered
s2022 <- s2022 %>%
  mutate(Time_Point = factor(Time_Point,
                             levels = c("Schmidt &\nHunter (1998)", "Sackett

# Create the plot
plot <- ggplot(s2022, aes(x = Time_Point, y = Correlation, group = Predictor)
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  geom_text(
    aes(label = Correlation,
        hjust = ifelse(Time_Point == "Sackett\net al. (2022)", -0.3, 1.3)),
    size = 3
```

```
  ) +
  facet_wrap(~ Predictor, ncol = 1, strip.position = "left") +
  scale_x_discrete() + # Maintain default order
  labs(title = NULL, x = NULL, y = NULL) + # Remove title and axis labels
  theme(
    strip.text.y.left = element_text(angle = 0), # Horizontal strip text
    strip.placement = "outside",                 # Move strips outside
    axis.text.y = element_blank(),               # Remove y-axis text
    axis.ticks.y = element_blank(),              # Remove y-axis ticks
    legend.position = "none",                    # Remove legend
    panel.border = element_rect(color = "black", fill = NA, linewidth = 0.5) # Add facet out
  )

ggsave(
  filename = here::here("images/sackett2022.png"), # File path
  plot = plot,
  width = 3,      # Width in inches
  height = 6,    # Height in inches (7 times the width)
  dpi = 300       # High resolution for websites
)
```


Assessment Centers
0.35
0.29
Cognitive Ability Tests
0.37
0.33
Conscientiousness
0.18
0.19
Empirical Biodata
0.48
0.38
Integrity
0.38
0.38
Interests
0.31
0.24
Job Experience
Job Knowledge
0.51
0.4
Structured Interviews
0.42
Unstructured Interviews
Work Sample
0.41
0.4
Schmidt & Hunter (1998)
Sackett et al. (2022)

2.2 Continuous Events

2.2.1 Line charts

2.2.2 Area charts

- Changes over time as ratios: Month over month (MoM), year over year (YoY), etc.
  - Key Performance Indicators (KPIs)
- Representing cycles and seasonality (seasonal plots, decomposition plots).

2.2.3 Survival Analysis

2.2.4 Communicating Uncertainty

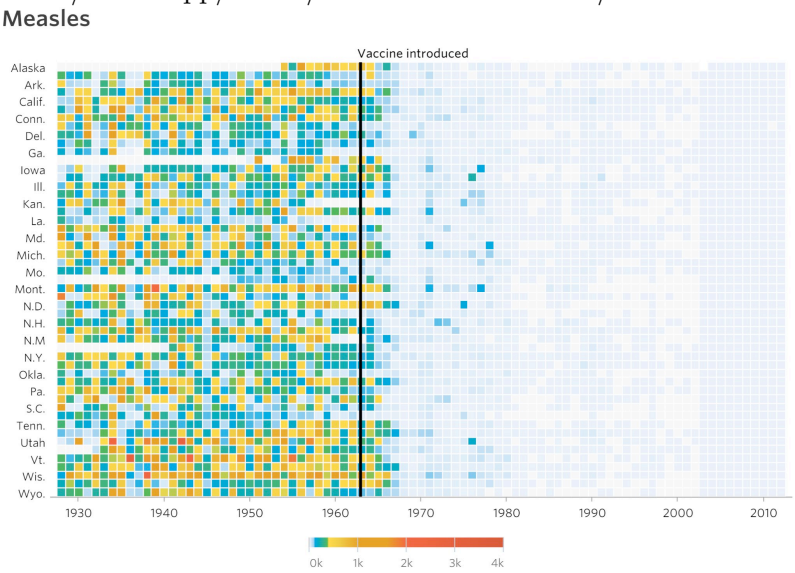
```
-   538 has a cool way of showing uncertainty "95% of people fall in this band"
-   ![](/images/fivethirtyeight_confidence_interval.png.png)
```

2.3 Categorical Events

2.4 Waffleplot

- Visualization of events
- I recommend for categorical only actually
  - https://vuorre.com/posts/2016-03-24-github-waffle-plot/
  - Cols: months, rows: days
  - Memento mori plot with 80 years of months

– https://x.com/AlecStapp/status/1826341717467115743/



photo/1

* This one is "bad" because I don't need it broken out by state if the cases do not meaningfully differ by state
* I also prefer to use line charts for continuous data on the y-axis
* If necessary I would do line charts as facets for each state
* Sinar (2018) suggested length and angle more accurate than color contrast, saturation, or hue in interpretative accuracy

2.4.1 Gantt Charts

2.5 Rank Order Changes

2.5.1 Alluvial Diagrams

- Alluvial diagrams

- Sankey diagrams

2.5.2 Bump Plot

- Bump plot for rank order changes over time
  - r-graph-gallery.com/package/ggbump.html
  - Sackett et al. 2022 meta-analysis of predictors: ![]sackett2022.png

2.6 Forecasting

- There was a SIOP 2024 talk I attended on this?
- Financial Times: Never ever make predictions Archive

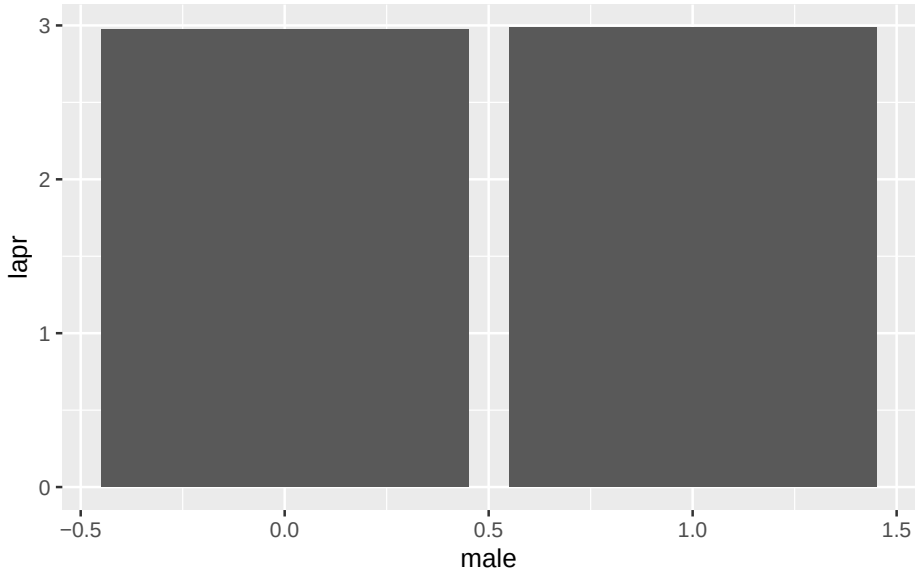
- Visualization of seasonality using births, but make something like this but for accidents at work to demonstrate seasonality
- Awesome strategies for visualizing changes over time
- https://x.com/aureliusltd28/status/1869416785684127783

3 Visualizing Categories

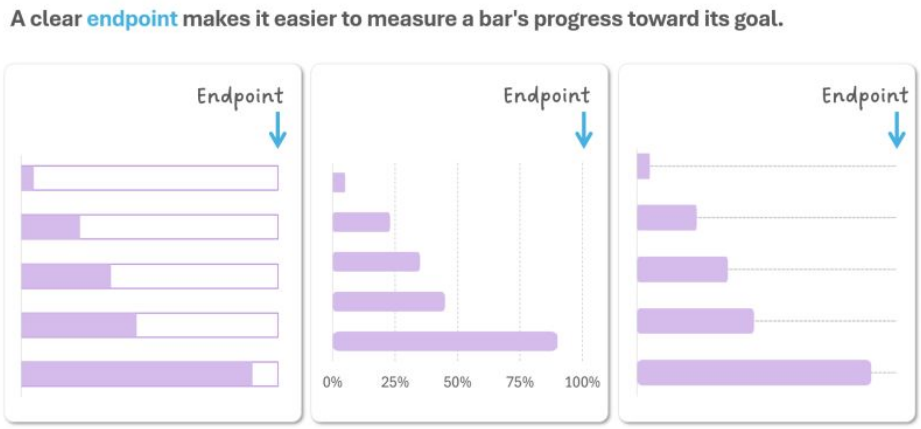
3.1 Frequencies

3.1.1 Barplots

```
library(ggplot2)
dat <- read.csv(here::here("data/validation.csv"), header = T)

ggplot(data = dat, aes(x = male, y = lapr)) +
    geom_bar(position = "dodge",
             stat = "summary",
             fun = "mean")
```

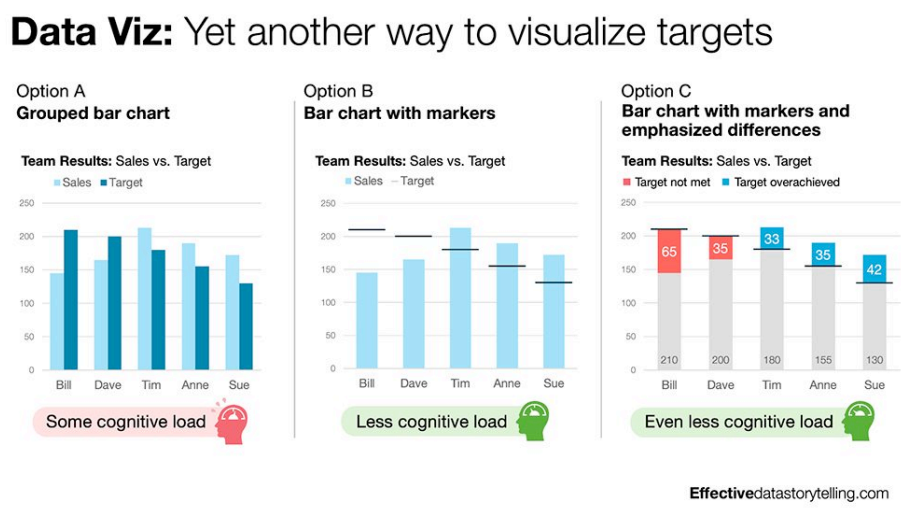

3
2
1
0
lapr
−0.5
0.0
0.5
1.0
1.5
male

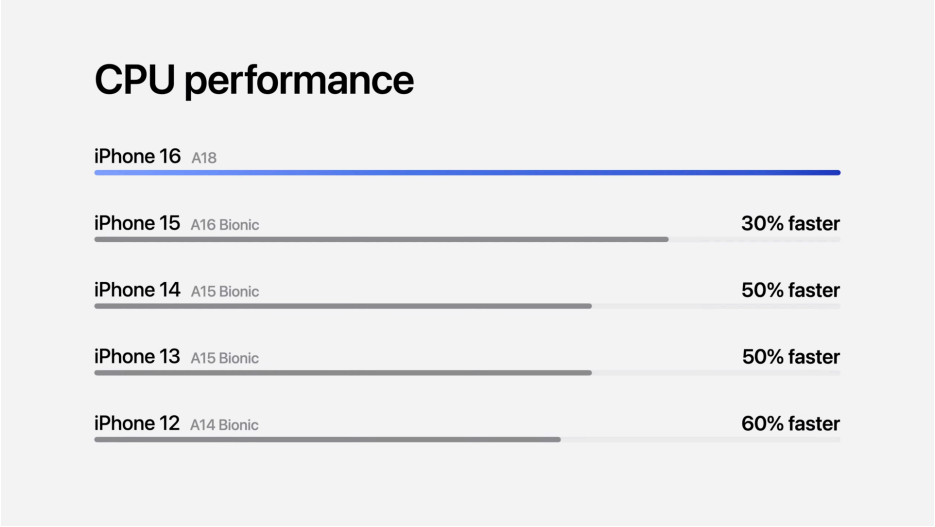
3.1.2 Endpoints



Credit: Salma Sultana (LinkedIn)

3.1.3 Targets


Data Viz: Yet another way to visualize targets
Option A
Grouped bar chart
Team Results: Sales vs. Target
Sales
Target
250
200
150
100
50
0
Bill
Dave
Tim
Anne
Sue
Some cognitive load
Option B
Bar chart with markers
Team Results: Sales vs. Target
Sales
Target
250
200
150
100
50
0
Bill
Dave
Tim
Anne
Sue
Less cognitive load
Option C
Bar chart with markers and emphasized differences
Team Results: Sales vs. Target
Target not met
Target overachieved
250
200
150
100
50
0
65
35
33
35
42
210
200
180
155
130
Bill
Dave
Tim
Anne
Sue
Even less cognitive load
Effectivedatastorytelling.com


CPU performance
iPhone 16 A18
iPhone 15 A16 Bionic
30% faster
iPhone 14 A15 Bionic
50% faster
iPhone 13 A15 Bionic
50% faster
iPhone 12 A14 Bionic
60% faster

3.1.4 Density Plot

- Demonstrating amounts (e.g., frequencies, distributions bar charts)
  - Bar charts, loading bar chart as whitespace is information as well
  - Communicating the lack of results is just as important as communicating the presence of results. Audiences are predispositioned to want to look for differences, but you must help them understand the MANIFOLD of differences, the larger context, that while there may be 2 differences among 10, the other 8 are not. You'd be surprised how often you need to state this explicitly.

3.1.5 Icon array

Icon arrays are useful alternatives to bar plots because they are empirically better understood by non-technical audiences (Zhang et al. 2018).

3.1.6 Jitterplot

3.2 Descriptives Statistics

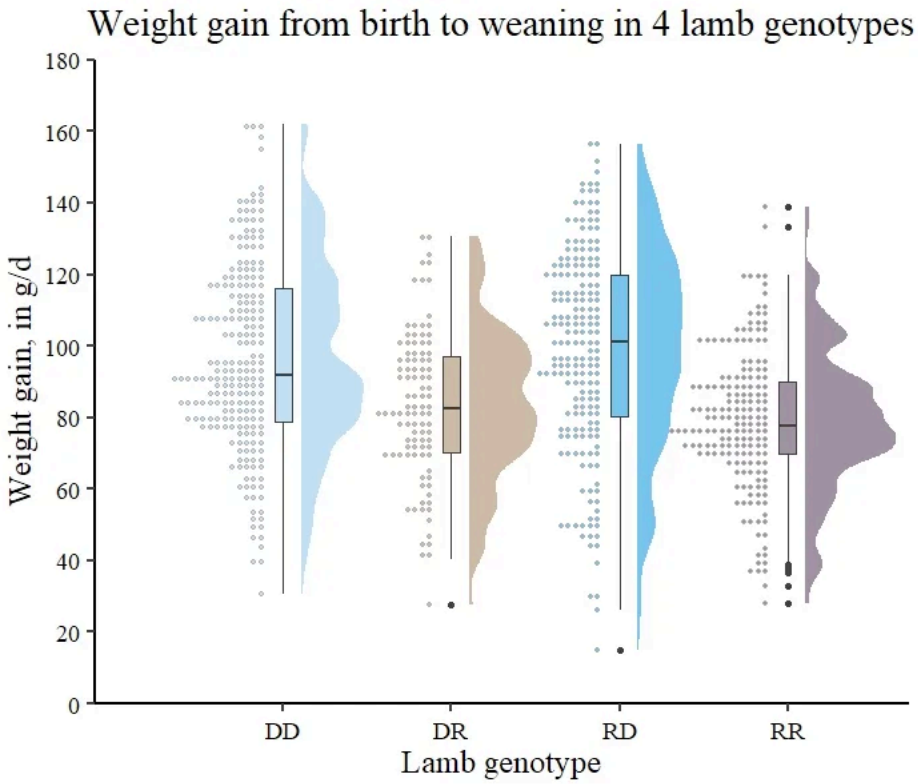
3.2.1 Barplot

3.2.2 Box and Whisker Plot

- Visualization of descriptive statistics and distributions
- Visualization of IQRs with density and quantiles overlaid

3.2.3 Raincloud plot

- Raincloud plots: https://medium.com/(**amorimfranchi/raincloud-plots-for-clear-precise-and-efficient-data-communication-4c71d0a37c23?**)



3.3 Parts of a Whole

3.3.1 Pie Charts

Do not make pie charts. They are very common but difficult to resolve as our brains struggle with how to resolve circles and are not adept at determining length from circumference.

Watch this video to see just how bad our brains are at this:

https://www.youtube-nocookie.com/embed/4kxI4EjL3rg

Figure 3.1: Which is longer? Credit: Modern Rogue (YouTube: 1 min)

If you're asked to make a pie chart, please show this video to your boss or whoever asked you to make the pie chart. You, too, can join the good fight against pie charts.

Another issue is that pie charts are essentially area charts.

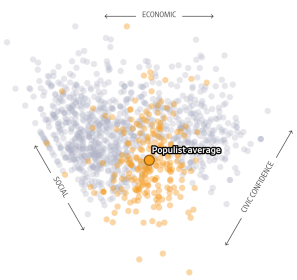
3.3.2 Knolling

3.3.3 Mosiac Plots

- Or parts of a whole

We don't do pie charts because of this:

-
-
  - Our brains are not built to process just how long things are around
- Tricolor/ternary/triangle plot: WSJ: What type of voter are you?



The conventional idea of right and left has given way to new divisions. An emerging group of populists has moved to the fore, in part because of the ways former President Donald Trump has reshaped traditional Republican orthodoxy. Populists are socially conservative but hold some traditionally liberal economic views, such as supporting Social Security and other government social programs.

– Instagram link

* https://github.com/jschoeley/tricolore

3.4 Appendix

https://www.youtube-nocookie.com/embed/wP8NWRR0Fdg

Figure 3.2: Make Beautiful Excel Charts Like The Economist Credit: Leila Gharani (YouTube: 19 minutes)

4 Visualizing Relationships: Describing Relationships

4.1 Flowcharts

- Mermaid graphs: https://towardsdatascience.com/introduction-to-mermaid-graphs-in-markdown-080d4377cab4

4.2 Group Differences

- Effect sizes here?
  - Cohen's D

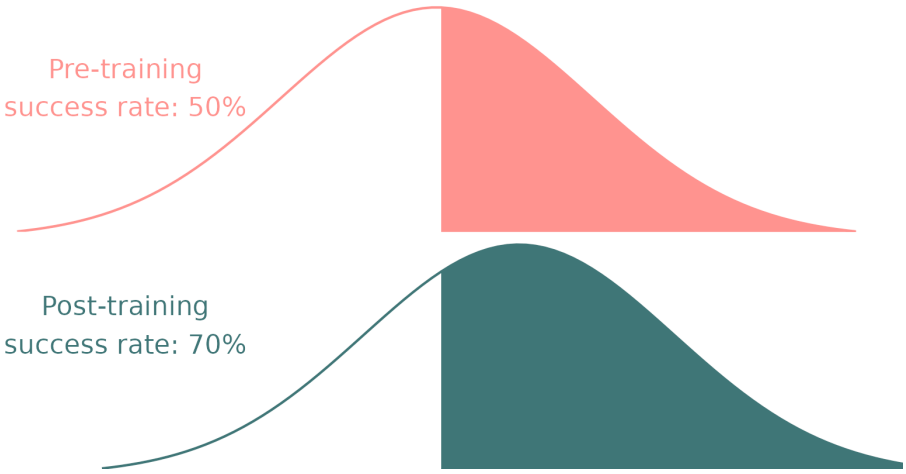
4.2.1 Dumbbell Plots

- Dumbbell plot ajthurston.com/dumbbell

4.2.2 Knolling

```
- ajthurston.com/knolling
- Alternative in Excel: Tables with distributions and barplots
- Simplified version in Excel: [Don't use regular bar charts! #excel #exceltutorial #excelti
- In R using gt_plt_dist() and gt_bar_plot(), examples here: <https://themockup.blog/posts/2
```

- Before and after with Cohen's U3 (ajthurston.com/u3) Inspired by Alliger (2024):



- Effect sizes (e.g., funnel and forest plots, alternative effect size displays)
  - Cohen's d
  - Common Language Effect Sizes (CLES)
    - Binomial effect size display
    - Cohen's U3
  - Elasticities

4.3 Correlation

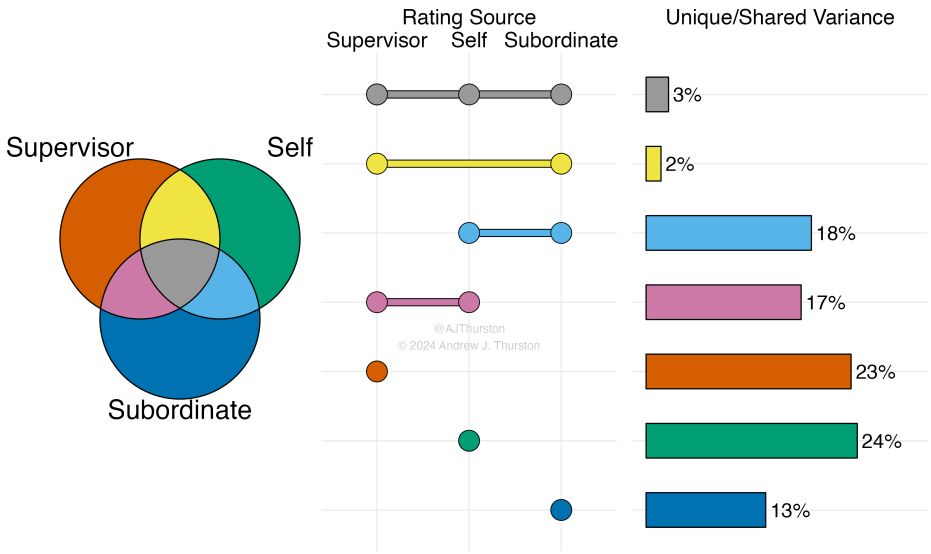
- Simple correlation

4.3.1 Scatterplots

4.3.2 Correlelogram

4.3.3 Ternary plots

4.3.4 Upset plots



Venn diagrams and Ballentine plots.

4.4 Simple linear regression

4.4.1 Expectancy charts (ajthurston.com/expectancy)

4.4.2 Relationship heatmap

- <https://x.com/clementychow/status/1831676878471807318>

- `\![\[curseword_heatmap.png \| 500\]\]`

  - Guttman chart, individuals vs. items

5 Visualizing Relationships: Explaining Relationships

5.0.1 Generalized Linear Modeling

- Logistic regression
  - Predicted probabilities ajthurston.com/predprobs
    - http://a.web.umkc.edu/andersonbri/InterpretinglogisticregressionPartII.html
    - https://info.umkc.edu/drbanderson/interpreting-logistic-regression-part-i/
    - http://a.web.umkc.edu/andersonbri/Interpretinglogisticregression.html
    - library.virginia.edu/data/articles/simulating-a-logistic-regression-model
    - https://stats.oarc.ucla.edu/r/dae/logit-regression/
  - Classification accuracy ajthurston.com/accuracy
- Multiple correlation
- Multiple regression
- Relative importance/dominance analysis/feature selection
- Connections (e.g., network analysis)

- Polynomial Regression and Response Surface Analysis (PRRSA) (Edwards & Parry, 1993)
  ![[response_surface_plots.png | 500]] ### Machine Learning
- Visualization of decision support techniques (e.g., CART/random forests)
- stats.stackexchange.com/questions/467697/visualizing-regression-with-categorical-variables ### Networks
- TED: A visual history of human knowledge - Manuel Lima (13 minutes)
  - Don't do radial convergence
  - Don't do arc diagrams

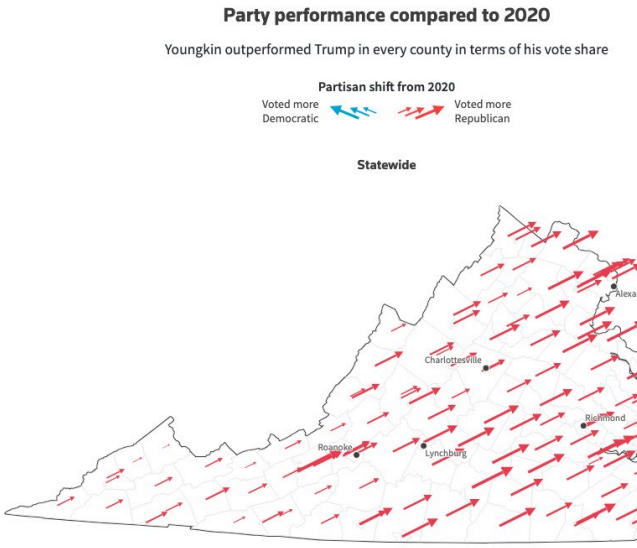
6 Visualizing Space

6.0.1 Maps

- YouTube: RealLifeLore: The Coastline Paradox Explained (6 minutes)
- Making data-based chloropleth and cartogram maps
- Visualization of zip code, county, state, and country data

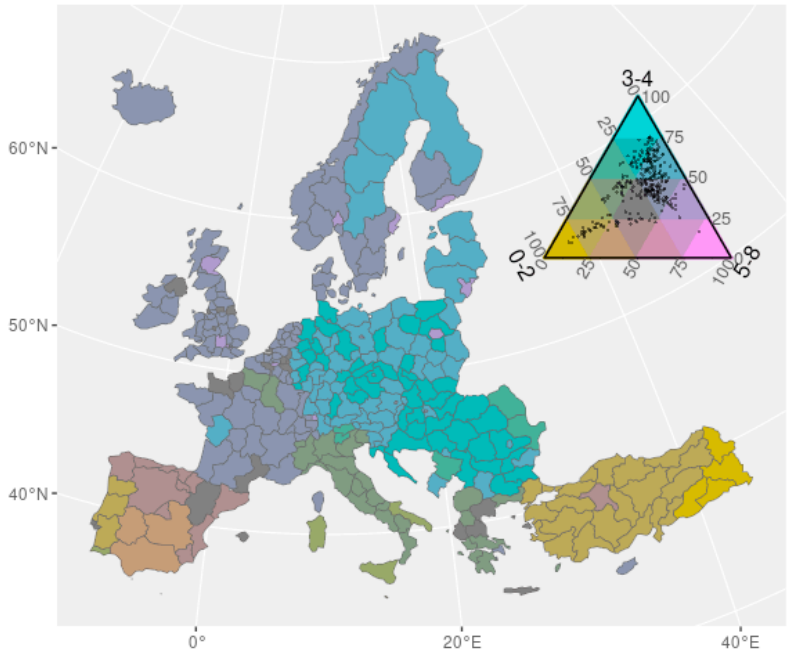
    - I need to teach them about that principal of how you can infinitely divide a coastline.

- Avoiding other mapping techniques and why
- Comparison to shapefiles and GIS data
- Maps with Leafdown
- https://cran.r-project.org/web/packages/tricolore/vignettes/choropleth_maps_with_tricolore.html



- Virginia 2021 Gubenatorial Shift

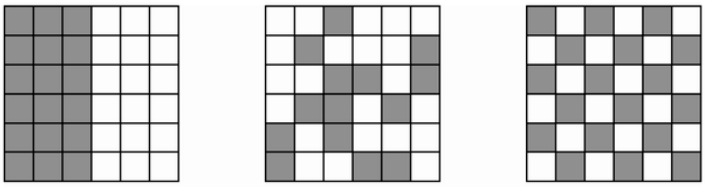
– https://www.reuters.com/graphics/USA-ELECTION/VIRGINIA/gkplgdmzavb/



  - https://github.com/jschoeley/tricolore

- Spatial autocorrelation

  - https://x.com/yohaniddawela/status/1868992061568917669



http://geostatsguy.github.io/MachineLearningDemos_Book/intro.html

7 Interactive Visualization

- Exposure to PowerBI, datawrappers, and advanced programming (e.g., D3.js)
- Hosting interactive tools (e.g., Github Pages, Netlify)
- Serverless apps with shiny live: https://youtu.be/RcwvG7dtMqU
- Automation of Quarto and Github pages: https://www.youtube.com/watch?v=arzBRW5XIkg
- Learn D3.js: Data Visualization using r2D3: R Interface to D3 Visualizations
- [An introduction to d3.js in 10 basic examples](https://d3-graph-gallery.com/intro_d3js.html
- I don't know if D3.js is all that worth it for a little bit of interactivity/animation

7.0.1 Dashboards

- Interactive dashboards/minimum via products using R Shiny
- Mastering Dashboard Design: From Good to Unmissable Data Visualizations Archive

7.0.2 Interactive

- Brown University - Seeing Theory
- Interactive maps with Leafdown

- Scrollytelling: https://georgios.quarto.pub/a-visual-journey-through-world-exhibitions/
- Scrollytelling tutorial: https://x.com/RosanaFerrero/status/1827224441312444629
- Scrollytelling R package: https://closeread.dev

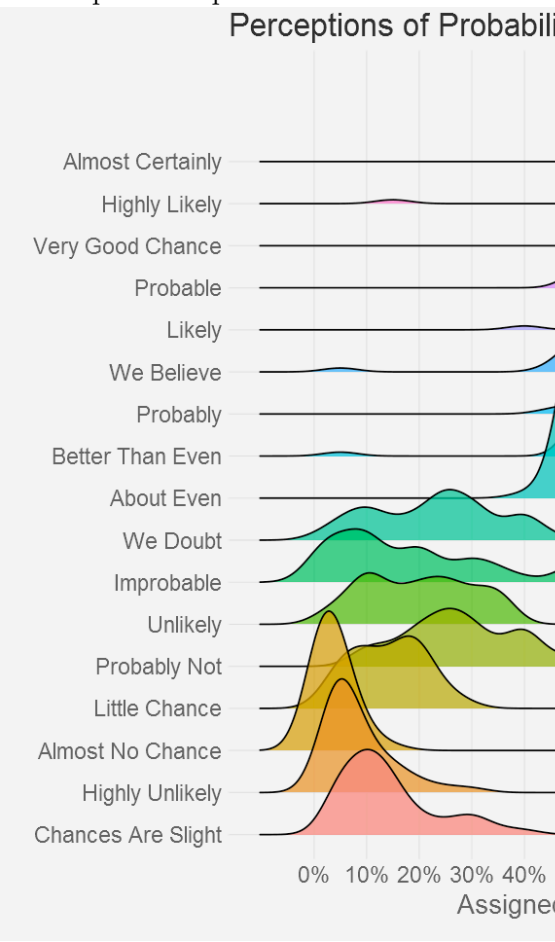
7.0.3 Animations

- Don't have animations playing in the background of your talks as our eyes are drawn to movement
- Using gganimate for animated data visualizations
- gganimate does not play nice with expression()
- Scrollytelling (closeread.dev) is an animation of sorts
- TED: The best stats you've ever seen - Hans Rolling (20 minutes)
  - "Data is often better than you think"
  - "The differences are much bigger than the weakness of the data"
  - "If you look at the average data of the countries – they are like this. That's dangerous to use average data, because there is such a lot of difference within countries."
  - "The improvement of the world must be highly contextualized"

8 Contextual Considerations

8.1 Words Matter(?)

The language used to describe the data matters - Perceptions of probability: https://github.com/zonination/perceptions



- How to communicate statistical concepts without sounding like a robot: https://www.linkedin.com/posts/morgandepenbusch_data-analyst-cheat-sheet-ugcPost-7275556209892446210-6Ilu/

8.2 Accessibility

- Accessibility: Colorblindness, ADA, Section 508 compliance

Here's a bunch of links on accessibility I got from Harvard: https://accessibility.huit.harvard.edu/data-viz-charts-graphs - Technique: Describing graphs - 10 Guidelines for DataViz Accessibility – Highcharts - An intro to designing accessible data visualizations by Sarah L. Fossheim - 5 easy ways to make your data visualization more accessible - Accessibility Considerations in Data Visualization Design - How to make charts and graphs more accessible – Pope Tech

8.3 Cognitive load

- Stephanie Evergreen: 10 Ways to Make Your Graphs More Accessible
- Annotations and best practices for data visualization workflows
- Ethical, legal, and cultural considerations in data visualization
- How to make figures and presentations friendly to colorblind people
- Obake Ito Color Palette
- Okabe-Ito (2008) color palette vs. viridis, I like the former for categorical/nominal data vs. viridus which is good for interval or ratio data, a lot of good detail here: https://stackoverflow.com/questions/57153428/r-plot-color-combinations-that-are-colorblind-accessible

- Alt-text - User centered design/human factors and gathering feedback
- Privacy and data security considerations in data visualization
- Data visualization isn't just about you visualizing your data but understanding and evaluating other people's visual presentation of data (or sniffing out bad dataviz bullshit)

8.3.1 Hosting

- Figshare: https://figshare.com/
- Protocols.io: https://www.protocols.io/
- VizHub is a platform for creating, sharing, and collaborating on interactive data visualizations https://vizhub.com/

8.3.2 Misleading Dataviz

- Misleading graphs
  - Y-axis manipulation, also known as truncation abuse
  - Example misleading graph: https://www.forbes.com/sites/trevorbutterworth/2014/10/02/when-data-journalism-goes-wrong/
  - https://www.linkedin.com/feed/update/urn:li:activity:7242897321909190656/
- From here: x.com/mikebeckhamsm/status/1818798532544917824
- Licensure (cc0)
- Keeping your code clean with Lintr and Stylr (Nicole Rennie video may have inspired this)
  - Tidyverse style guide: https://style.tidyverse.org/index.html
- Hosting
- Data Management

  - Data management slides
  - Levels of documentation: https://x.com/cghlewis/status/1826269873875345875
  - Codebook comparisons - Crystal Lewis

8.3.3 Bad Dataviz

- Long list of bad examples: https://x.com/mpfix1/status/1828109846912372940
- Another long list of bad examples: https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/
- How charts can inadvertently mislead: https://www.asimov.press/p/charts
- linkedin.com/posts/gregdata_data-dataliteracy-activity-7228710352056659968-Rl9e
- https://x.com/chendabo/status/1799614807878095359
- x.com/cremieuxrecueil/status/1825934608137072826
- https://www.businessinsider.com/paul-krugman-inflation-chart-food-energy-gas-housing-prices-economy-2023-10

Resources

Books

- Applied Machine Learning in Python: a Hands-on Guide with Code
- Big Book of R
- Graphic methods for presenting facts - Willard C. Brinton
- Fundamentals of Data Visualization - Claus O. Wilke
- How Charts Lie - Alberto Cairo
- The Visual Display of Quantitative Information - Edward R. Tufte
- Envisioning Information - Edward R. Tufte
- Data Visualization: A Practical Introduction - Kieran Healy
- Factfulness - Hans Rosling
- The Truthful Art: Data, Charts, and Maps for Communication - Alberto Cairo
- Calling Bullshit: The Art of Skepticism in a Data-Driven World - Carl T. Bergstrom and Jevin D. West
- How to Lie with Statistics - Darrell Huff
  - Census: 19 chances out of 20 the figures have a degree of precision
  - Giselle's norms
  - Normal != desirable
  - Armchair correspondent
  - How they use language to couch statistics, over 3/4 have power available, available is not powered
  - What is probable error and why is it no longer used

  - When numbers in tabular form are taboo and words will not do the work well. as is often the case. there is one answer left: Draw a picture.
  - Beware of pictographs with disproportionate widths/volumes
  - Suspiciously precise
  - Clear weather is more dangerous that foggy weather
  - Safer to be in the Navy than NYC
  - Have students go to spurious correlation website
  - The darkening shadow first national bank of Boston and Newsweek
  - Shifting percentages based on different bases. %lift as a percent of before or after. To offset a pay cut of 50%, you need a raise of 100%
  - Percent, percentage point, percentile
  - Statistics is as much art as it is practice

- Weapons of Math Destruction - Cathy O'Neil
- The Data Detective - Tim Harford
- May Contain Lies

  - Knowing is have the battle, so even with the knowledge, we're only halfway there

- Harvard ebook: Intro to Data Science: Intro to Data Visualization
- Visual Complexity: Mapping Patter of Information - Manual Lima

Blogs

- Krisztina Szucs' blog

Cheatsheets

- medium.com/responsibleml/data-visualization-cheat-sheets-1c12ba8a7671
- Posit cheat sheets: https://posit.co/resources/cheatsheets/
- Ggplot2 data visualization cheat sheet: https://rstudio.github.io/cheatsheets/data-visualization.pdf
- Data Viz cheat sheet: https://s3.amazonaws.com/assets.datacamp.com/email/other/Data+Visualizations+-+DataCamp.pdf
- Core principles of datavis: https://policyviz.com/wp-content/uploads/2018/08/PolicyViz-DataViz-Cheatsheet.pdf

Courses

- UC Berkley: Visualization enhances exploratory analysis as well as efficient communication of data results. This course focuses on the design of visual representations of data in order to discover patterns, answer questions, convey findings, drive decisions, and provide persuasive evidence. The goal is to give you the practical knowledge you need to create effective tools for both exploring and explaining your data. Exercises throughout the course provide a hands-on experience using relevant programming libraries and software tools to apply research and design concepts learned.
- Coursera dataviz course: https://www.coursera.org/learn/jhu-data-visualization-r

Datasets

- https://r-packages.io/datasets
- ajthurston.com/validation
- IBM Human Resources Attrition Dataset

Glossary

- Data visualization is the practice of communicating quantitative information through visual means. It's goal is to take detailed individual or aggregate data to identify trends, group differences, and relationships among data to make the most salient findings readily accessible to technical and non-technical audiences alike. Data tables are not data visualization as they do not summarize data they simply present it.
- Organizational data visualization is communicating quantitative information visually for the purposes of aiding in work-related decision-making.
- This is contrasted from data art, which is the creation of aesthetically enjoyable data displays principally for their novelty or enjoyment.

Sources

- https://hbr.org/data-visuals
- https://hbr.org/offers/visual-library
- https://medium.com/tag/data-visualization
- https://www.modernstatisticswithr.com/
- https://policyviz.com/
- https://www.nytimes.com/spotlight/graphics
- https://www.qlik.com/blog
- https://stephanieevergreen.com/blog/
- https://www.storytellingwithdata.com/blog
- https://tellingstorieswithdata.com/
- http://www.thefunctionalart.com/
- https://visualcapitalist.com
- https://visualisingdata.com/

Tools

- Tool for adding live coding in course: https://r-wasm.github.io/quarto-live/
- Surveydown for collecting data: https://x.com/johnhelveston/status/1828124022514577432
- Chart pickers:
  - From Data to Viz visualization chooser: https://www.data-to-viz.com/
  - https://datavizcatalogue.com/
- Esquisse package R drag and drop ggplot2

Videos

- Nicola Rennie Spaghetti code talk: https://nrennie.rbind.io/talks/useR-spaghetti-code/slides.html#/section; video: https://www.youtube.com/watch?v=wbhWl5-xR10
- Bayron design hour long video on data visualization: https://youtu.be/tnikYc0O0-o
- x.com/SpencrGreenberg/status/1746562746206863595/photo/2
- clearerthinking.org/post/we-tested-the-predictive-power-of-astrology-here-are-the-results

Random

- Games:
  - Data Land Board Game
  - Charty Party Card Game
- Newsweek fairness meter: https://www.newsweek.com/fairness-meter

Resources

- Visualizing only inferential uncertainty, compared to outcome variability, can result in significant overestimates of treatment effects, even among highly trained experts (Zhang, et al., 2023). Inferential uncertainty (how precisely a statistic is estimated) and outcome variability (how much individual outcomes differ). This confusion can lead to an "illusion of predictability," where people overestimate the certainty and significance of scientific findings.
- Data analysis workfow: https://archive.is/T1TsC
- Citedrive for bibliographic references for RStudio: citedrive.com/en/quarto
- Book Title: Organizational Data Visualization: Communicating clearly, concisely, and convincingly.
- Blue dot effect in picking colors for continuous variable scale like viridis, see cognitive biases notes in obsidian
- https://pyoflife.com/data-visualization-in-r-with-100-examples/
- https://illustratingdata.com/tag/teagan-white/
- https://handsondataviz.org/
- https://datavizs24.classes.andrewheiss.com/lesson/01-lesson.html
- Videos on cognitive load
  - https://youtube.com/shorts/nmNMasnNBQY
  - https://www.youtube.com/watch?v=Pmr_DeM815w
- It would be tough to visualize a latent state–trait model with multiple indicators and autoregression across four measurement occasions in a simple, succinct, and "beautiful" way. Instead, the best visuals, in terms of effectively communicating and persuading the audience, focus on not much more than means (e.g., bar charts), variances (e.g., scatterplots), and groupings (e.g., social networks) (Ahmad & Zhou)
- To form positive judgments of numerical information, people must be able to evaluate the "goodness" of a number (Voss & Lake, 2020)
- Storytelling with data seminar: The art of exploring and explaining data
- Ted Talk (great statistics lecture): Hans Rosling: The best stats you've ever seen

- https://www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty?language=en
- https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen
- R Discovery App
- It would be tough to visualize a latent state–trait model with multiple indicators and autoregression across four measurement occasions in a simple, succinct, and "beautiful" way. Instead, the best visuals, in terms of effectively communicating and persuading the audience, focus on not much more than means (e.g., bar charts), variances (e.g., scatterplots), and groupings (e.g., social networks) (Ahmad & Zhou)
- To form positive judgments of numerical information, people must be able to evaluate the "goodness" of a number (Voss & Lake, 2020)
- Can grab attention tho: https://x.com/szucsi/status/1798700114690597353
- Misleading graphs are another crucial topic in data visualization. These are charts that misrepresent data, leading to incorrect conclusions. They can be intentionally deceptive or result from errors in data handling or graph construction. Common methods of creating misleading graphs include truncating axes, improper scaling, using 3D effects, omitting data, and biased labeling. These practices can distort data interpretation, making small differences appear significant or large differences appear minor. To avoid misleading graphs, strategies include clear labeling, appropriate scaling, and avoiding unnecessary complexity. Proper data visualization aims to present data accurately without manipulation, ensuring the audience can interpret the information correctly.
- How to make figures and presentations friendly to colorblind people: https://jfly.uni-koeln.de/color/
- Modern data viz in R: https://pyoflife.com/modern-data-visualization-with-r/
- Dataviz with R course: https://cognitiveclass.ai/courses/data-visualization-with-r#about-course
- Multiple courses: https://www.datavis.ca/courses/

Resources

- Charts to make sense of 2024 from HBR: https://archive.ph/EJy0O
- NYT: From Inflation to Bitcoin, 9 Charts That Explain 2024: archived link

References

Zhang, D. C., S. Highhouse, M. E. Brooks, and Y. Zhang. 2018. "Communicating the Validity of Structured Job Interviews with Graphical Visual Aids." *International Journal of Selection and Assessment*. https://doi.org/10.1111/ijsa.12220.