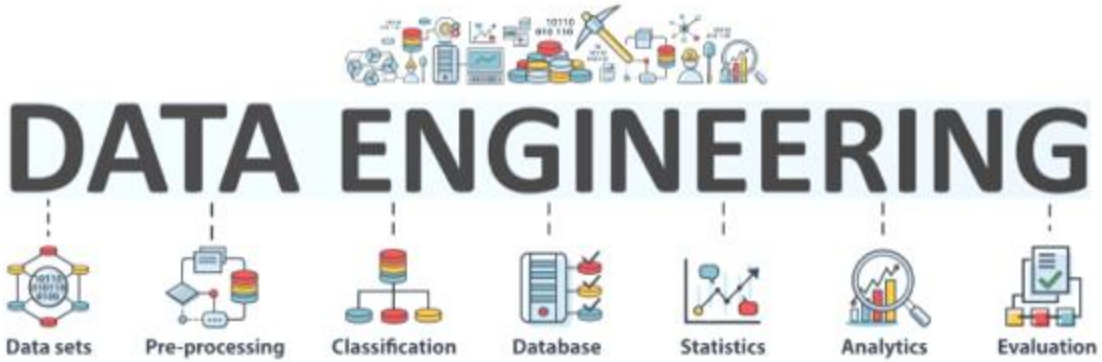


香港城市大學
City University of Hong Kong

Lecture 4: Data Visualization

CS5481 Data Engineering

Instructor: Linqi Song

Outline

1. Introduction to Data Visualization
2. Conventional Visualizations
3. Data Visualization Skills
4. Tips and Tricks
5. Data Visualization Traps
6. Visualization and Dashboard Software

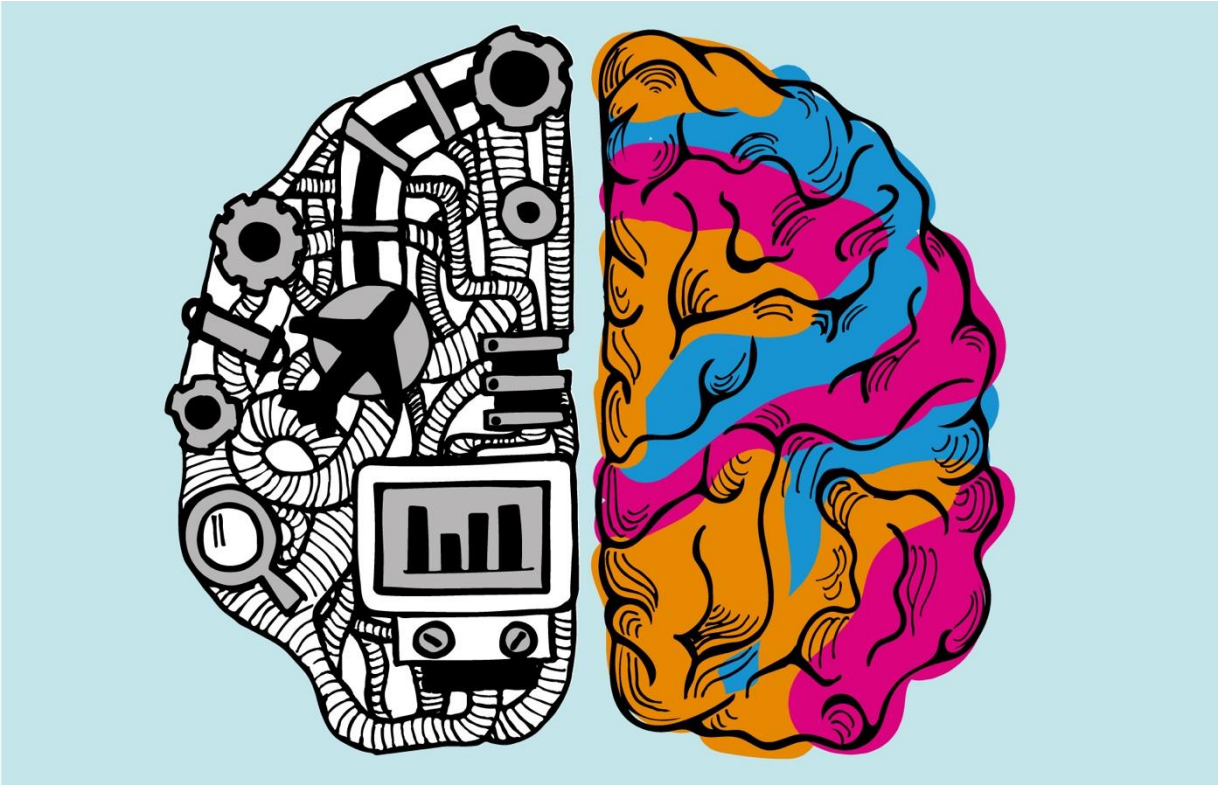
Why do we need data visualization?

- The power of the unaided mind is highly overrated. The real powers come from devising external aids that enhance cognitive abilities.
  - Without external aids, memory, thought, and reasoning are all constrained.
  - But human intelligence is highly flexible and adaptive, superb at inventing procedures and objects that overcome its own limits, e.g., pencil and paper.

- How have we increased memory, thought, and reasoning?
  - By the invention of external aids: Some assistance comes through cooperative social behavior;
  - some arises through exploitation of the information present in the environment; and
  - some comes through the development of tools of thought – cognitive artifacts – that complement abilities and strengthen mental powers.

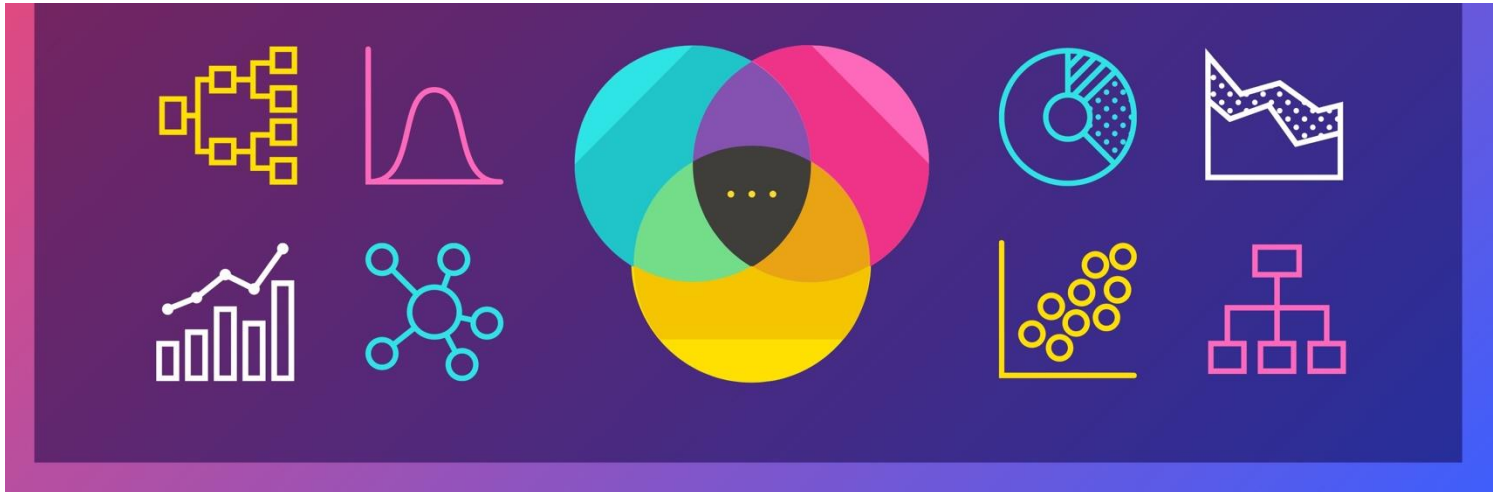
There is a gap between human brain and data

- **Human recognition cannot directly process computerized data.**
- **Human brains are good at visual cognition.**

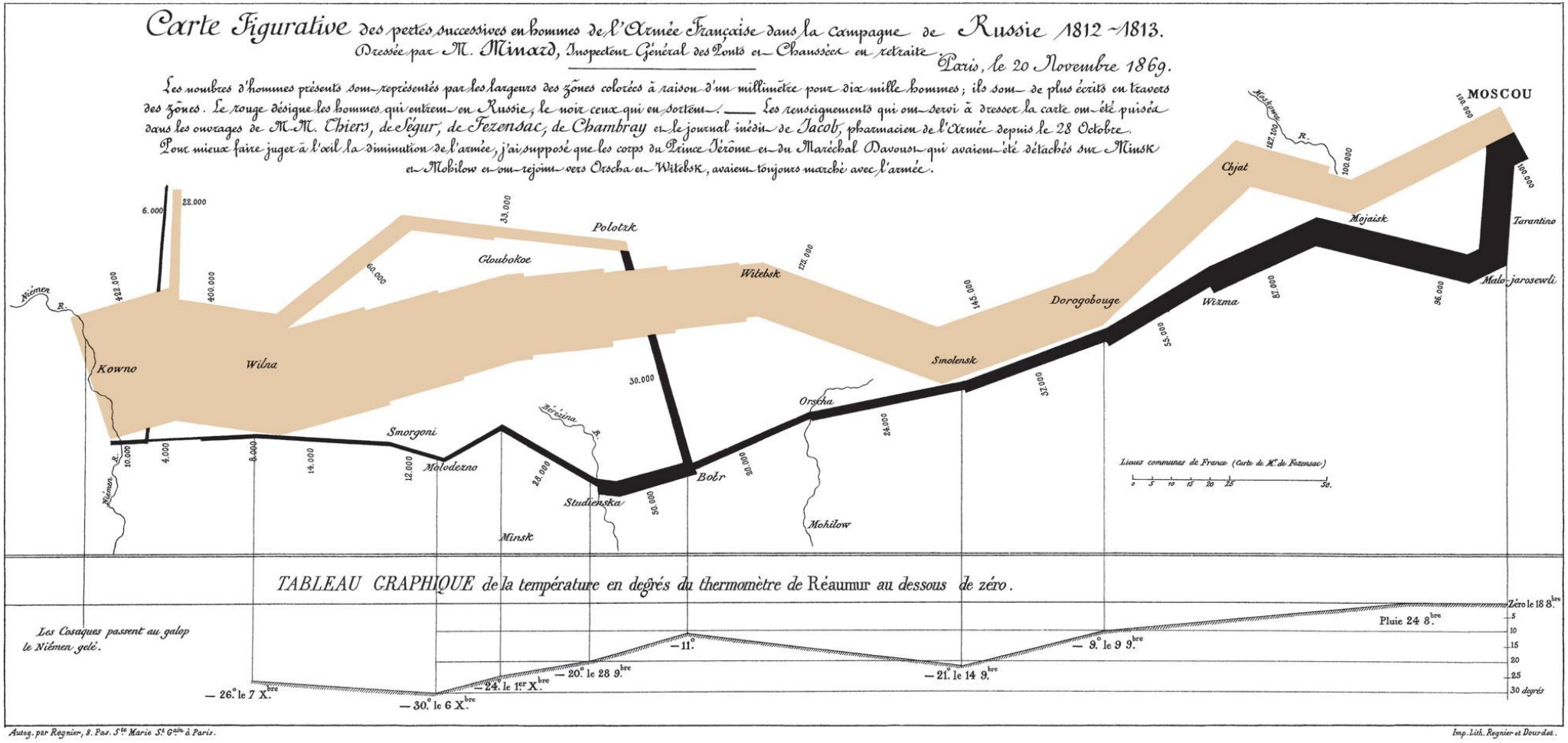


Data visualization

- **Data visualization is the study of visual representations of abstract data to reinforce human cognition.** Its goal is to make information easy to comprehend, interpret, and retain.
- The discipline of communicating information through the use of visual elements such as graphs, charts, and maps.



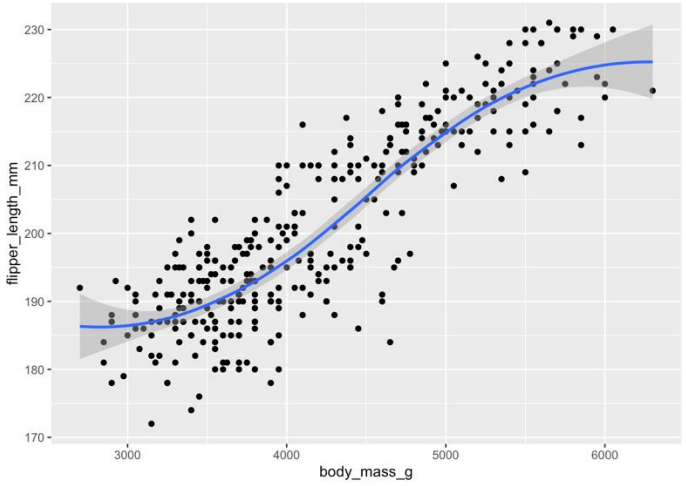
A data visualization example



Charles Minard's map of Napoleon's march time and time again – almost to the point of exhaustion

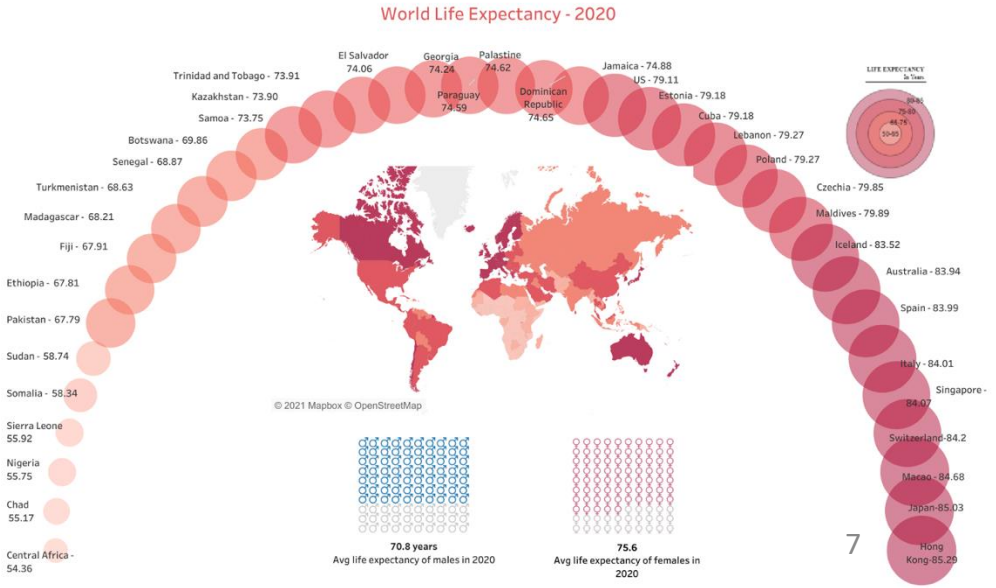
Exploratory vs. explanatory data visualization

- Exploratory data visualization
  - Exploratory visualizations are used when you want or need to explore data to find insights. You use these types of visualizations to help better understand your underlying data.
  - Sharing is limited to internal teams/research groups and not published or presented to larger groups



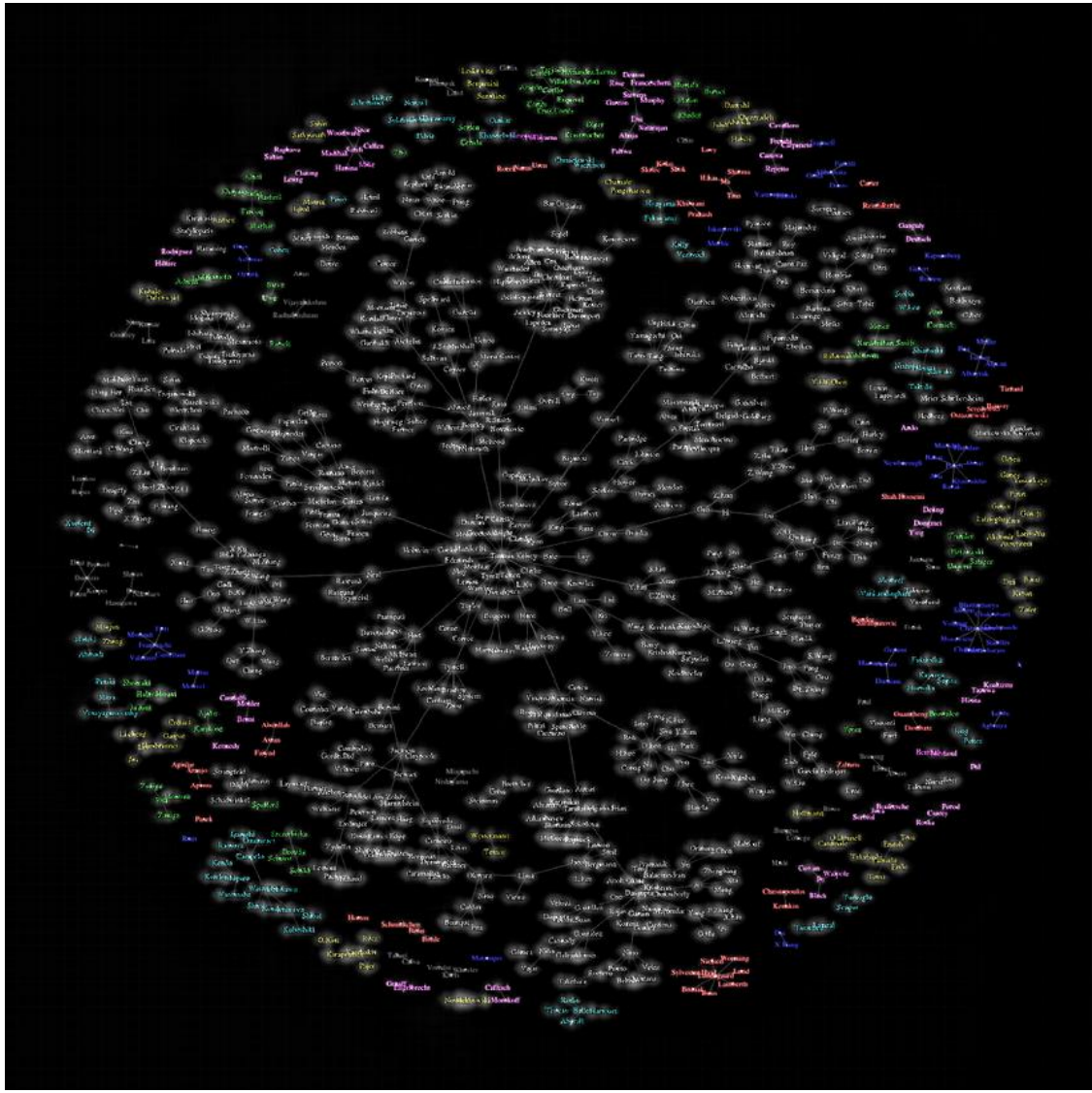
Body mass vs. flipper length of chickens

- Explanatory data visualization
  - Explanatory visualizations are used to communicate the results of your analyses.
  - Use for data storytelling/presentations



Data visualization – a combination of science and art

- Science – skills
  - A skill to handle “the graphical display of data”, to select a good way to present, to use specific software, etc.
- Art – the beauty of data
  - The main way complicated problems are explained to decision makers.
  - A good graphic tells a story. To help identify patterns in a data set and explain those patterns to a wider audience.
  - Everything should be made as simple as possible, but no simpler.
  - Ink is cheap. Electrons are even cheaper.



Choosing appropriate visualizations - statistics

For data visualization to be of value, choose the visualization that effectively delivers your findings to your audience.

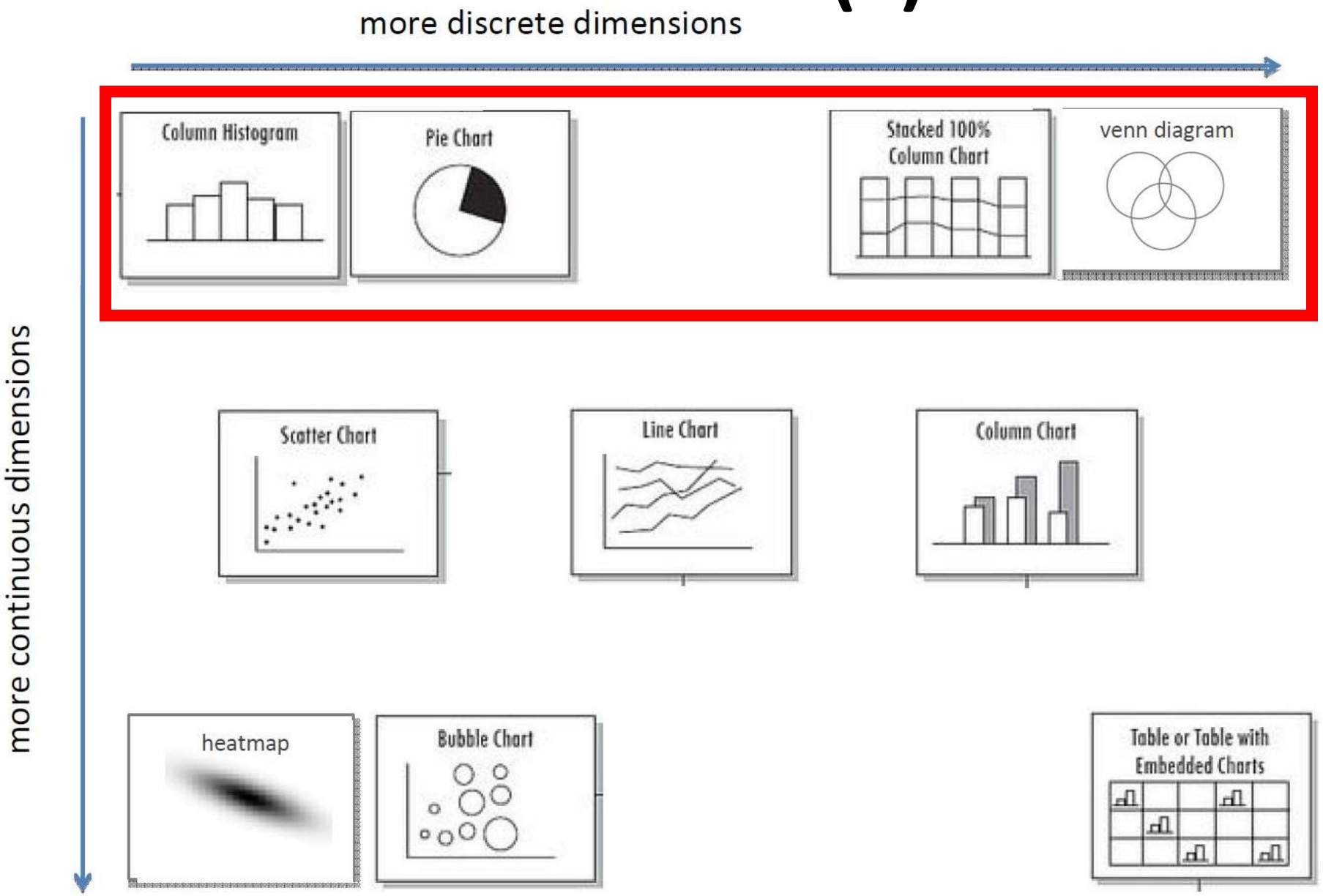
- What is the relationship that I am trying to establish?
- Do I want to compare multiple values?
- Do I need my audience to see the correlation between two variables?
- Do I want to detect anomalies in data?

Choosing appropriate visualizations - audiences

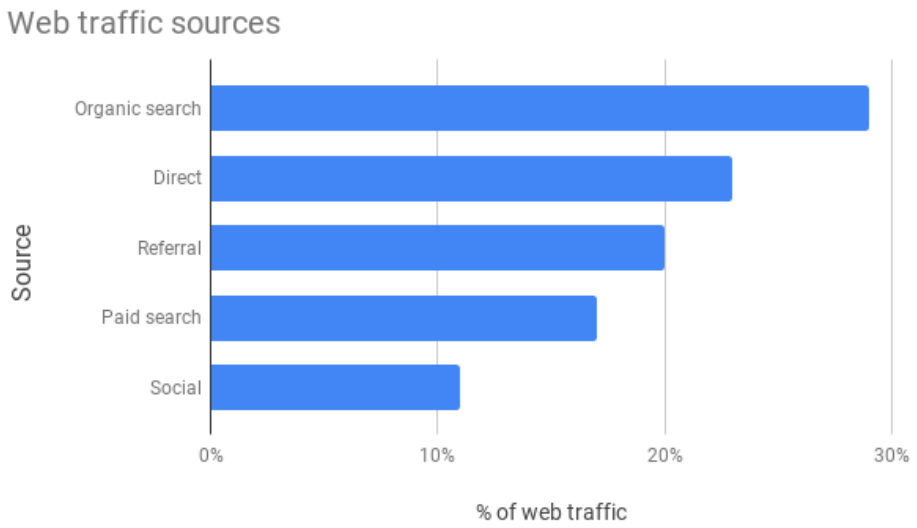
You need to be able to answer this question for your audience with every dataset and information that you visualize.

- What should be the key takeaway for my audience?
- What does my audience need to know?
- What are the questions they have?

Conventional visualizations (1)

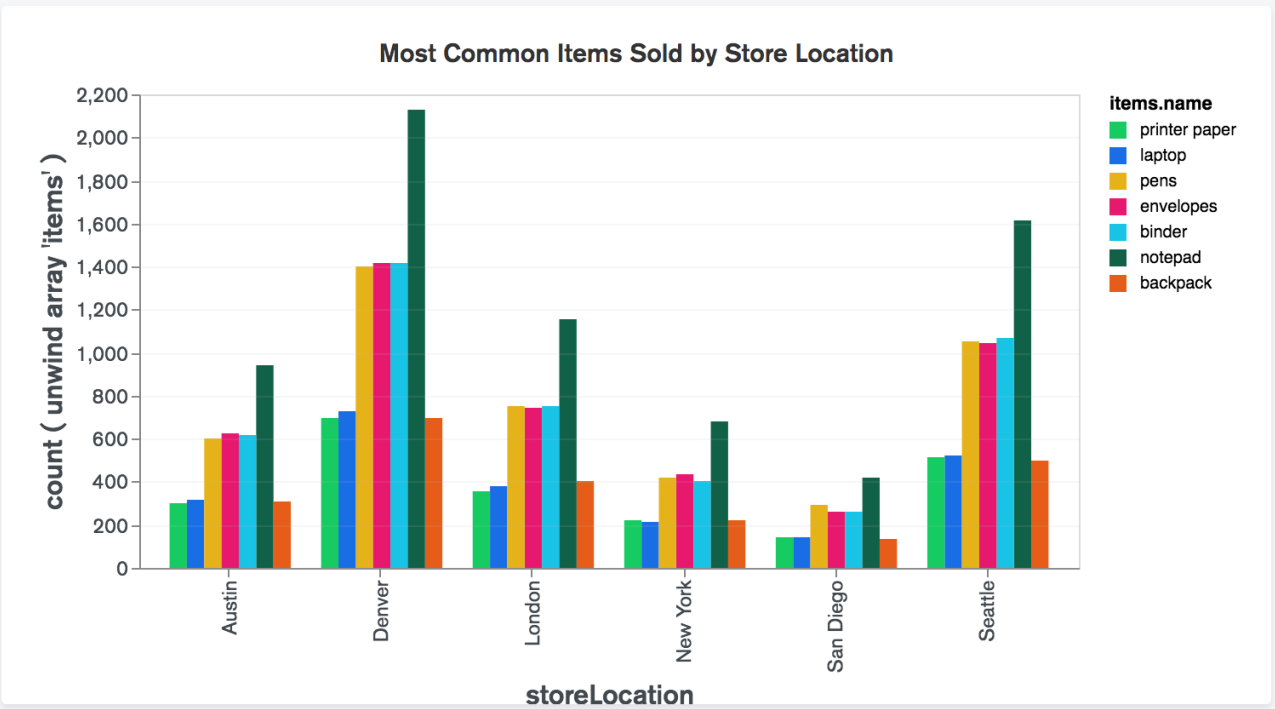


Visualizations – charts



Bar chart

Great for comparing related data sets or parts of a whole.



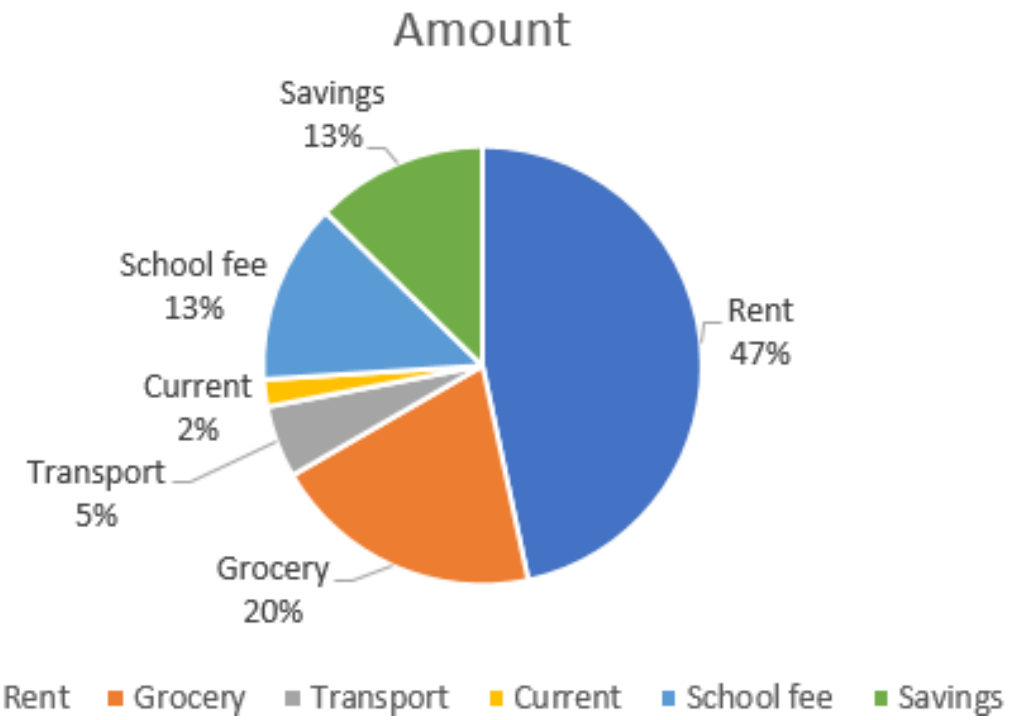
Column chart

Compare values side-by-side. You can use them quite effectively to show change over time

Visualizations – pie charts

Pie chart

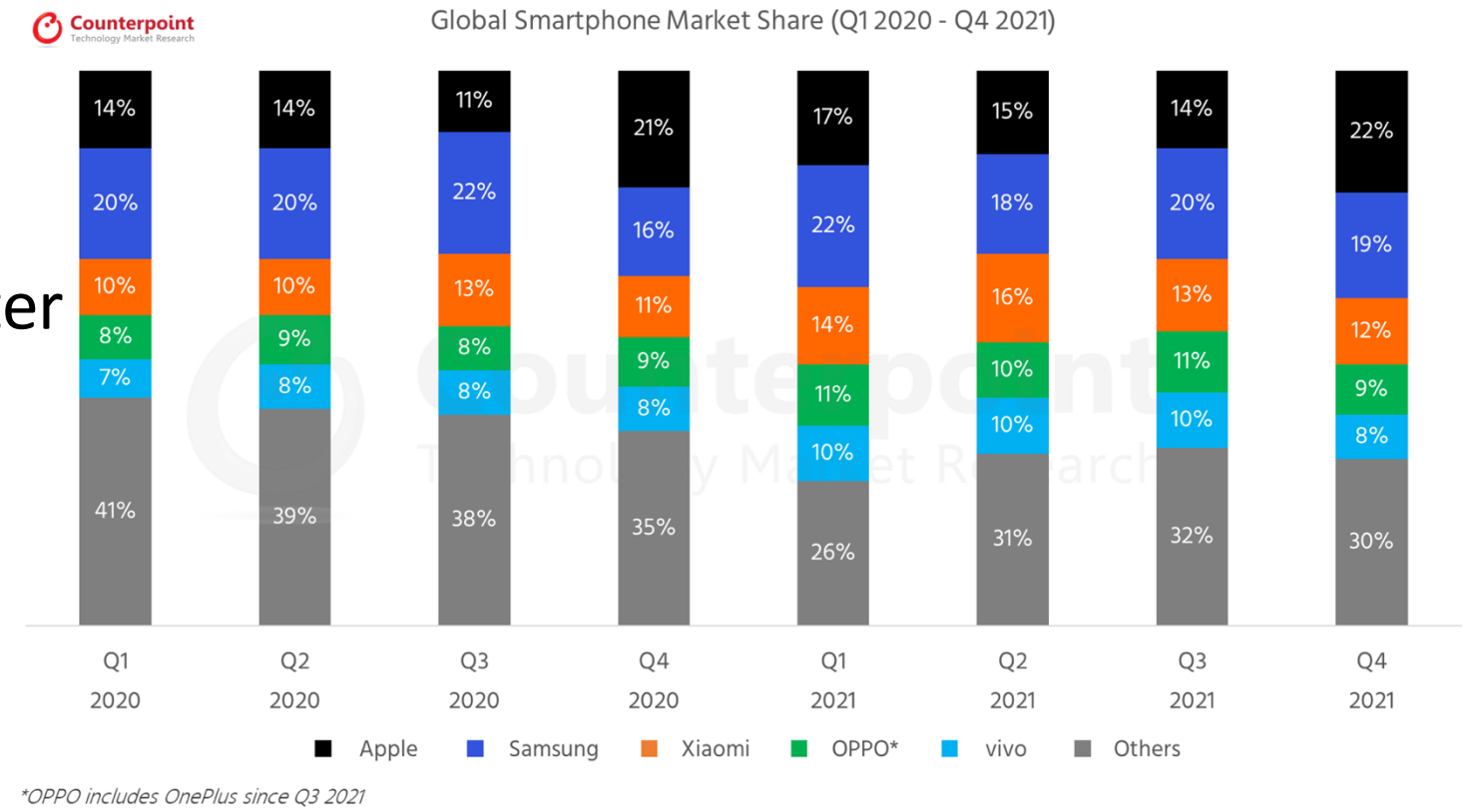
- Show the breakdown of an entity into its sub-parts and the proportion of the sub-parts in relation to one another.
- Each portion of the pie represents a static value of category, and the sum of all categories is equal to hundred percent.



Visualizations – stacked bar graphs

Stacked bar graph

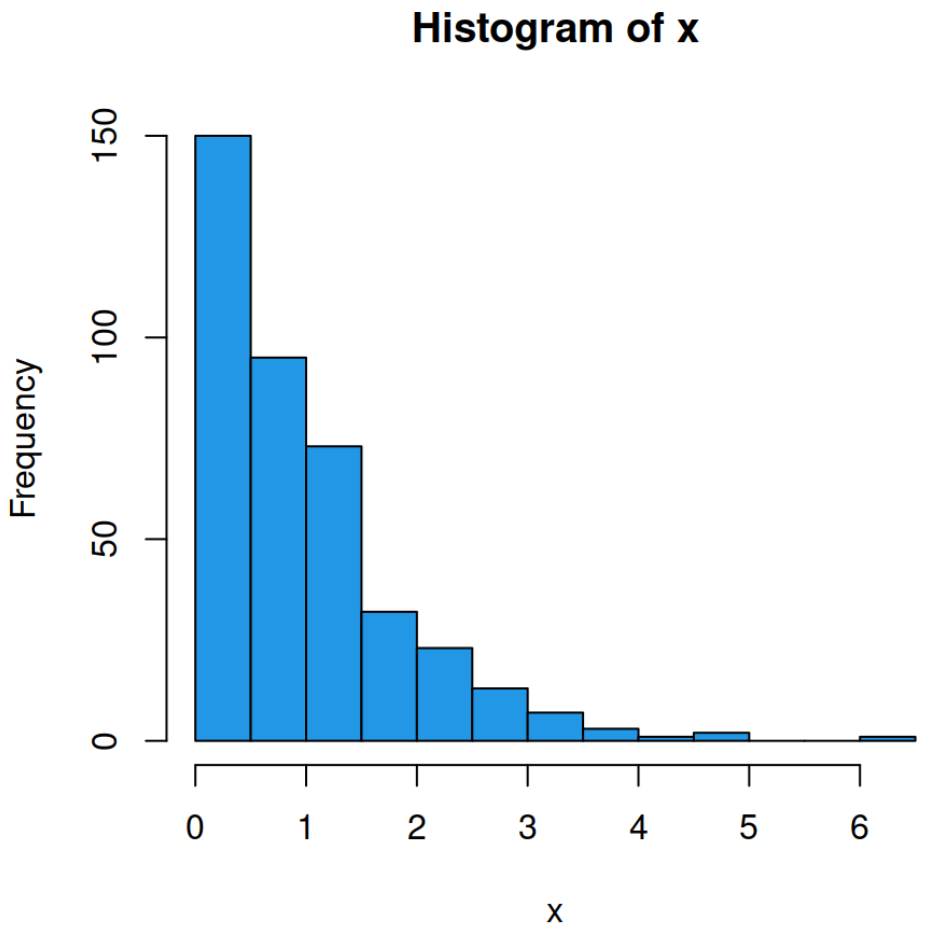
- Wholes split into parts.
- Easy to compare – often better than pie chart.
- Can have multiple discrete dimensions.



Visualizations – histograms

Histogram

- Important first way of looking at your data.
- One dimensional
- Shows shape by binning a continuous distribution

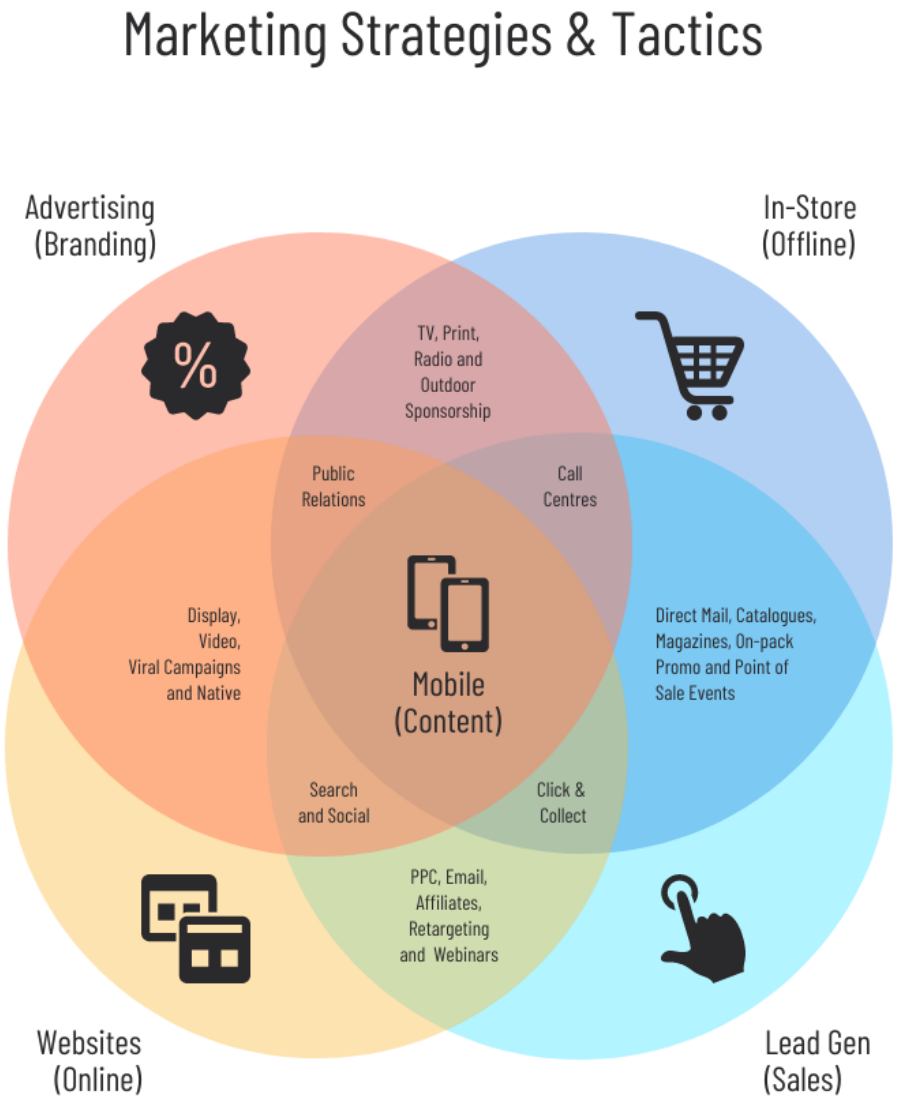


Frequencies of search queries

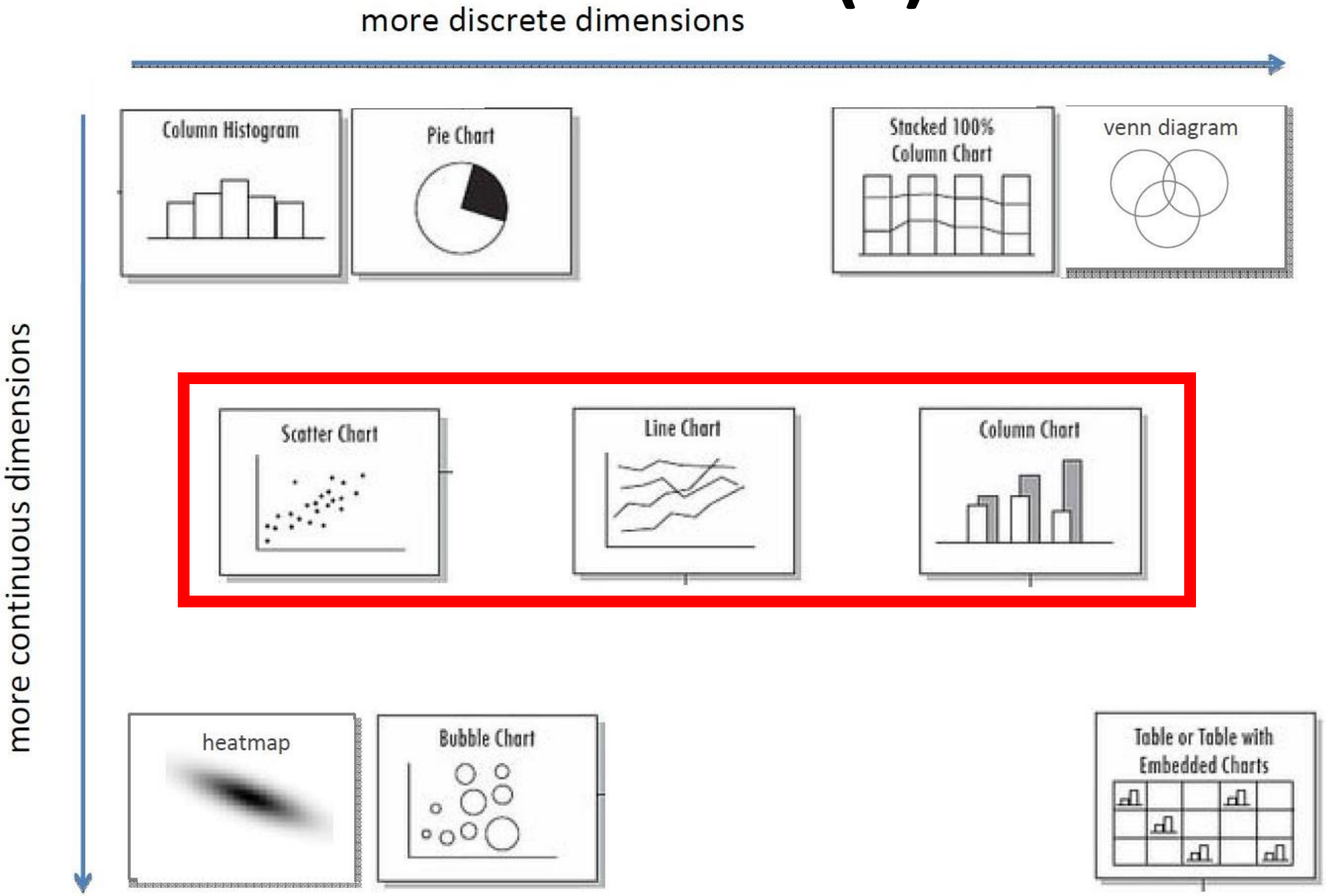
Visualizations – venn diagrams

Venn diagram

- Shows overlap between discrete groups.
- Sometimes the only way to display overlapping sets.
- Unintuitive – no “popout”



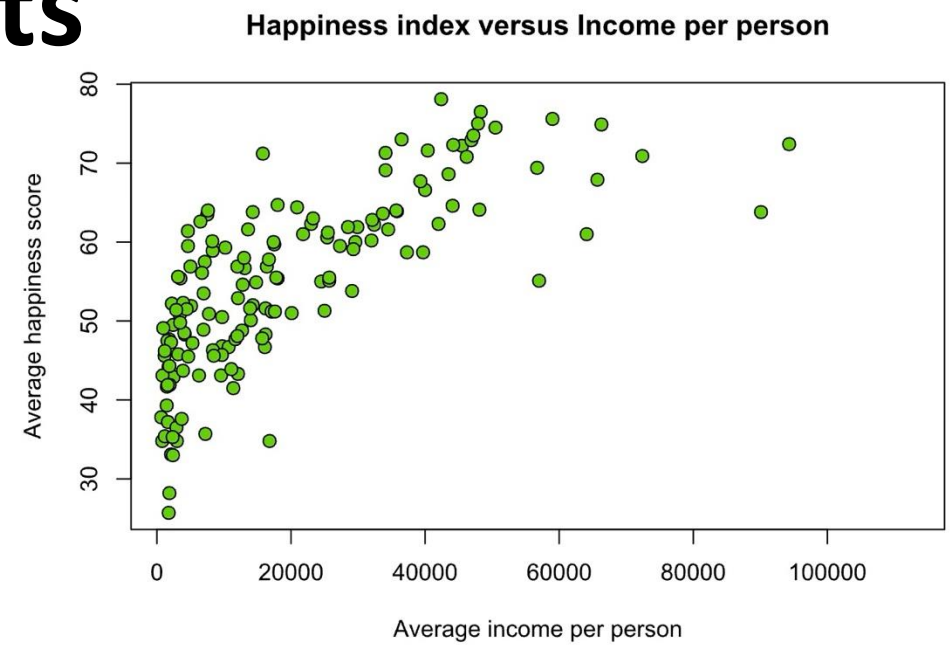
Conventional visualizations (2)

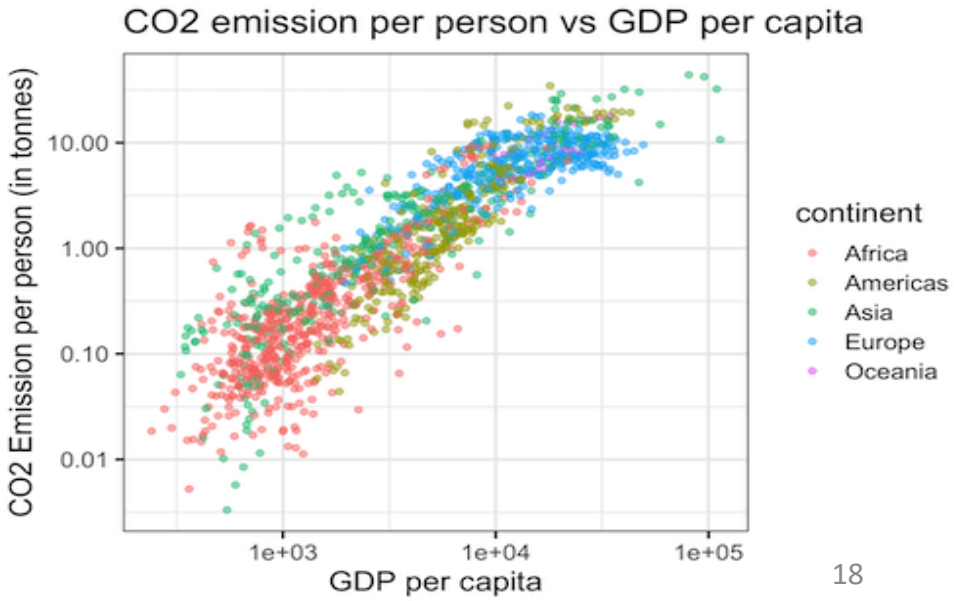


Visualizations – scatter plots

Scatter plot

- Relationship between observations on two continuous dimensions
- Can show multiple groups
- Can show trend lines etc.
- Uninformative with too much data

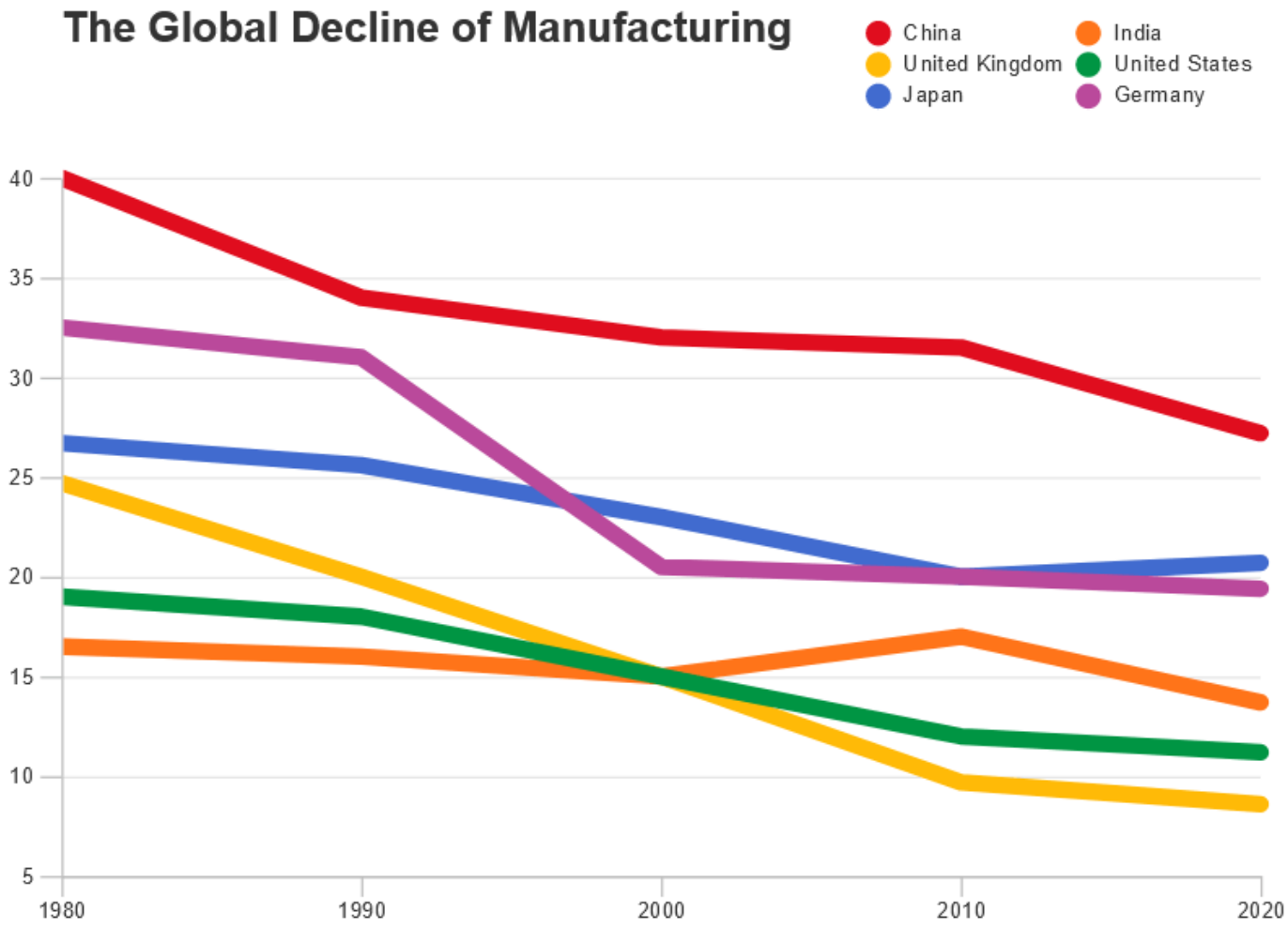




Visualizations – line graphs

Line Graph

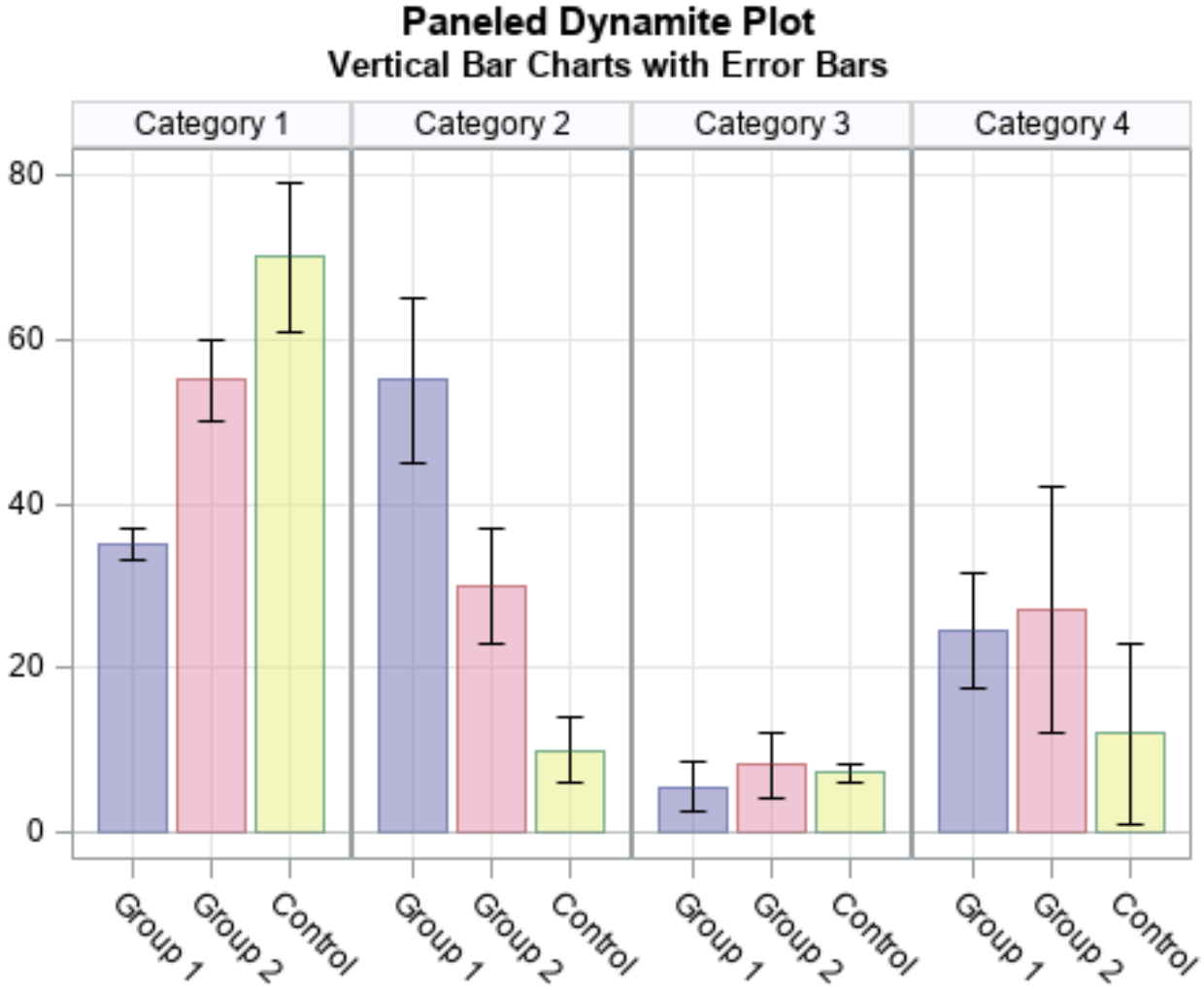
- Also ubiquitous.
- Good for showing one variable as continuous even though you have discrete measures.
- Can compare several discrete groups.

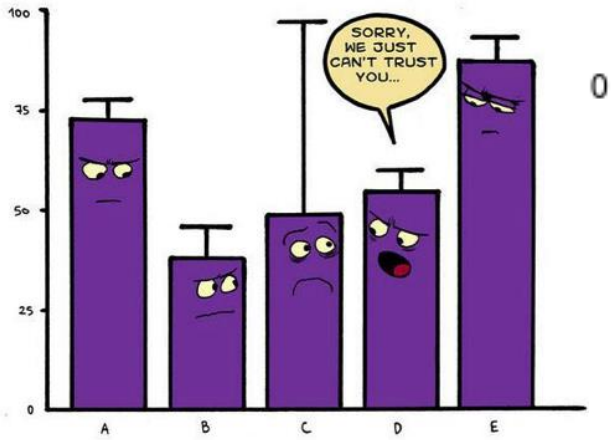


Visualizations – dynamite plots

Dynamite Plot

- Can be used for lots of discrete grouping factors.
- Natural semantics of grouping
- Conceals data.

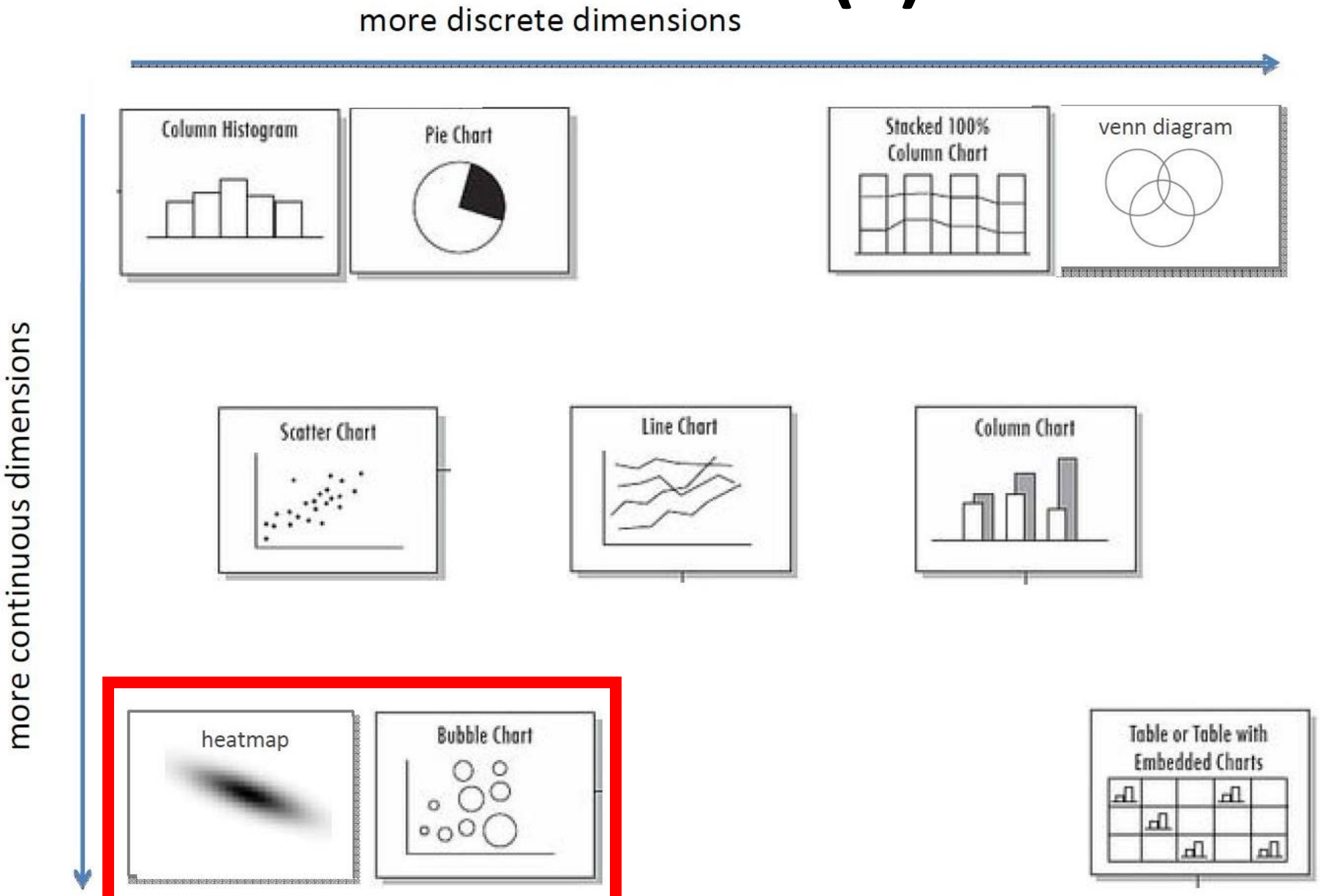




The Vanderbilt University Department of Biostatistics has a formal policy discouraging use of these plots, stating that:

Dynamite plots often hide important information. This is particularly true of small or skewed data sets. Researchers are highly discouraged from using them, and department members have the option to decline participation in papers in which the lead author requires the use of these plots.

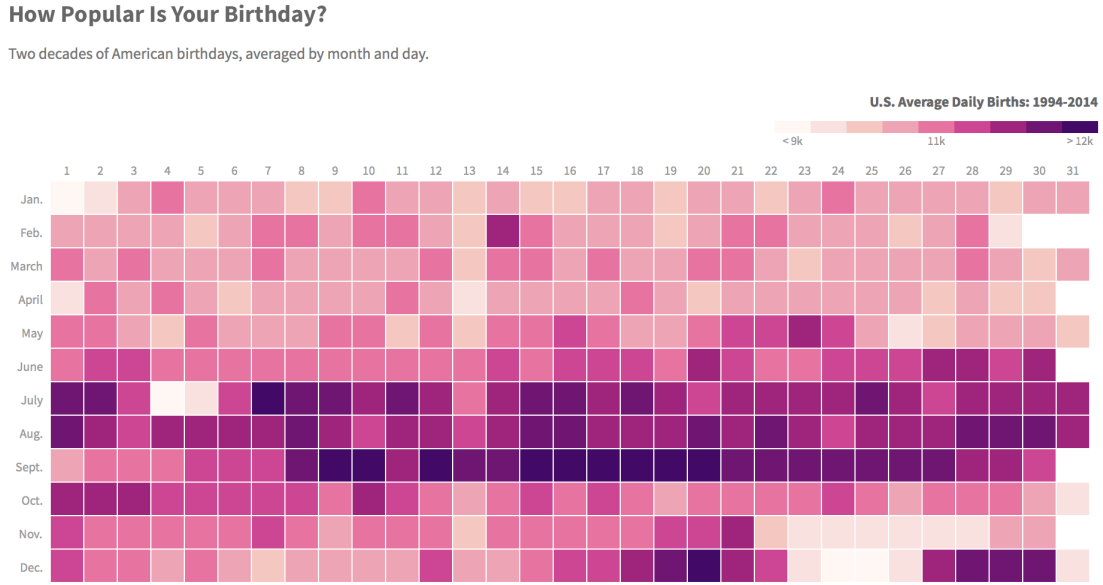
Conventional visualizations (3)

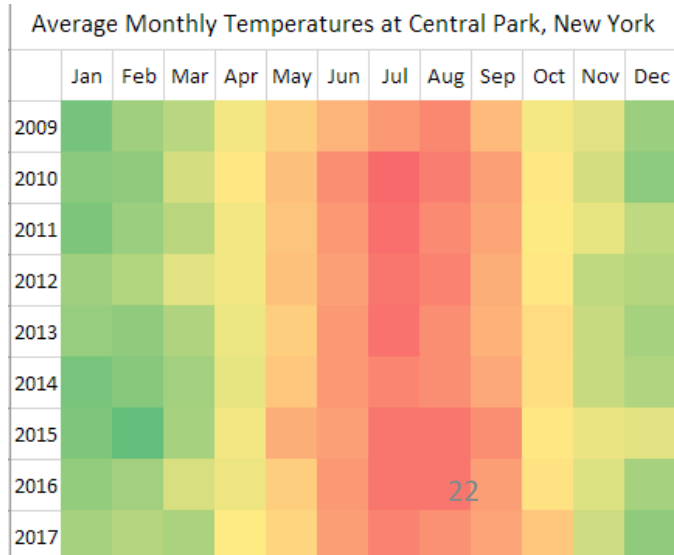


Visualizations – heat maps

Heat map

- Works very well when there are natural semantics.
- Color mapping can be problematic – grayscale usually fine.
- Can be unintuitive.

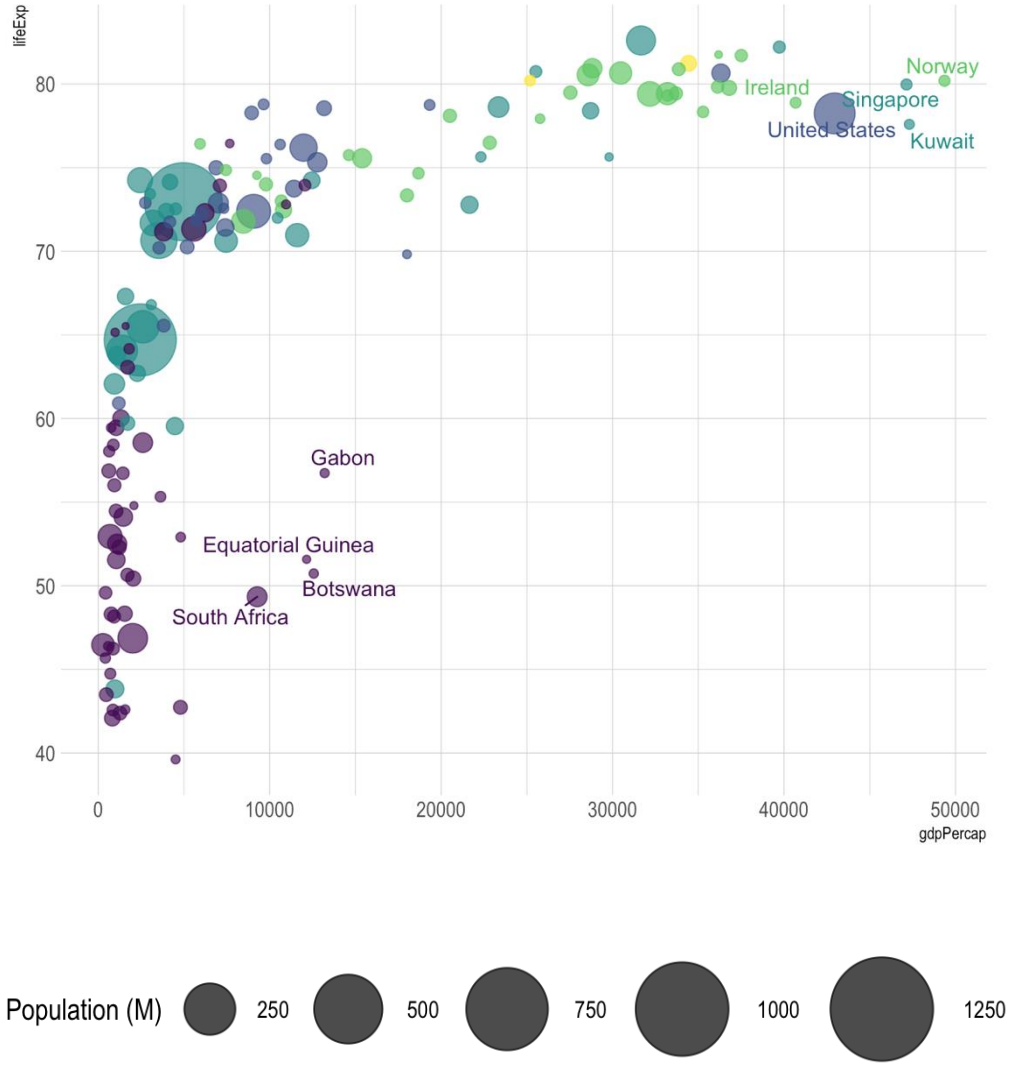




Visualizations – bubble plots

Bubble plot

- Can be very intuitive.
- Size is not perfectly quantitative.

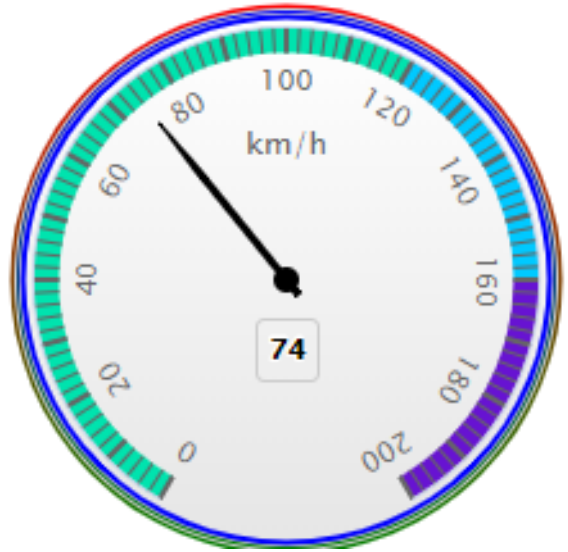


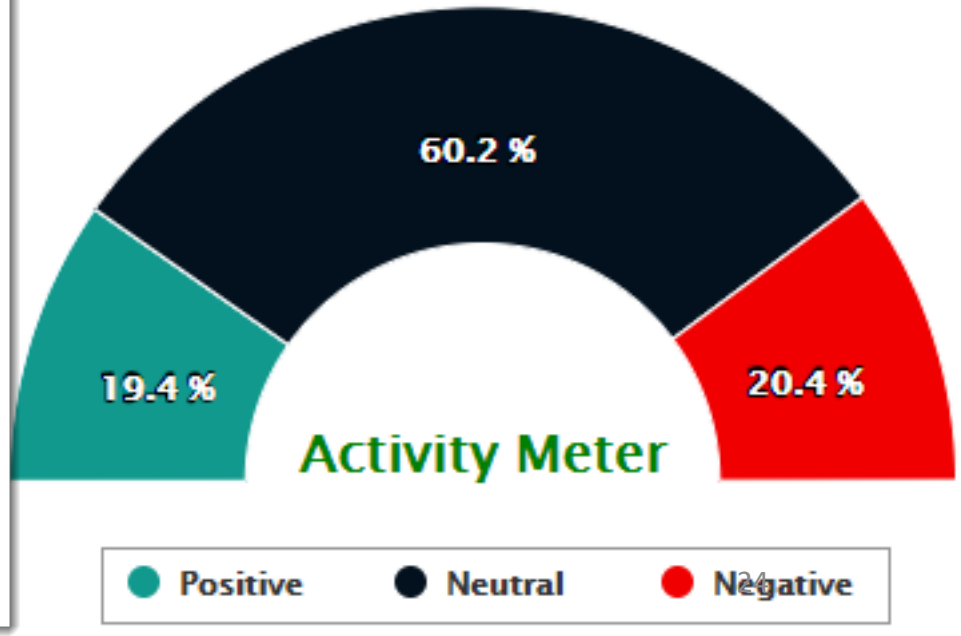
Different bubbles represent different countries; colors represent continents; sizes of the bubbles represent the size of population.

Others – gauges

Gauge

- Creates a gauge that indicates its metric value along an arc.

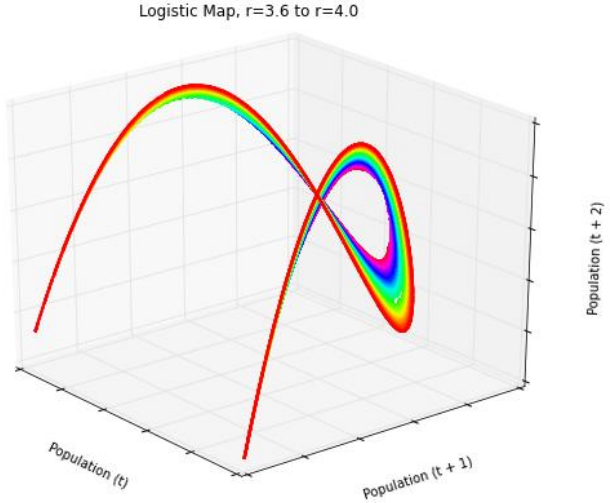


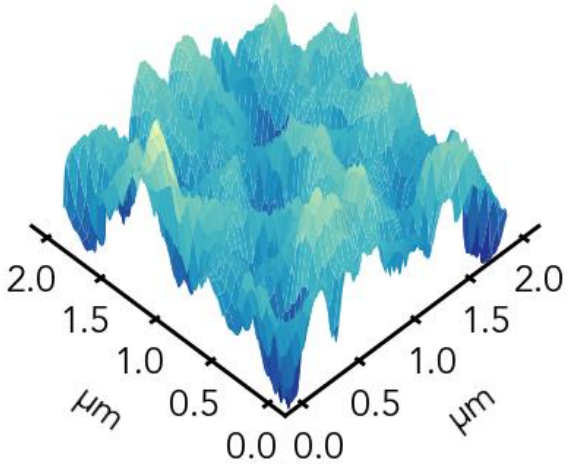


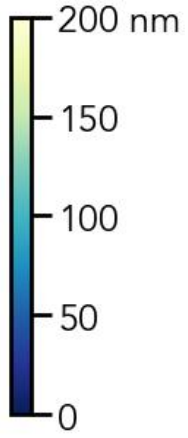
Others – 3D visualization

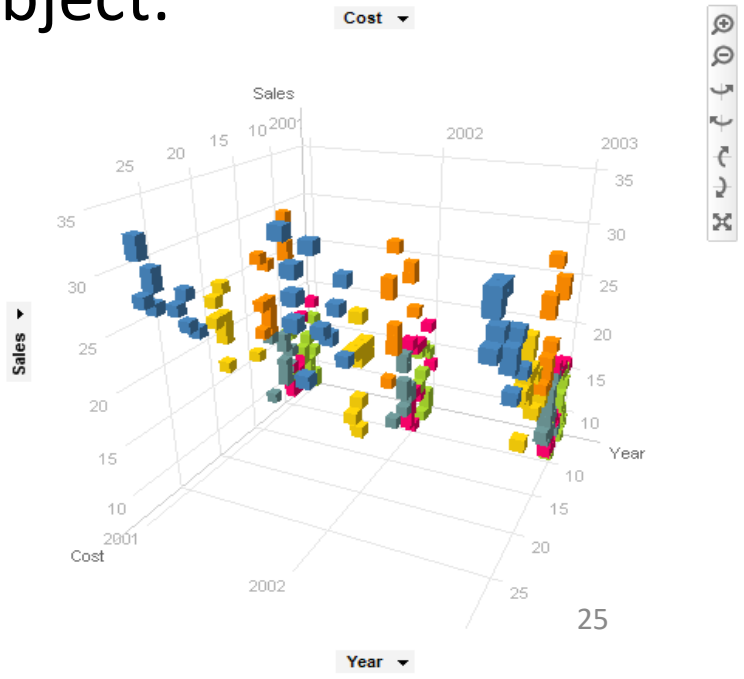
3D plot

- 3-D Dimensional data which provides the perception of depth, breadth, and height. Three-dimensional visualizations developed to provide both qualitative and quantitative information about an object.









Data visualization skills (1)

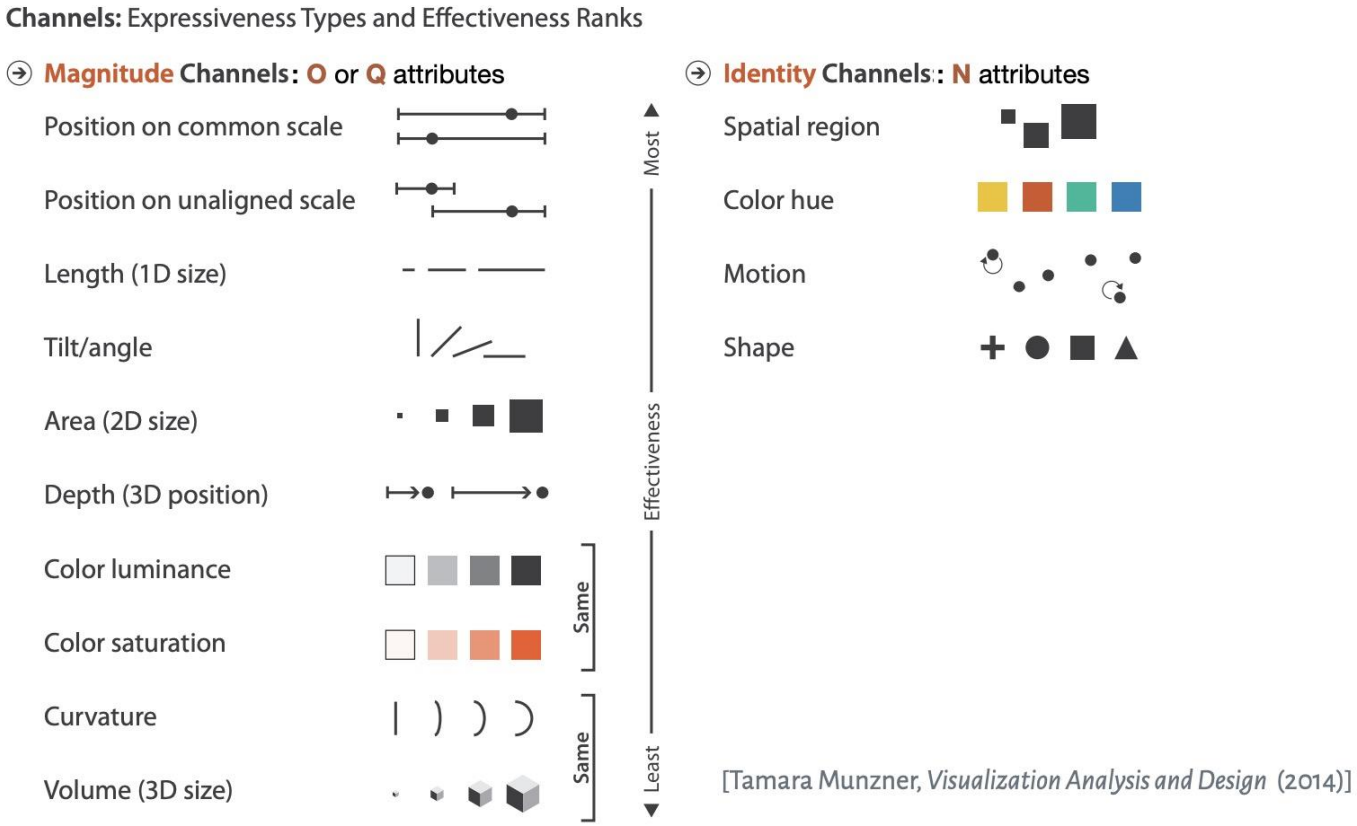
Find out what you want to say

This is the first and most important step in visualization preparation. You must ask yourself, “what is the purpose of this chart?”. Once we know the clear reason why the chart should exist, we will naturally be able to select the correct chart type for that reason.

Effectiveness vs. expressiveness

Effectiveness

To effectively map data to visuals, we need a level of abstraction. Data abstraction allows us to consistently encode the same "types" of data, even if different domains use different terminology to describe it



[Tamara Munzner, *Visualization Analysis and Design* (2014)]

Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Data visualization skills (2)

There are usually 6 reasons to make a chart:

1. To **compare**
2. To show the **distribution**
3. To explain **parts of the whole**
4. To tell the **trend over time**
5. To find out the **deviations**
6. To understand the **relationship**

To compare

- **What it means?**

  - You want to compare one set of value(s) with another.

- **Examples:**

  - Performance of Product A vs. Product B in 5 regions.

  - Interview performance of various candidates.

- **Charts that can be used for this reason:**

To show the distribution

- **What it means?**

  - You want to show the distribution of a set of values (to understand the outliers, normal ranges etc.)

- **Examples:**

  - Distribution of call waiting times in a call center

- **Charts that can be used for this reason:**

Parts of whole

- **What it means?**

  - You want to show how various parts comprise the whole.

- **Examples:**

  - Individual product sales as a percentage of whole revenue.

  - Browser types of customers visiting our website.

- **Charts that can be used for this reason:**

Trend over time

- **What it means?**

  - You want to understand the trend over time of some variable(s).

- **Examples:**

  - Customer footfalls on the last 365 days.

  - Share price of MSFT in the last 100 trading sessions.

- **Charts that can be used for this reason:**

To find out the deviations

- **What it means?**

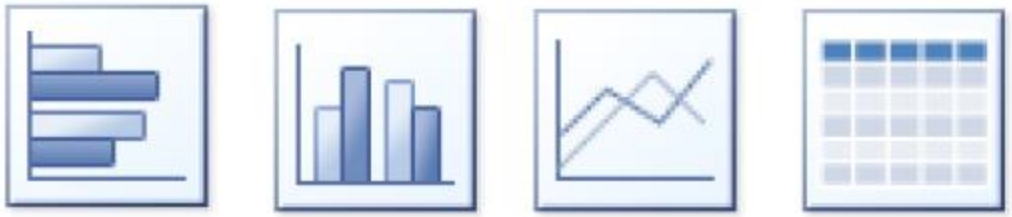
  - You want to see which values deviate from the norm.

- **Examples:**

  - Failures (or bugs) in the context of quality control.

  - Sales in various stores.

- **Charts that can be used for this reason:**



To understand the relationship

- **What it means?**

  - You want to establish (or show) relationship between 2 (or more) variables.

- **Examples:**

  - Relationship between Search Phrases and Product Purchases in your website.

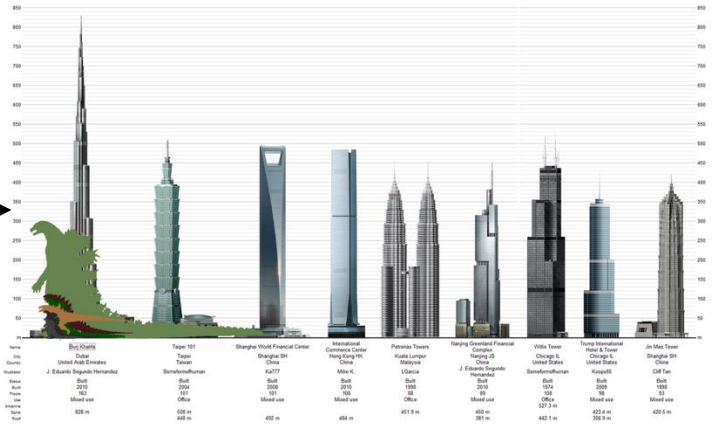
  - Relationship between in-store sales and holidays.

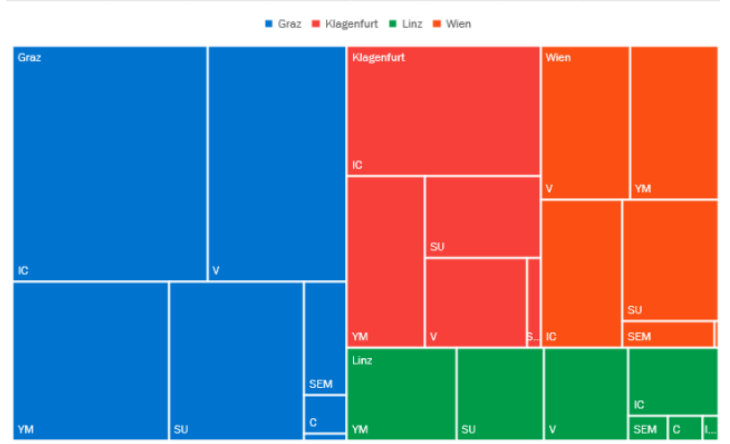
- **Charts that can be used for this reason:**

Data visualization skills

Skills

1. **Use graph to represent number: I**nstead of bars and lines, use some related figures. Integrating and clear.
2. **Comparison**
3. **2D instead of 1D:** Use area instead of bars to represent data
4. **Metaphor:** Give a similar concept to tell the story
5. **Flow:** Circle, Arrow



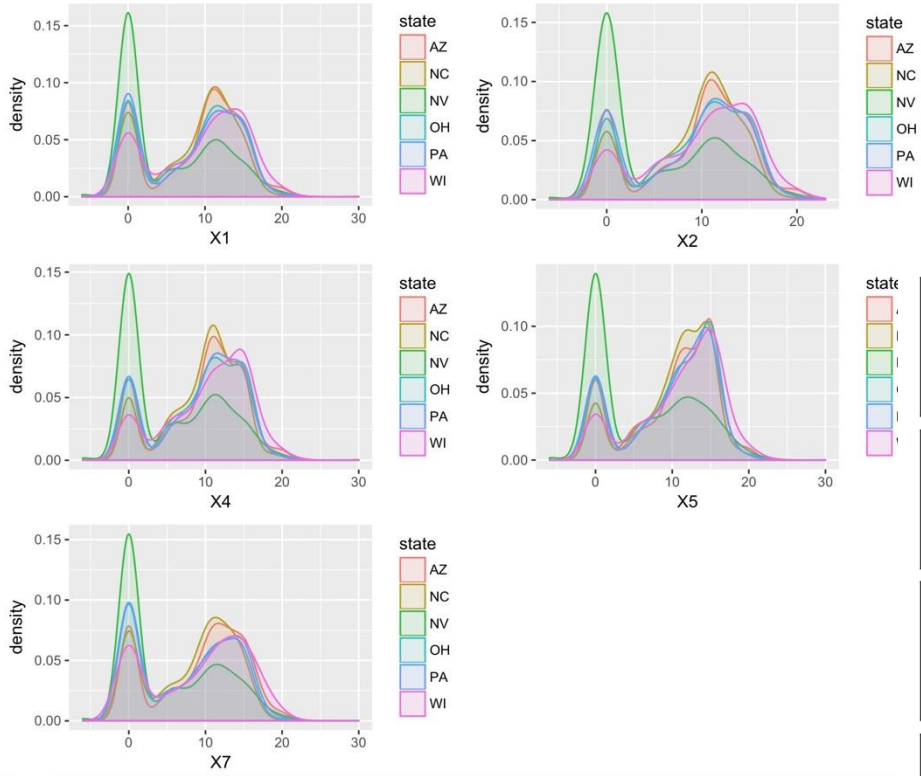


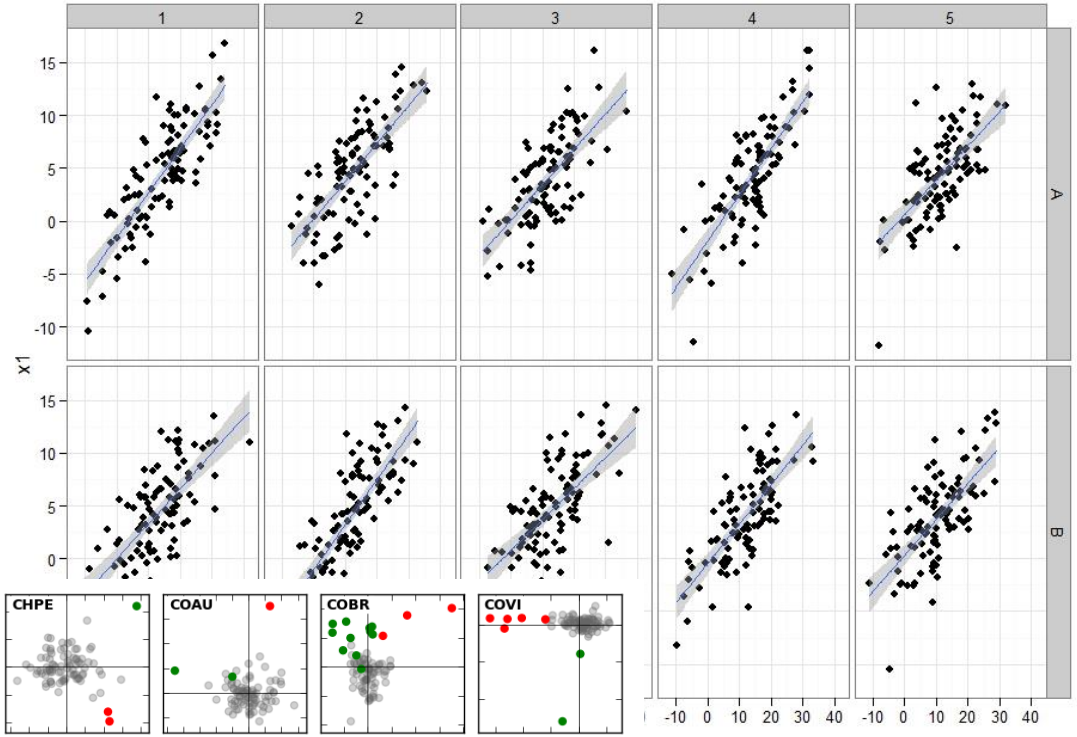
Tips and Tricks

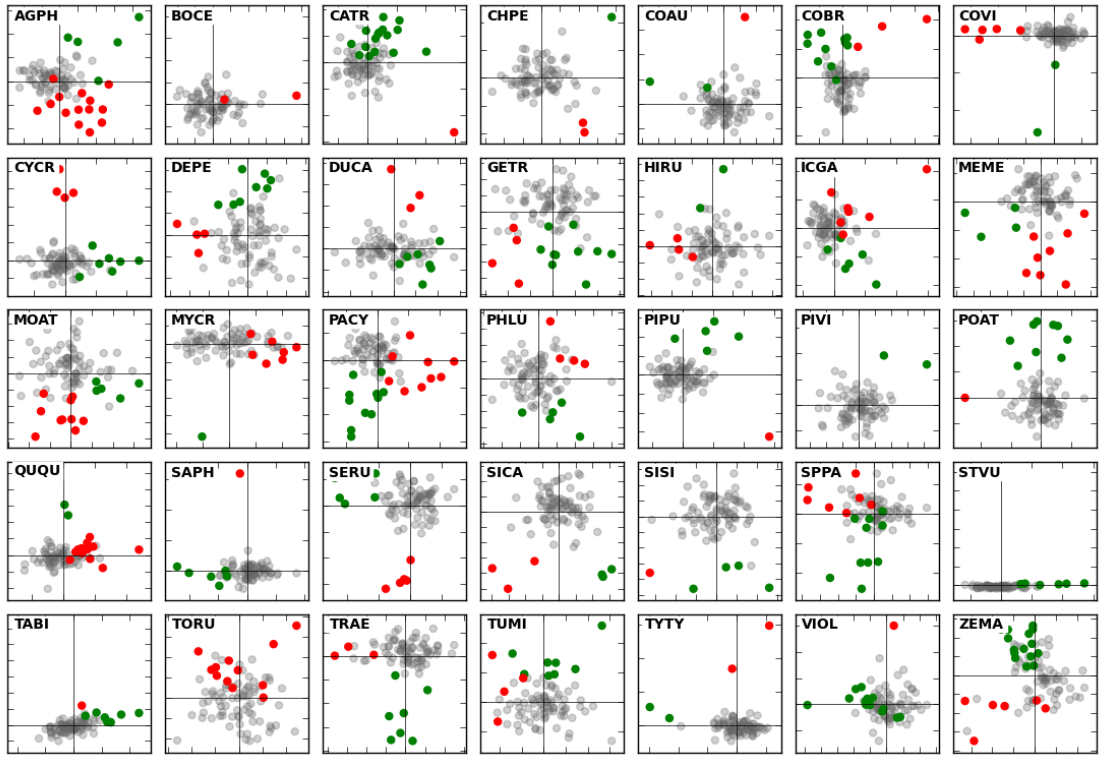
3 tricks for doing more with less

- **Multiple Plots**
  - Simple, easily interpretable subplots.
  - Can be beautiful but overwhelming.
- **Hybrid Plots**
  - A scatter plot of histograms.
  - Or a venn-diagram of histograms, etc.
- **Multiple Axes**
  - Plot two (or more) different things on one graph.

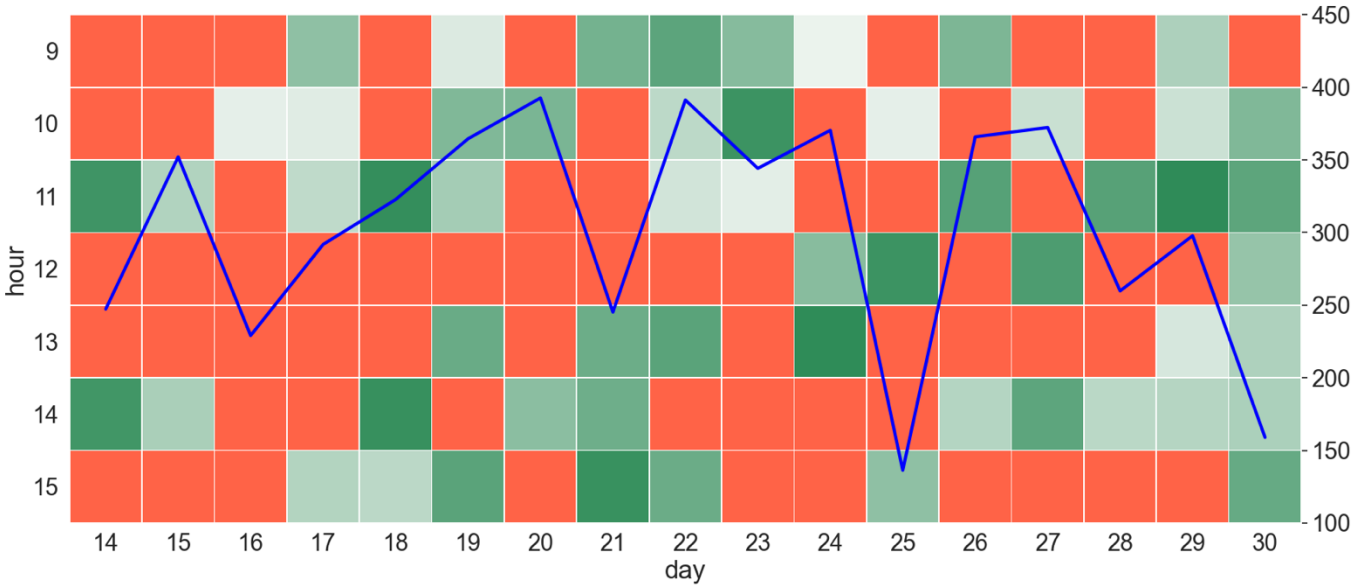
Multiple plots

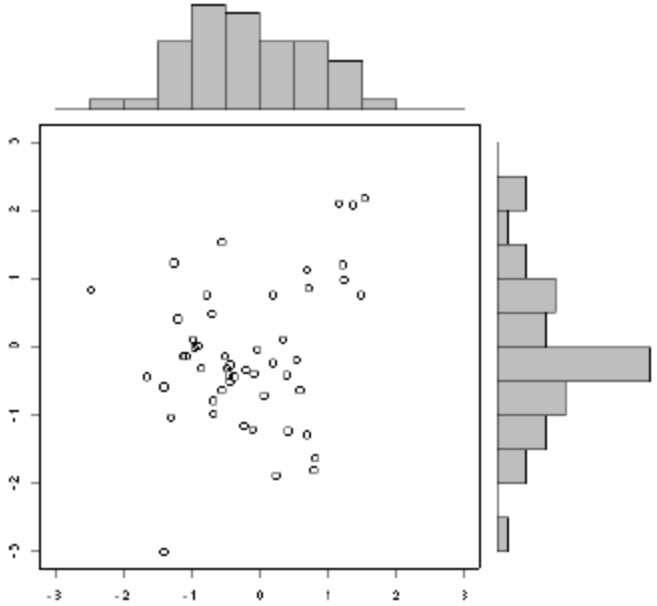


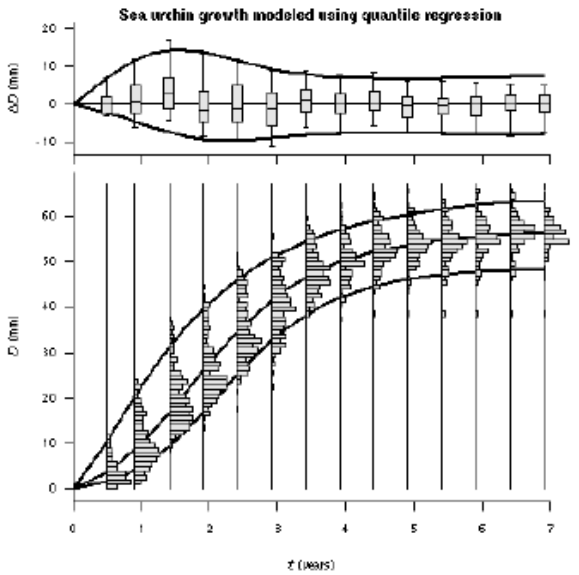




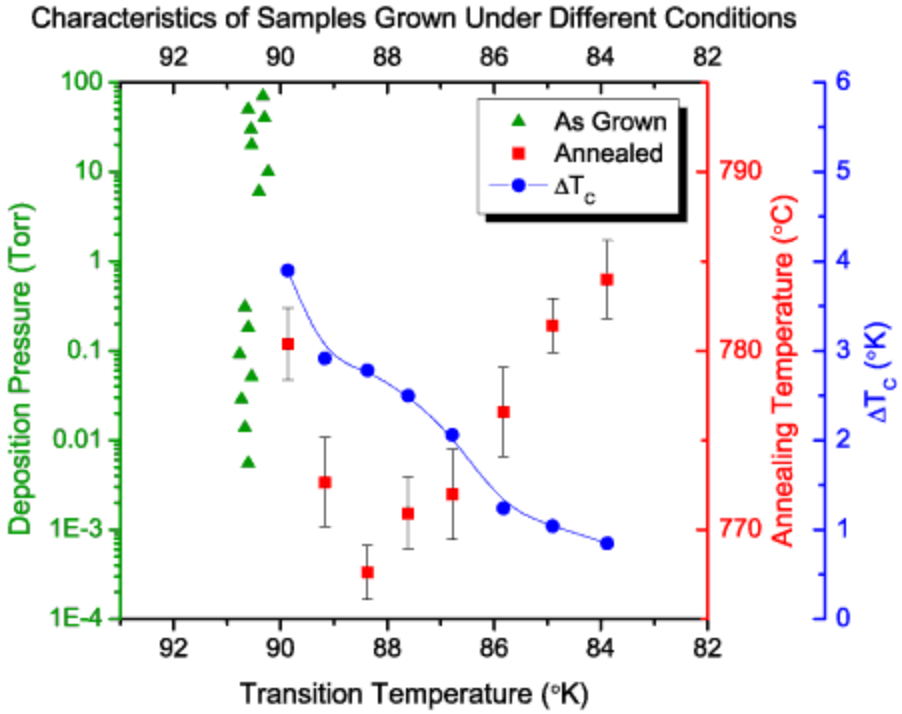
Hybrid plots

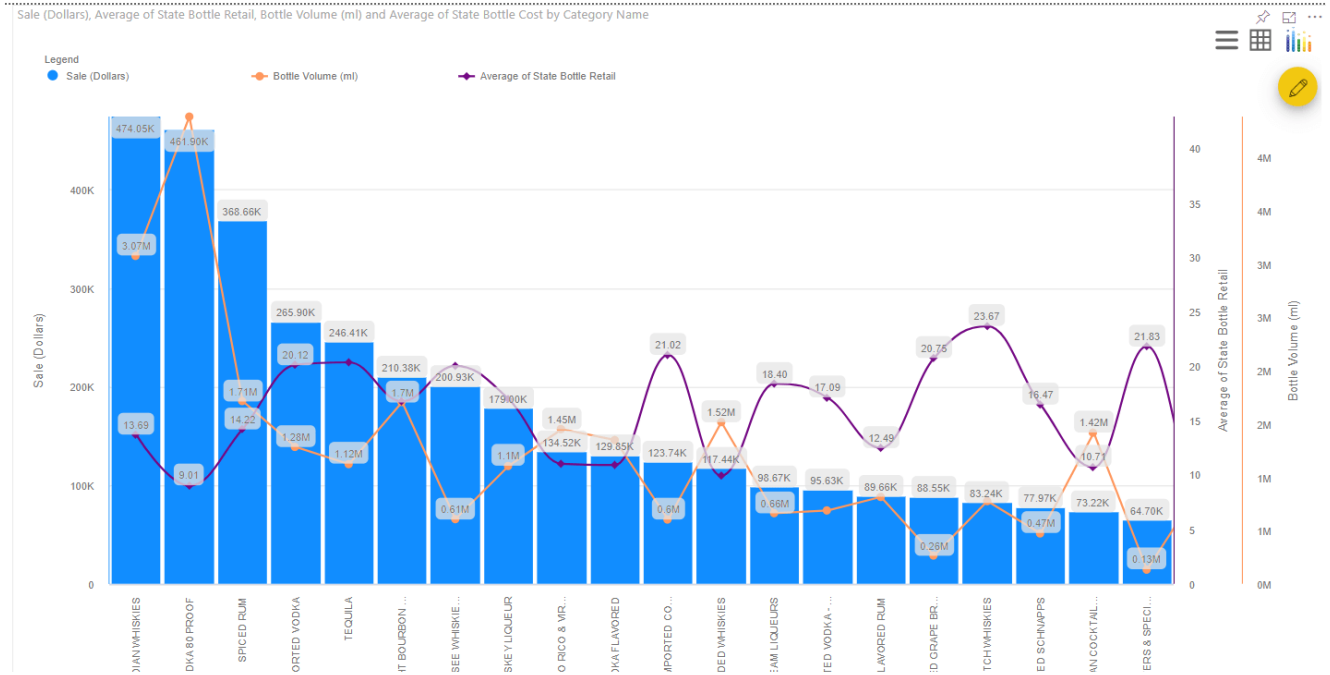


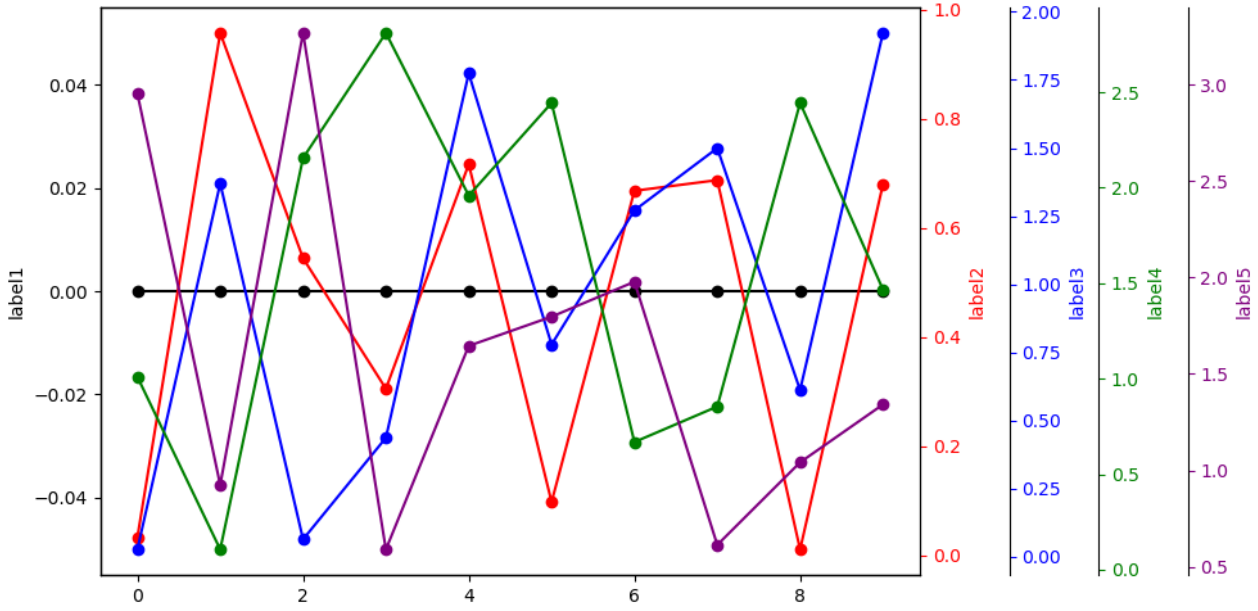




Multiple axes







2 tradeoffs

- **Informativeness vs. readability**
  - Too little information can conceal data.
  - But too much information can be overwhelming.
  - Possible solution: hierarchical organization.
- **Data-centric vs. viewer-centric**
  - Viewers are accustomed to certain types of visualization.
  - But novel visualization can be truer to data.

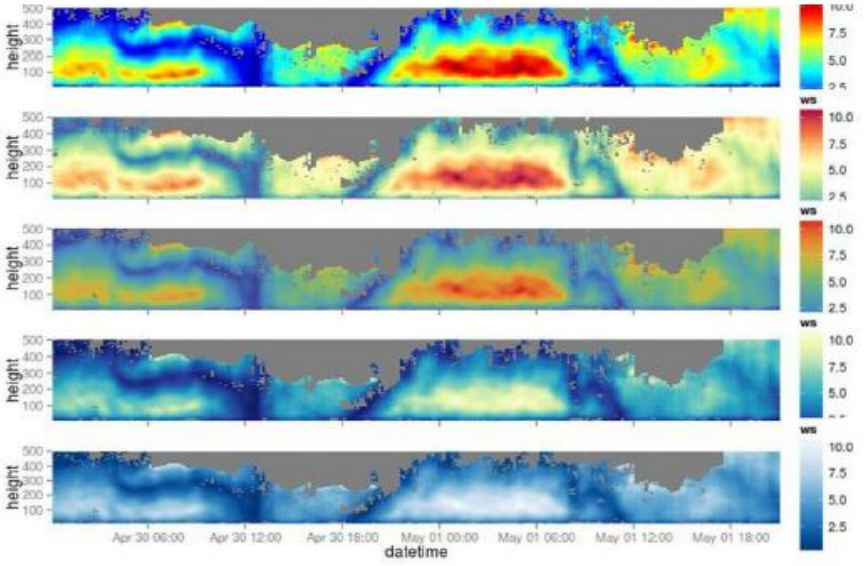
Data visualization traps (1)

- **Wrong plot types**

  Be careful with interpretations and keep in mind that the x-axis generally shows the independent variable while the dependent variable is on the y-axis.

- **Different color usages result in different impressions**

  Depending on the actual scheme, certain features are highlighted while others fade away. In addition, not all color pallets are b/w compatible.



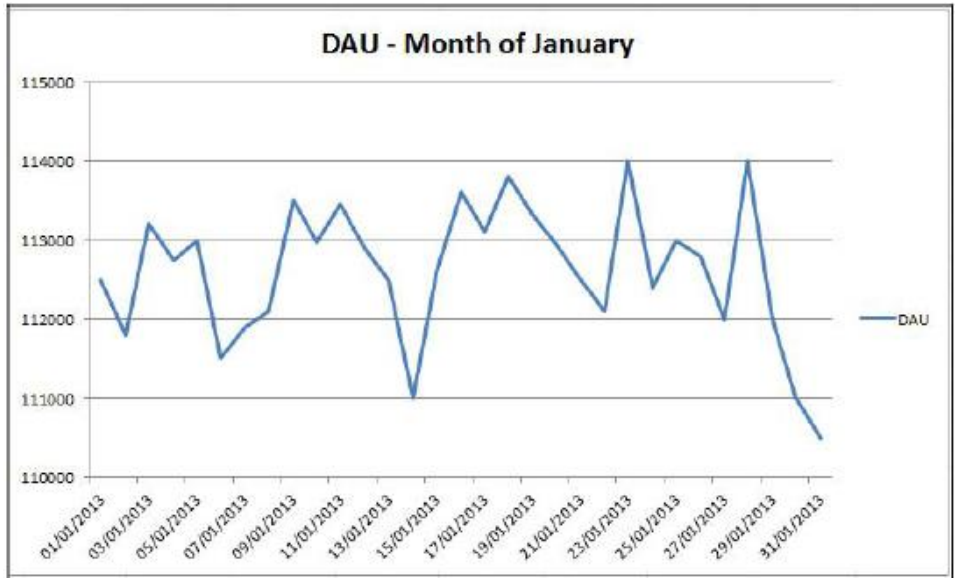
Data visualization traps (2)

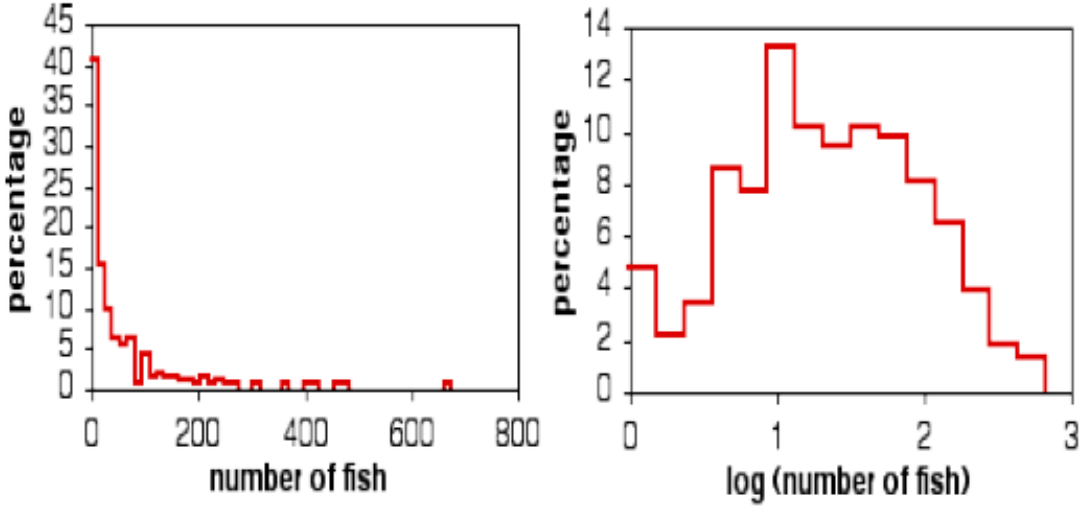
- **Exaggerated plot scale**

  - The chart displays the evolution of Daily Active Users (DAU) on a timeline. As we can see, the evolution seems to have lots of variations, and the curve also has a worrying decrease at the end of the month.

  - Now let's have a glimpse at the y axis. It definitely does not look so frightening.

- **Take log-scale to show full information**





Data visualization traps (3)

Common Errors

1. No title on graph
2. Source of data not given
3. No figure legend
4. Scales are interrupted
5. Scales are not labeled
6. Symbols are not in the same size or equally spaced on graph
7. Use of 3 dimensional objects to compare data (area/volume)
8. Scales do not start at zero
9. Numbers on axis are not equally spaced
10. Scale is selected to produce desired result

Visualization and dashboard software

Commonly used data visualization software and tools include:

- Spreadsheets
- Jupyter Notebook and Python libraries
- R-Studio and R-Shiny
- IBM Cognos Analytics
- Tableau
- Microsoft Power BI

Spreadsheets

- **Excel**
  - ❑Provides several chart types – bar charts, line charts, pie charts, pivot charts, scatter charts, trendlines, Gantt charts, Waterfall charts, and combination charts.
  - ❑Provides recommendations on visual representation.
  - ❑Can add chart title, change colors of elements, and add labels to data.
- **Google sheets**
  - ❑Suggest visualization best suited for your data set.
  - ❑Preferred over Excel for its collaboration features.

Jupyter Notebook and Python libraries (1)

- **Jupyter Notebook**

  An open-source web application that provides a great way to explore data and create visualizations.

- **Matplotlib**

  ❑Widely used Python data visualization library.

  ❑Provides different kinds of 2D and 3D plots and the flexibility to create plots in several different ways.

  ❑Helps create high-quality interactive graphs and plots with just a few lines of code.

Jupyter Notebook and Python libraries (2)

- **Bokeh**
  - ❑Provides interactive charts and plots.
  - ❑Delivers high-performance interactivity over large of streaming datasets.
  - ❑Offers flexibility for applying interaction, layouts, and different styling options to visualization.
  - ❑Can transform visualizations written in other Python libraries, such as Matplotlib, Seaborn, and ggplot.

Jupyter Notebook and Python libraries (3)

- **Plotly Dash**
  - ❑A Python framework for creating interactive web-based visualizations.
  - ❑Helps build highly interactive web interactive web applications using Python code.
  - ❑Does not require knowledge of HTML and JavaScript.
  - ❑Is easily maintainable, cross-platform, and mobile-ready.

R-Studio and R-Shiny

- **R-Studio**
  - ❑Can create basic visualization such as histograms, bar charts, line charts, box plot, and scatter plots.
  - ❑Advanced visualizations such as heat maps, mosaic maps, 3D graphs, and correlograms.
- **R-Shiny**
  - ❑Shiny is an R package that helps build interactive web apps that can be hosted as standalone apps on a webpage.
  - ❑You can also build dashboards using Shiny.

IBM Cognos Analytics

IBM Cognos Analytics – an end-to-end analytics solution

Some of the visualization features provided by Cognos include:

- Importing custom visualizations.
- A forecasting feature that provides time-series data modeling and forecasts.
- Recommendation for visualizations based on your data.
- Conditional formatting which allows you to see the distribution of your data and highlight exceptional data points.

Cognos is known for its superior visualizations and overlaying data on the physical world using its geospatial capabilities.

Tableau

Tableau – a software company that produces interactive data visualization products.

Tableau products allow you to:

- Create interactive graphs and charts in the form of dashboards and worksheets, with drag and drop gestures.
- Publish results in the form of stories.
- Import R and Python scripts.
- Compatible with Excel files, Text files, Relational databases, Cloud database sources such as Google Analytics and Amazon Redshift.

Power BI

Power BI – a cloud based business analytics service from Microsoft that enables you to create reports and dashboards.

- A powerful and flexible tool known for its speed and efficiency.
- Has a drag and drop interface.
- Is compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories.
- Provides the ability to collaborate and share dashboards and reports securely.

香港城市大學
City University of Hong Kong

Thanks for your attention!

Appendix

1.http://vis.csail.mit.edu/classes/6.859/

2. https://www.coursera.org/learn/introduction-to-data-analytics/lecture/z4tlk/introduction-to-visualization-and-dashboarding-software