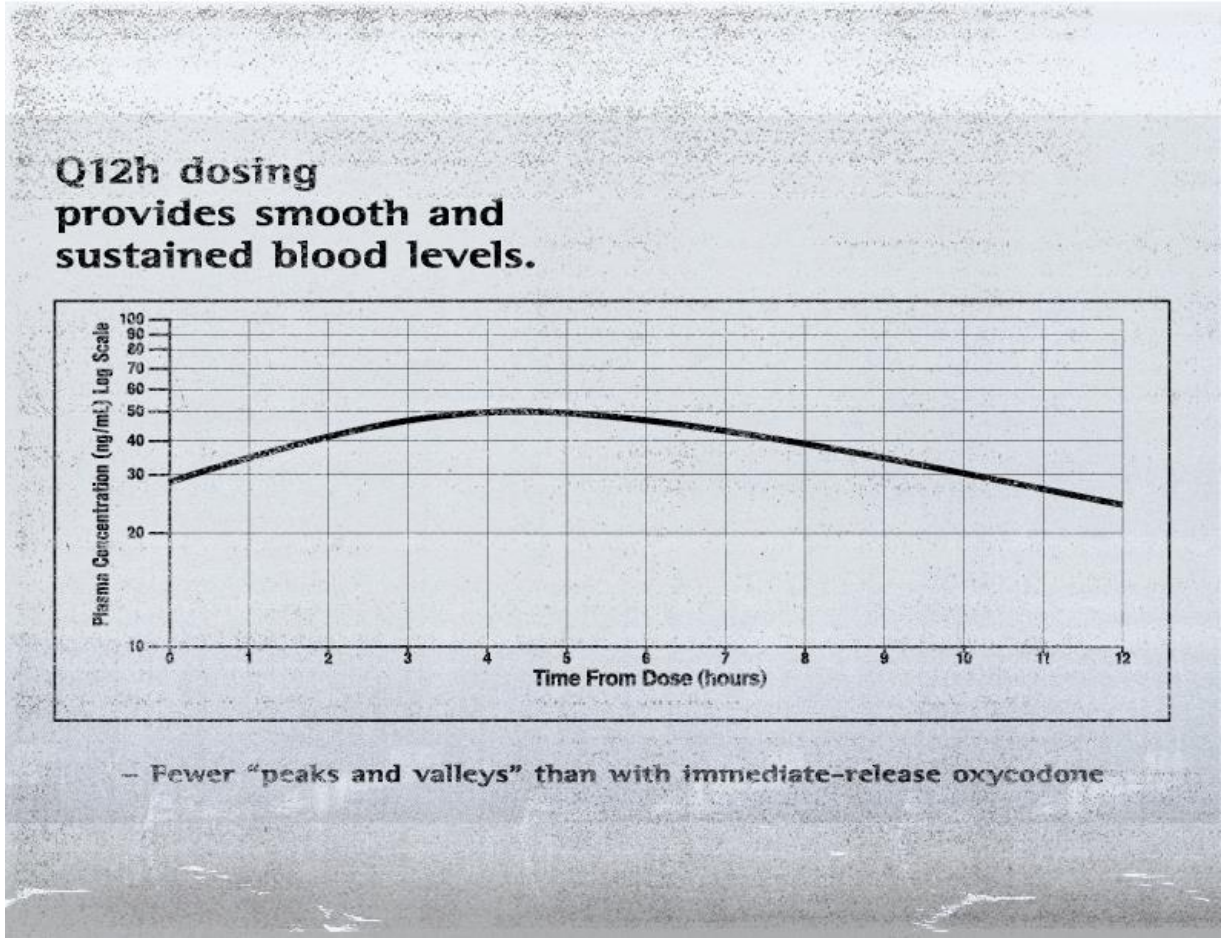
香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Data Visualization

Instructor: Dr. Xiaoyi Fu

Anything wrong in this chart?



Who am I

Xiaoyi Fu, PhD

IT Professional – Computer Science Researcher

- Lecturer
  - Department of Journalism- Hong Kong Baptist University

新聞系 DEPARTMENT OF JOURNALISM

香港浸會大學 HONG KONG BAPTIST UNIVERSITY

- Background
  - Bachelor of Engineering in Computer Science & Technology, Zhejiang University
  - PhD in Computer Science, HKBU

Outline

- Introduction to Data Visualization
- Dos and Don’ts
- Hands-on Practice

A perfect example of data analysis and visualization

- 200 years in 4 minutes https://www.youtube.com/watch?v=Z8t4k0Q8e8Y

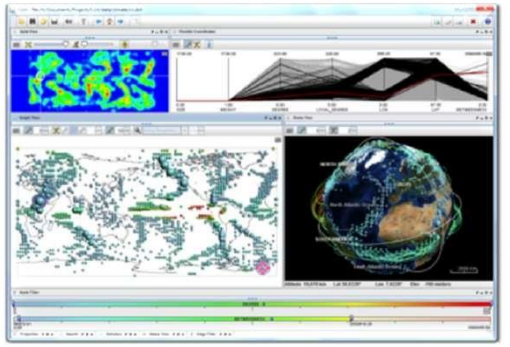
What is Data Visualization

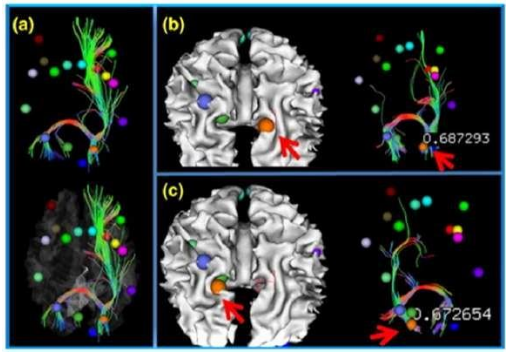
- In 1987
  - The National Science Foundation (of the U.S.) started “Visualization in scientific computing” as a new discipline, and a panel of the ACM coined the term “scientific visualization”
  - Scientific visualization, briefly defined: The use of computer graphics for the analysis and presentation of computed or measured scientific data.
- Oxford Engl. Dict., 1989
  - To form a mental vision, image, or picture of (something not visible or present to the sight, or of an abstraction); to make visible to the mind or imagination
- Visualization transforms data into images that effectively and accurately represent information about the data.
  - Schroeder et al. The Visualization Toolkit, 2nd ed. 1998

Tool to enable a User insight into Data

What Does Visualization Do?

- Three types of goals for visualization
  - … to explore
    - Nothing is known,
    - For data exploration
  - … to analyze
    - There are hypotheses,
    - For Verification or Falsification
  - … to present
    - "everything" known about the data,
    - For Communication of Results





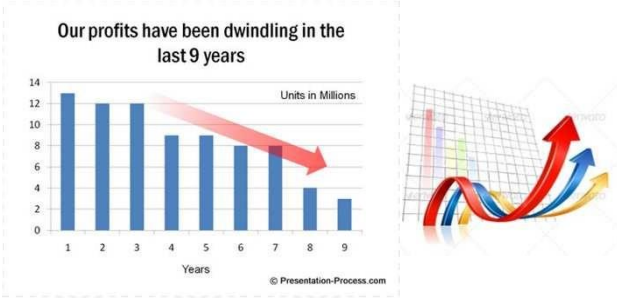


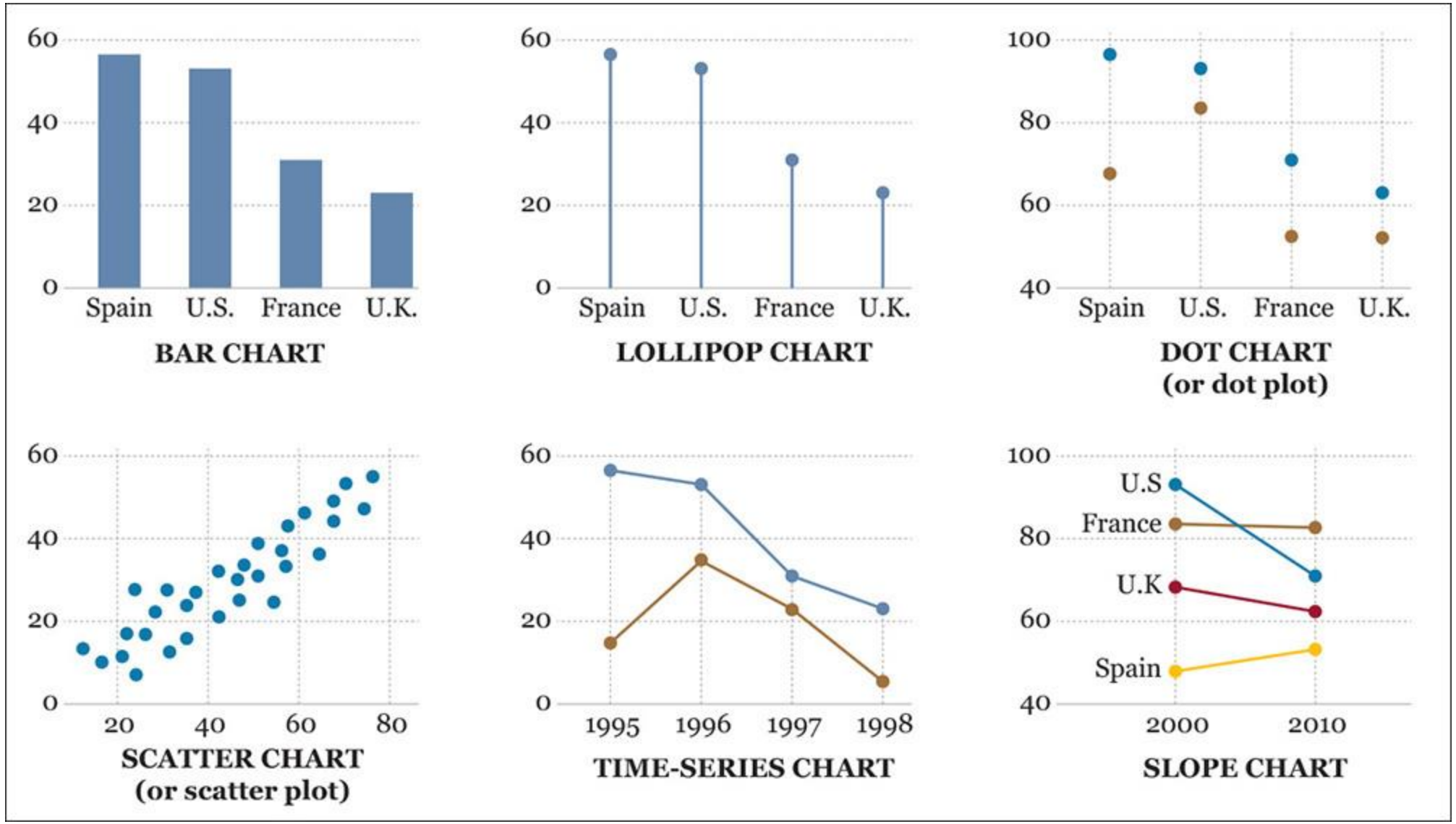
Image source: Google images

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

—John W. Tukey, *Exploratory Data Analysis*

Chart

- A chart is a display in which data are encoded with symbols that have different shapes, colors, or proportions.
  - In many cases, these symbols are placed within a Cartesian coordinate system.
  - The word “plot” is a synonym of “chart” in this course, as it’s commonly used to refer to a few specific charts in the professional literature (“scatter plot” sounds more familiar than “scatter chart”).


60
40
20
0
Spain
U.S.
France
U.K.
BAR CHART
LOLLIPOP CHART
100
80
60
40
DOT CHART
(or dot plot)
SCATTER CHART
(or scatter plot)
1995
1996
1997
1998
TIME-SERIES CHART
U.S
France
U.K
Spain
2000
2010
SLOPE CHART

Data Types

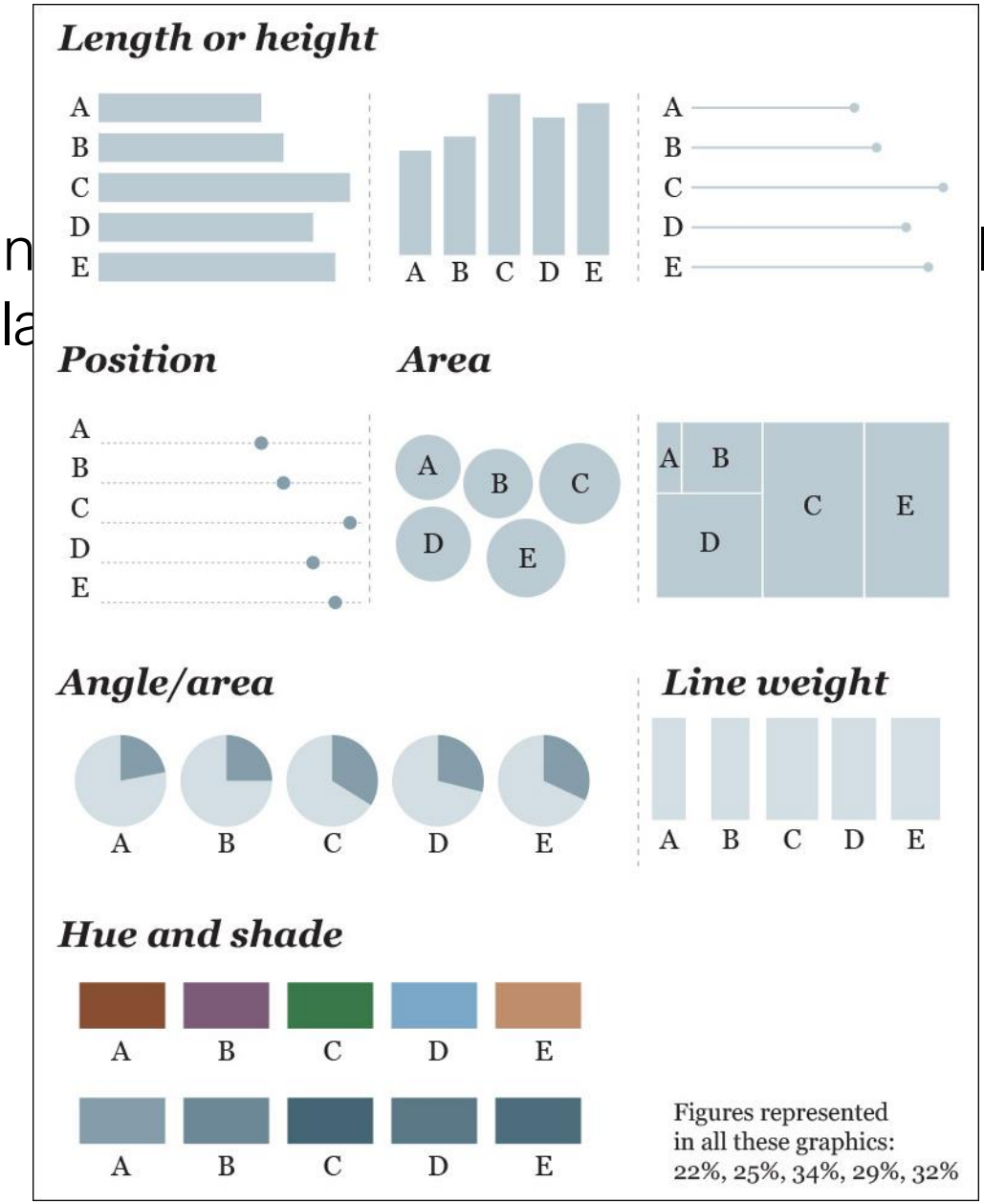
	Discrete (no between values)	Continuous (values between)
Ordered (values are comparable)	**Ordinal,** e.g. size: S,M,L,XL,… **Quantitative,** e.g. counts: 1,2,3,…	**Fields,** e.g. altitude, temperature
Unordered (values not comparable)	**Nominal,** e.g. shape: ☐○△ **Categories,** e.g. nationality	**Cyclic values,** e.g. directions, hues

Mapping

- The mapping part consists of choosing properties that will let readers accomplish a particular goal (“comparing accurately”) without being forced to read all numbers.

Mapping

- The mapping part con let readers accomplish a particula out being forced to read all numbers.



Visual encoding

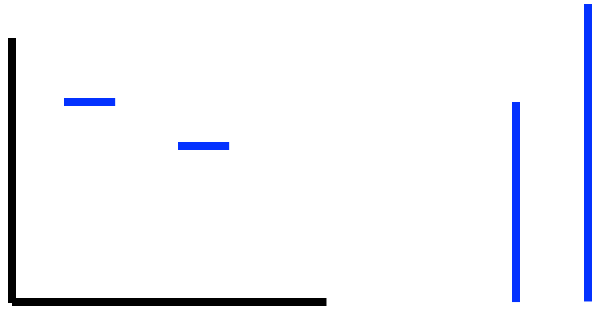
Mapping Quantitative Values

- Position
- Length
- Angle/Slope
- Area
- Volume
- Color/Density

Perceptual accuracy

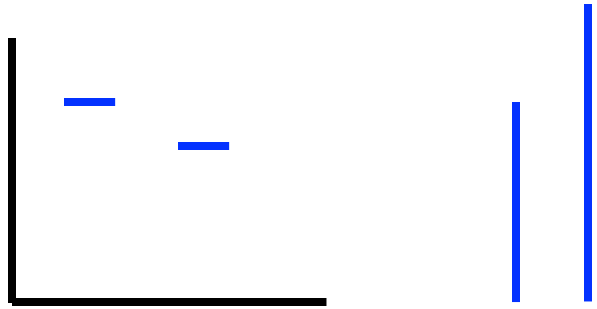
CLEVELAND, W. S., AND MCGILL, R. Graphical perception: Theory, experimentation and application to the development of graphical methods. Journal of the American Statistical Association, 79(387) 1984

Mapping Quantitative Values

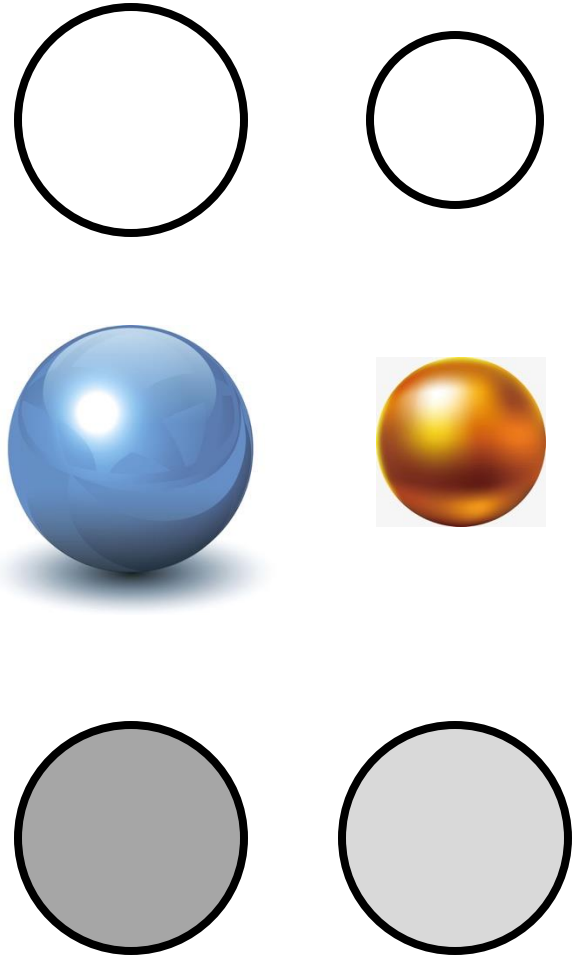


- Position
- Length
- Angle/Slope

Mapping Quantitative Values



- Position
- Length
- Angle/Slope
- Area
- Volume
- Color/Density



Mapping Quantitative Values

- Position
- Length
- Angle/Slope
- Area
- Volume
- Color/Density

1-D

2-D

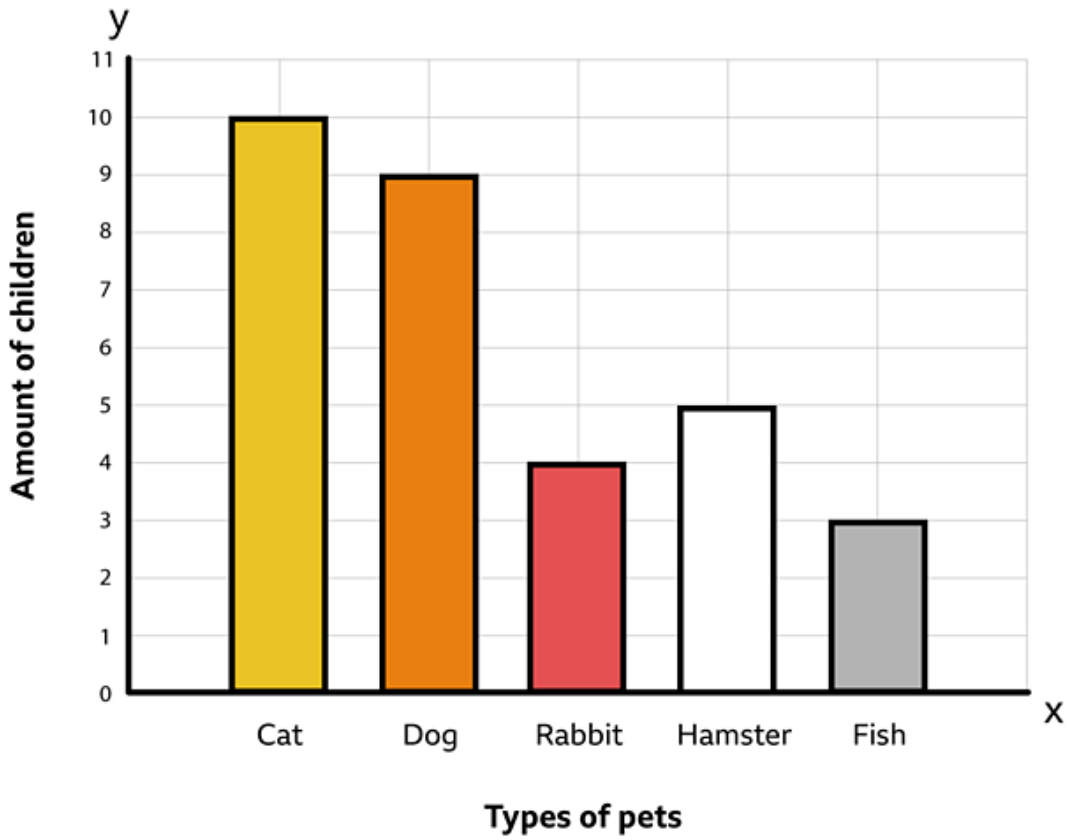
3-D

Independent/dependent variable

- Research Question: Does the level of education affect income level?
- Independent Variable: Level of education (e.g., high school diploma, bachelor's degree, master's degree, etc.).
  - This is the variable that you think might influence the outcome, and you can categorize or quantify the different levels of education that participants have completed.
- Dependent Variable: Income level (e.g., annual income).
  - This is what you measure to see if there is an effect from the different levels of education. You're interested in whether changes in education level lead to changes in income.

Bar Chart

Quantitative dependent variable



Discrete/nominal independent variable

Benefits from both position (top of bar) and length (size of bar)

https://www.bbc.co.uk/bitesize/topics/zqgrd2p/articles/z9kbp4j

Line Chart

Quantitative continuous dependent variable

Quantitative continuous independent variable

Benefits from position but not length

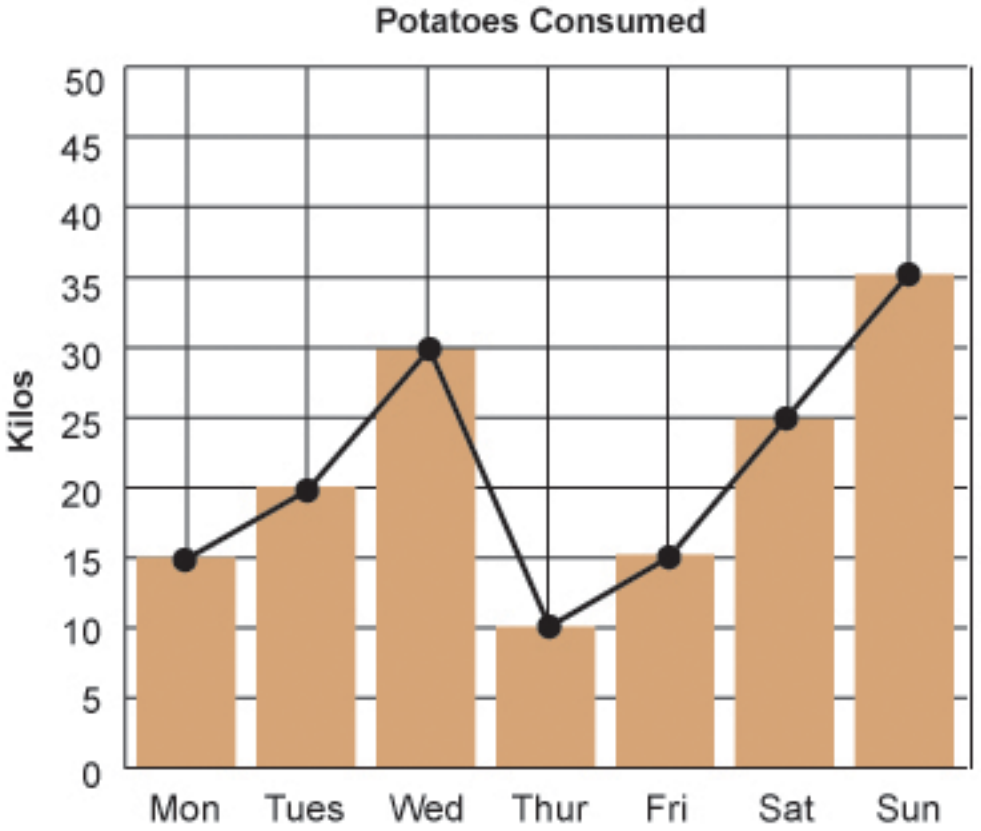
Line Chart

Quantitative continuous dependent variable

Benefits from position but not length

Quantitative continuous independent variable

Line Chart



Quantitative continuous dependent variable

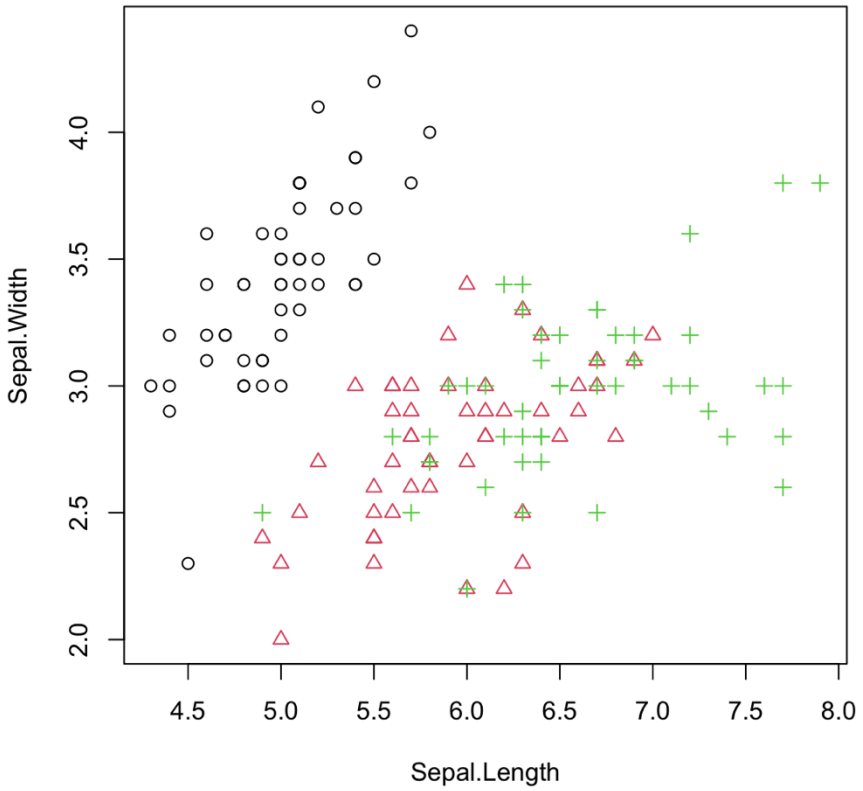
Quantitative continuous independent variable

Benefits from position but not length

https://github.com/naver/billboard.js/issues/847

Scatter Plot

Quantitative independent variable

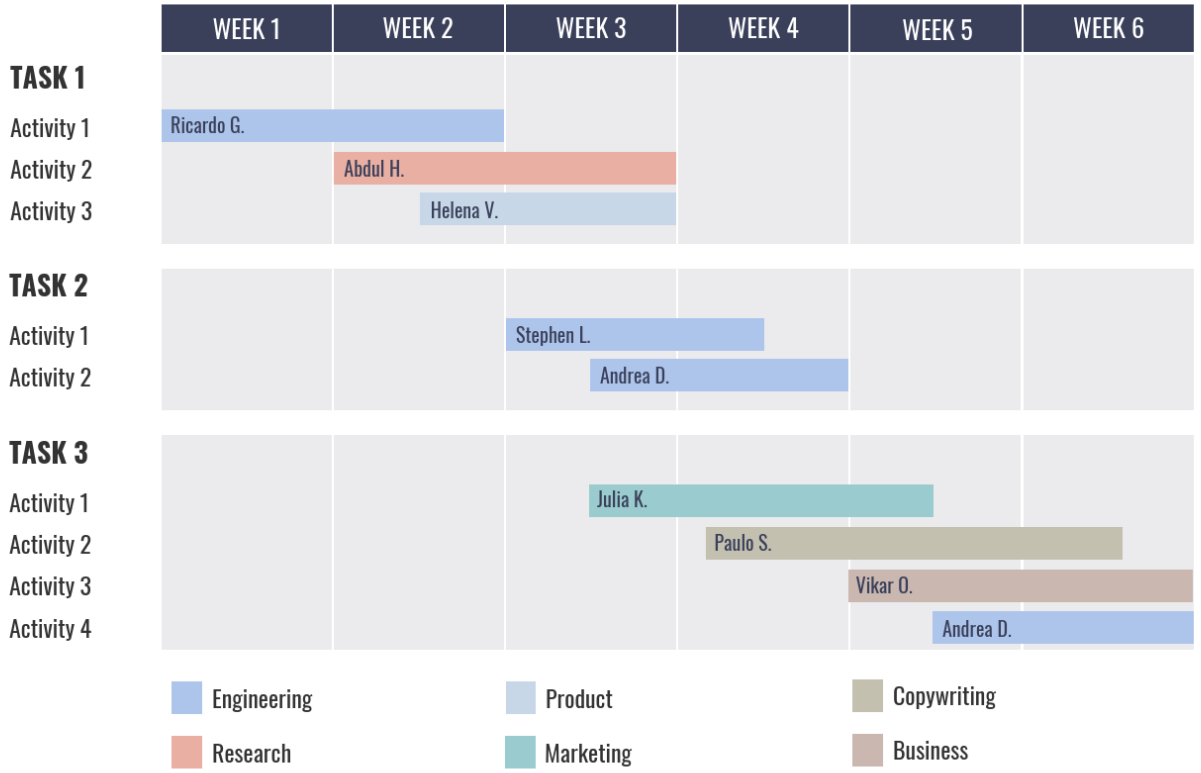


Quantitative independent variable

Relies mostly on position, but clusters also yield density

Gantt Chart

Discrete/nominal independent variable



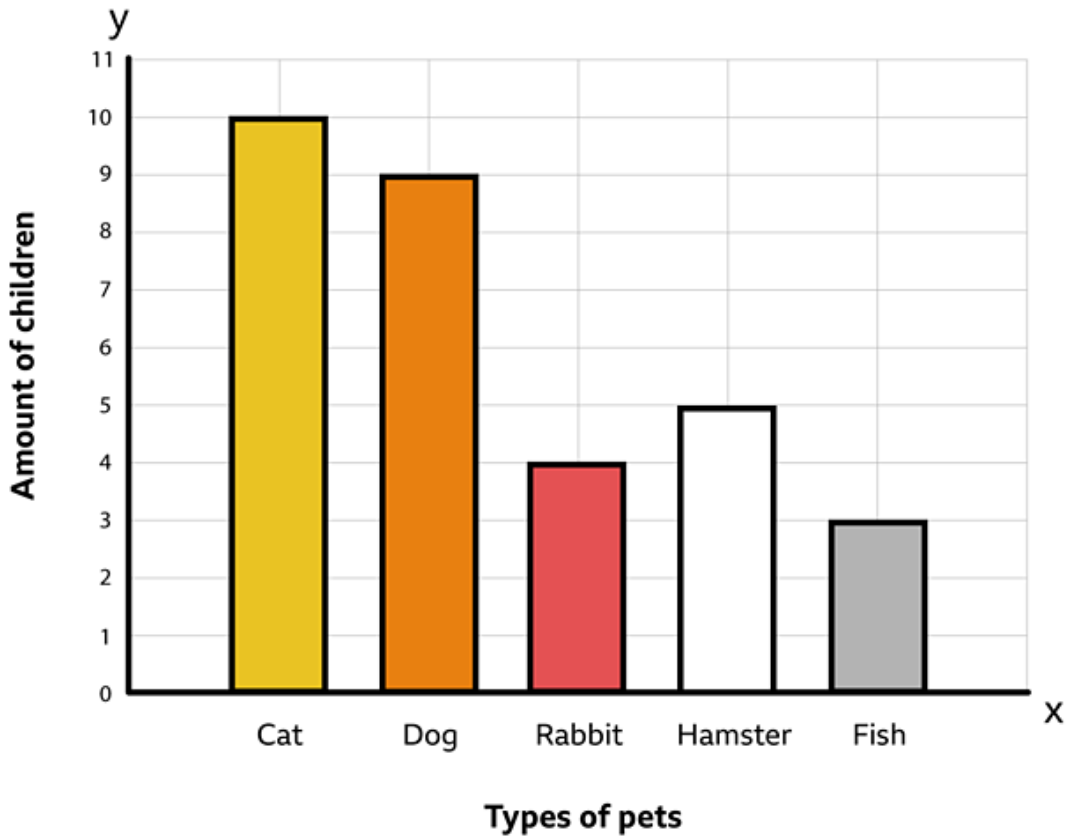
Benefits from both position and length

Quantitative independent variable

https://venngage.com/blog/gantt-chart-example/

Bar Chart

Quantitative dependent variable



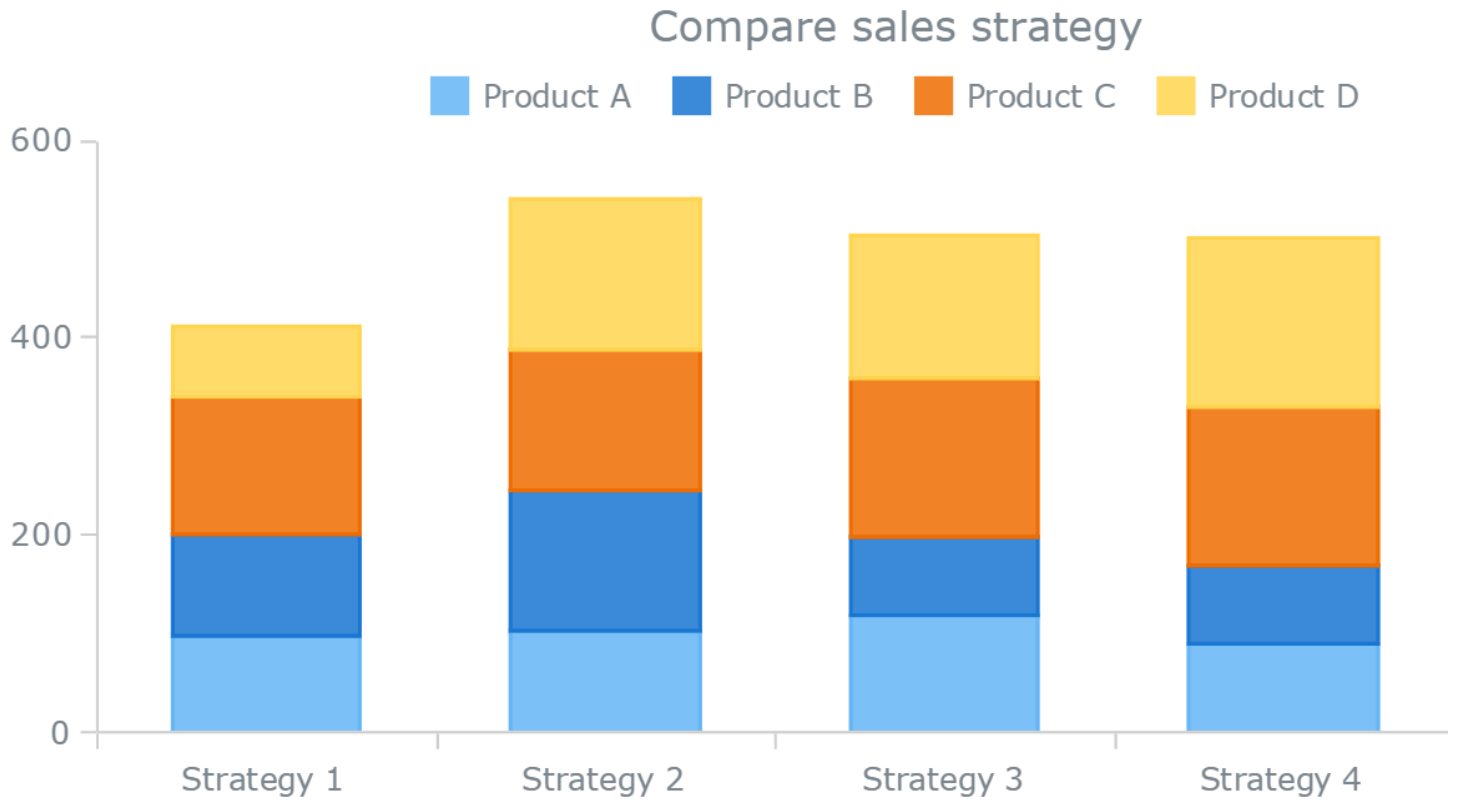
Discrete/nominal independent variable

Benefits from both position (top of bar) and length (size of bar)

- For a bar chart, we may want to look at more than just a single quantitative dependent variable measured across an independent variable dimension.

Stacked Bar Chart

Multiple (accumulating) dep. variables



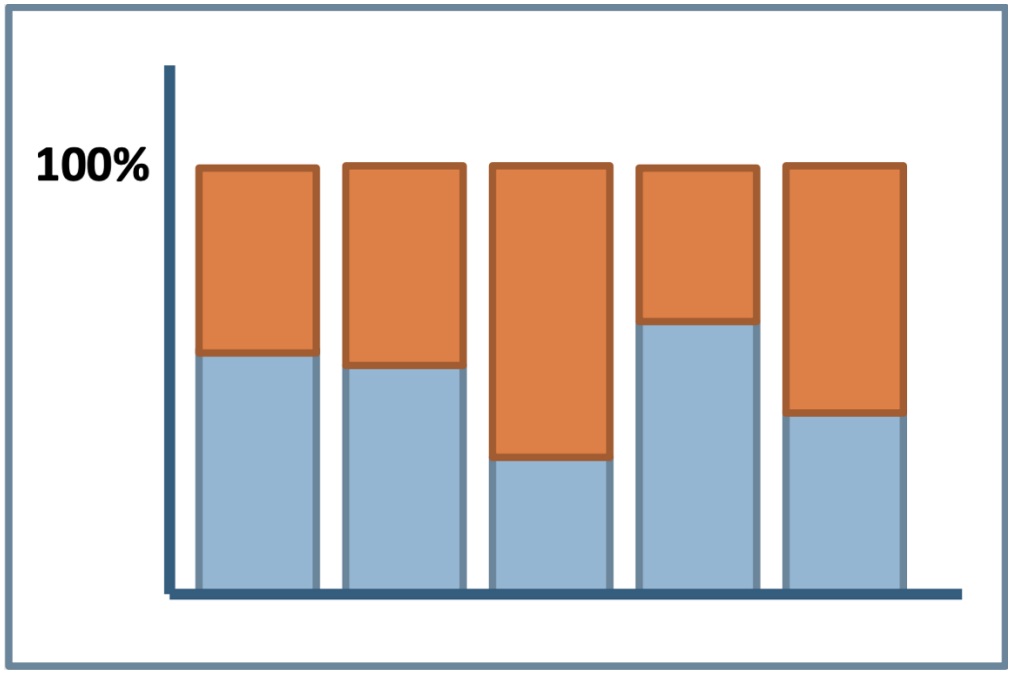
as more bars are added, sums will vary less

independent variable

https://github.com/kalkih/mini-graph-card/issues/79

Relative Stacked Bar Chart

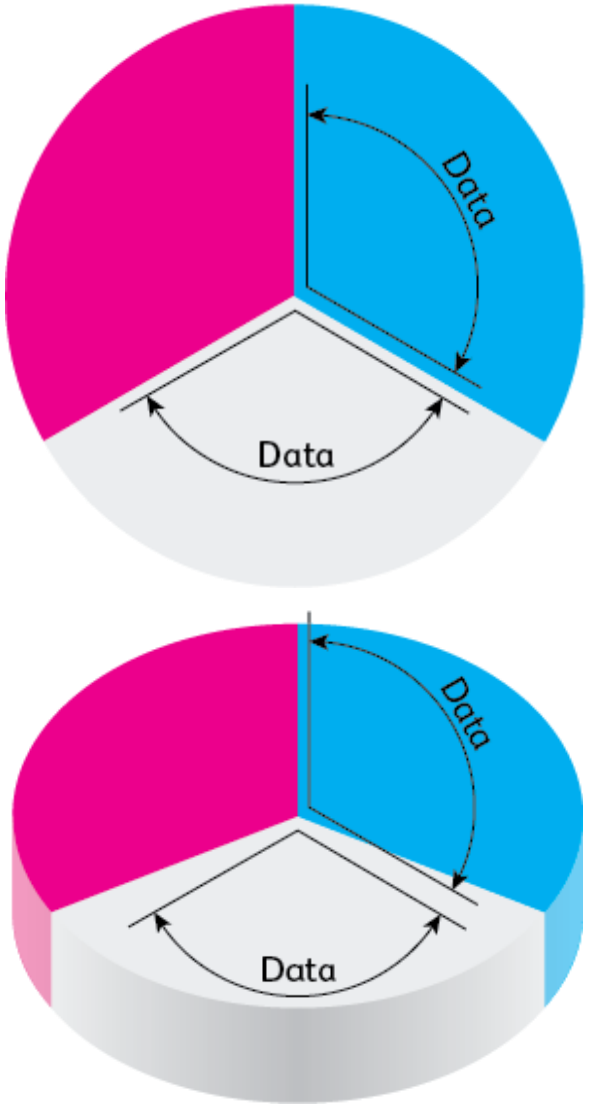
Two quantitative dependent variables



independent variable

Pie Chart

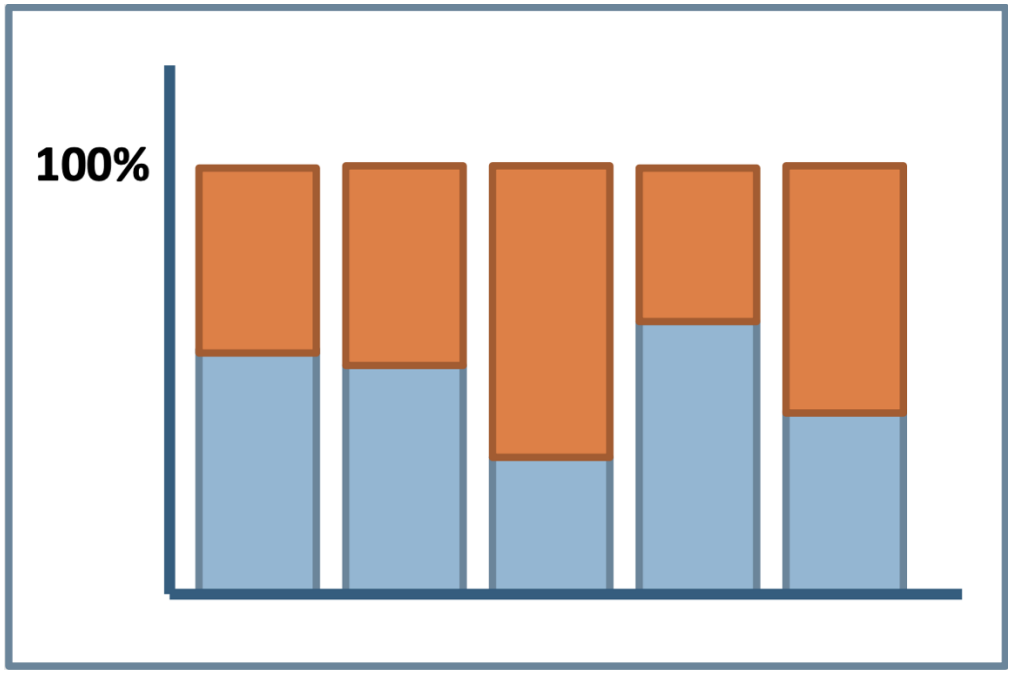
- Used to indicate relative portions of a quantitative dependent variable of a single dimension
- Maps percentage of total to angle of wedge arc
- Perspective (both distortion and foreshortening) confounds perception of angle
- NOT to use three dimensions for pie charts



https://rockcontent.com/blog/2ds-company-3ds-a-crowd/

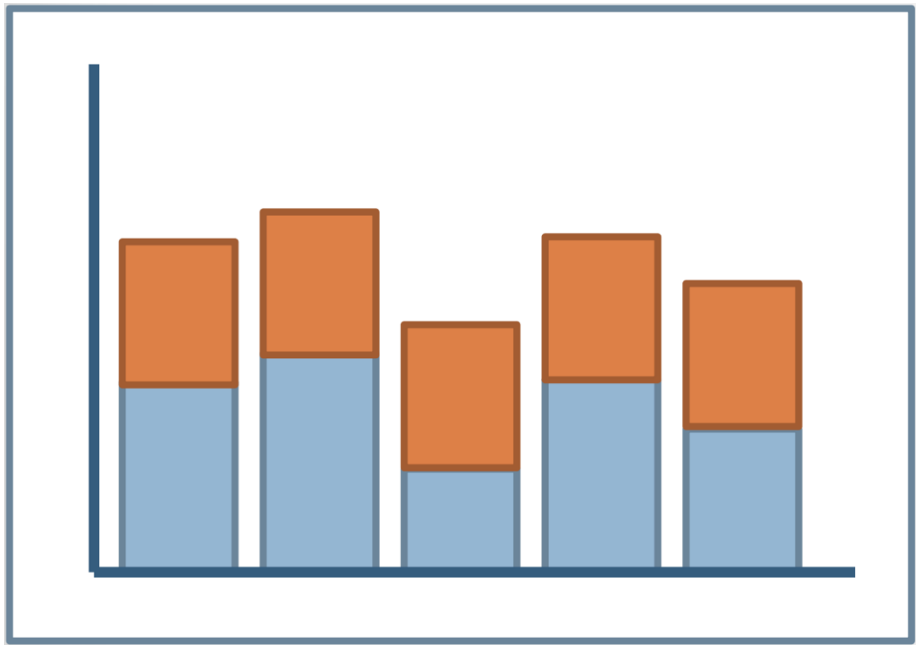
Angle

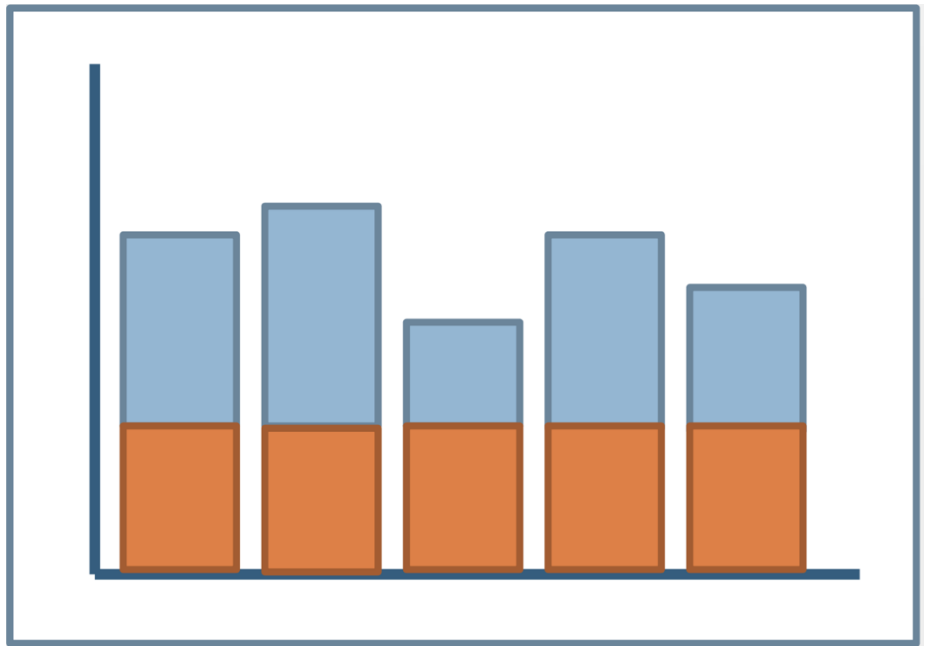
Two quantitative dependent variables



independent variable

Stacking Order Matters

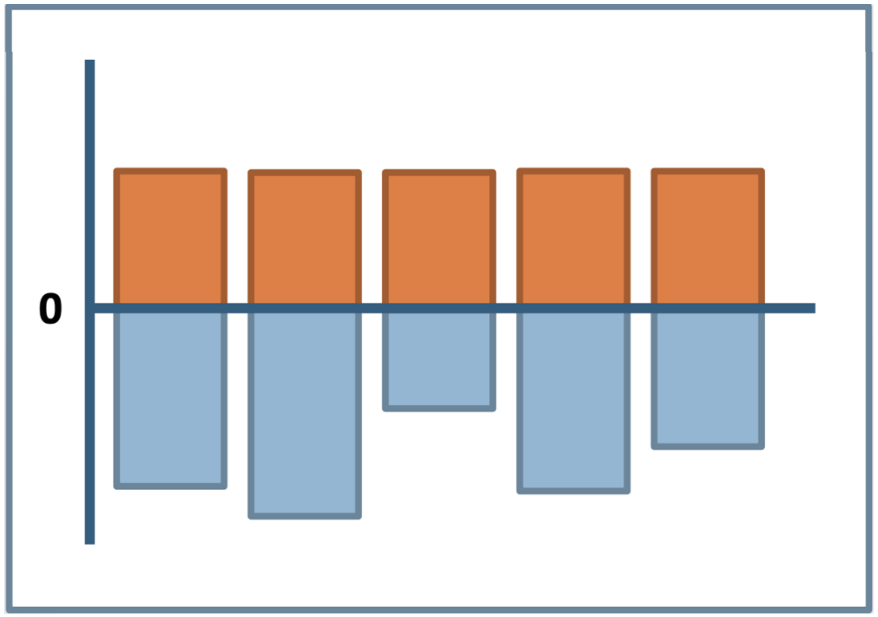




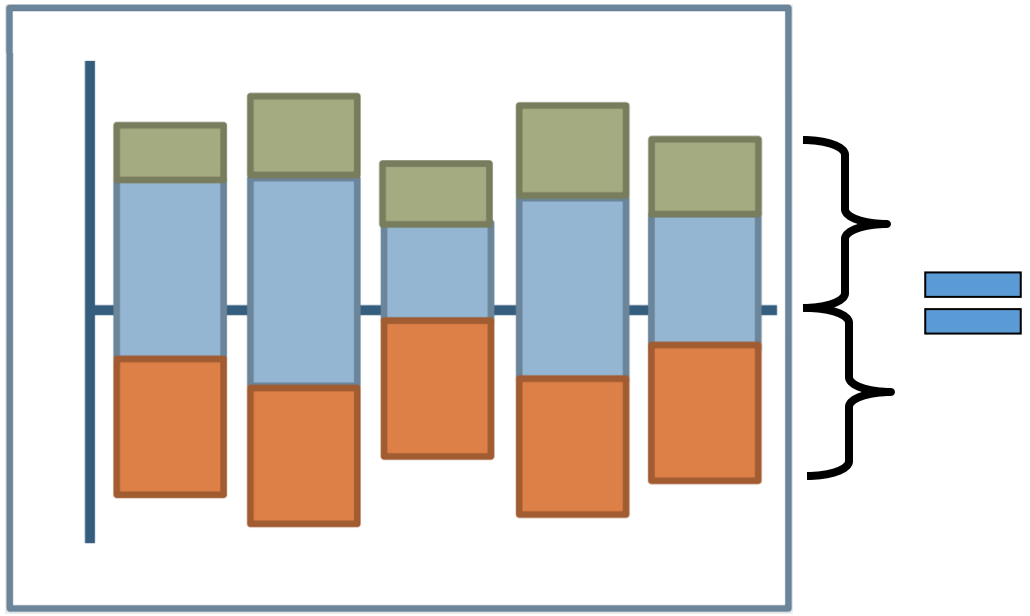
position > length

- Variance of lower stack elements influences perception of upper stack elements

Diverging Stacked Bar Charts

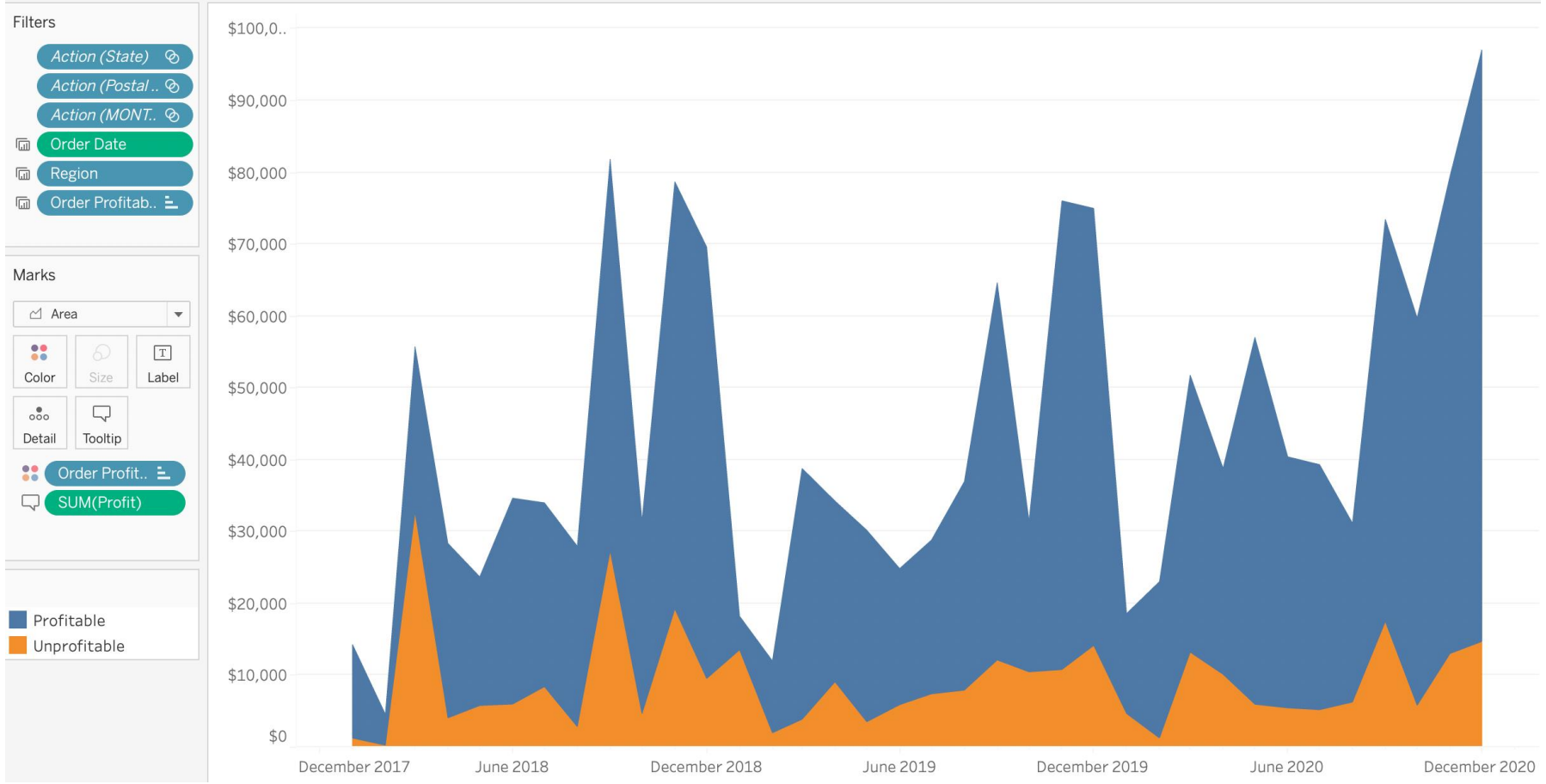


- Benefits from pos. & length
- Only works for two variables
- Negative connotation for lower bars



- Only indicates length
- Works for many variables
- Bar trends can still be obscured by neighboring bar variance

Stacked Line Graphs



- Appropriate for continuous data over a continuous independent variable
- Can smooth regions using curves instead of line segments

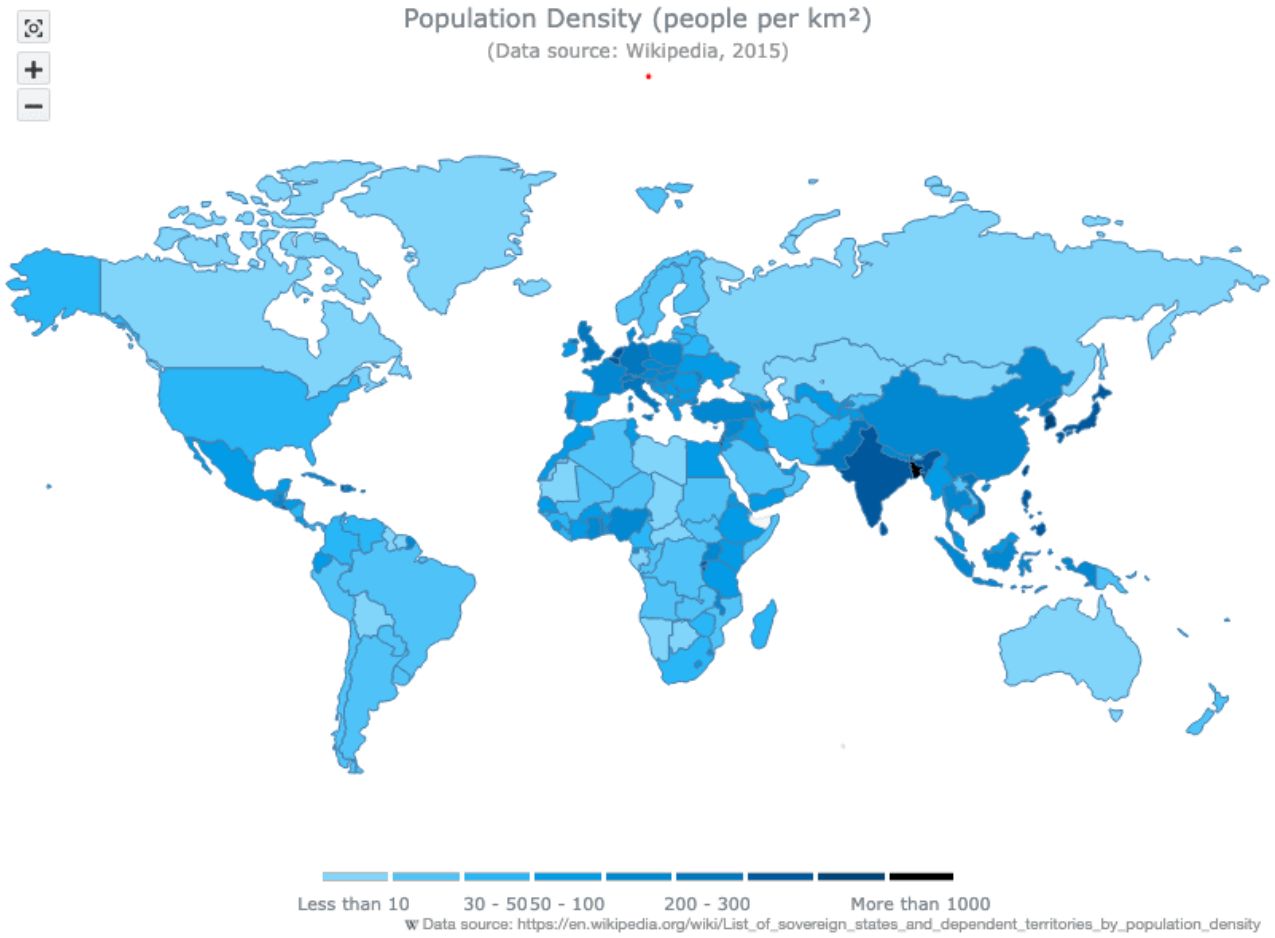
Map

- Plotting data geographically can help you understand and show your data findings in new and insightful ways.
- When should you use a map?
  - Maps are a good way to represent your data if some of the variables relate to geographical areas.

Choropleth

- Choropleth maps convey information by shading regions in different colors.
  - The hue can be determined by a category (which political party has won a local election) or by a number (percentage of arable land per state).
- Downside:
  - larger areas on the map will bring more attention to themselves and smaller areas will get lost.

Choropleth



World Population Density

Ordinal

Position

Density

Saturation

Hue

Texture

Connection

Containment

Length

Angle

Slope

Area

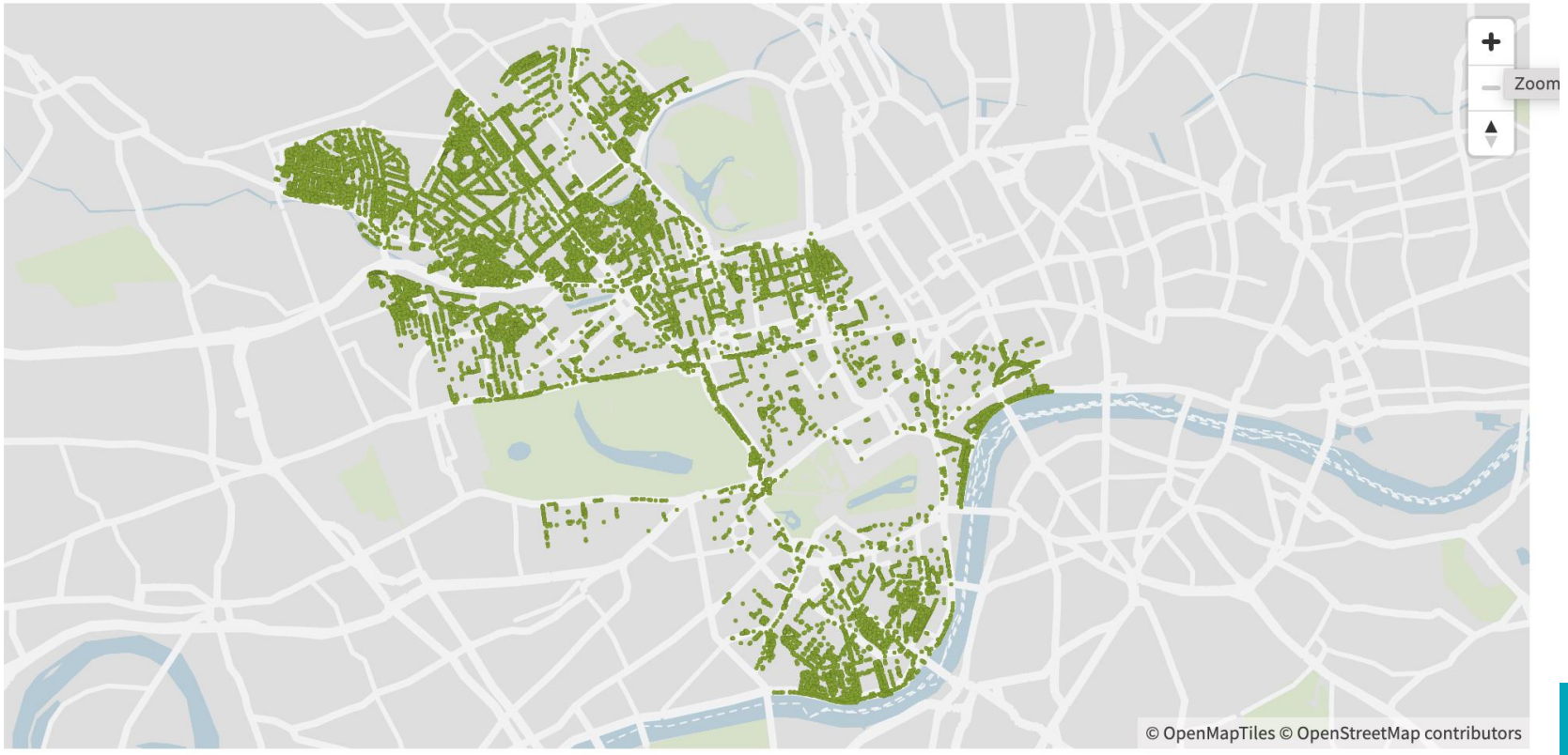
Volume

Point Map

- Point maps show information by positioning symbols on top of a specific space.
- The ma ements.

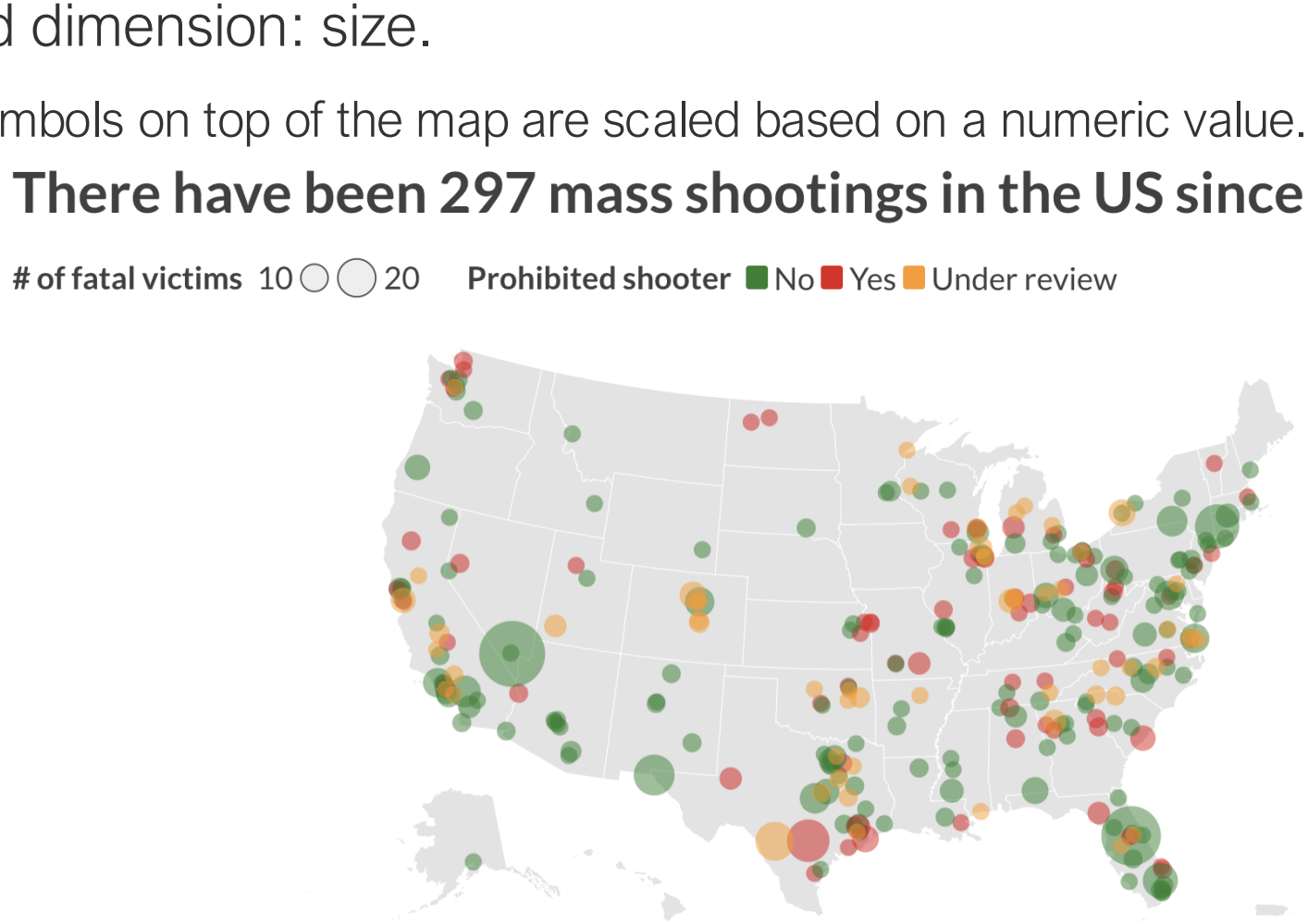
Here's every tree in the City of Westminster

There are over 19,000 trees in the London borough



Proportional symbol map

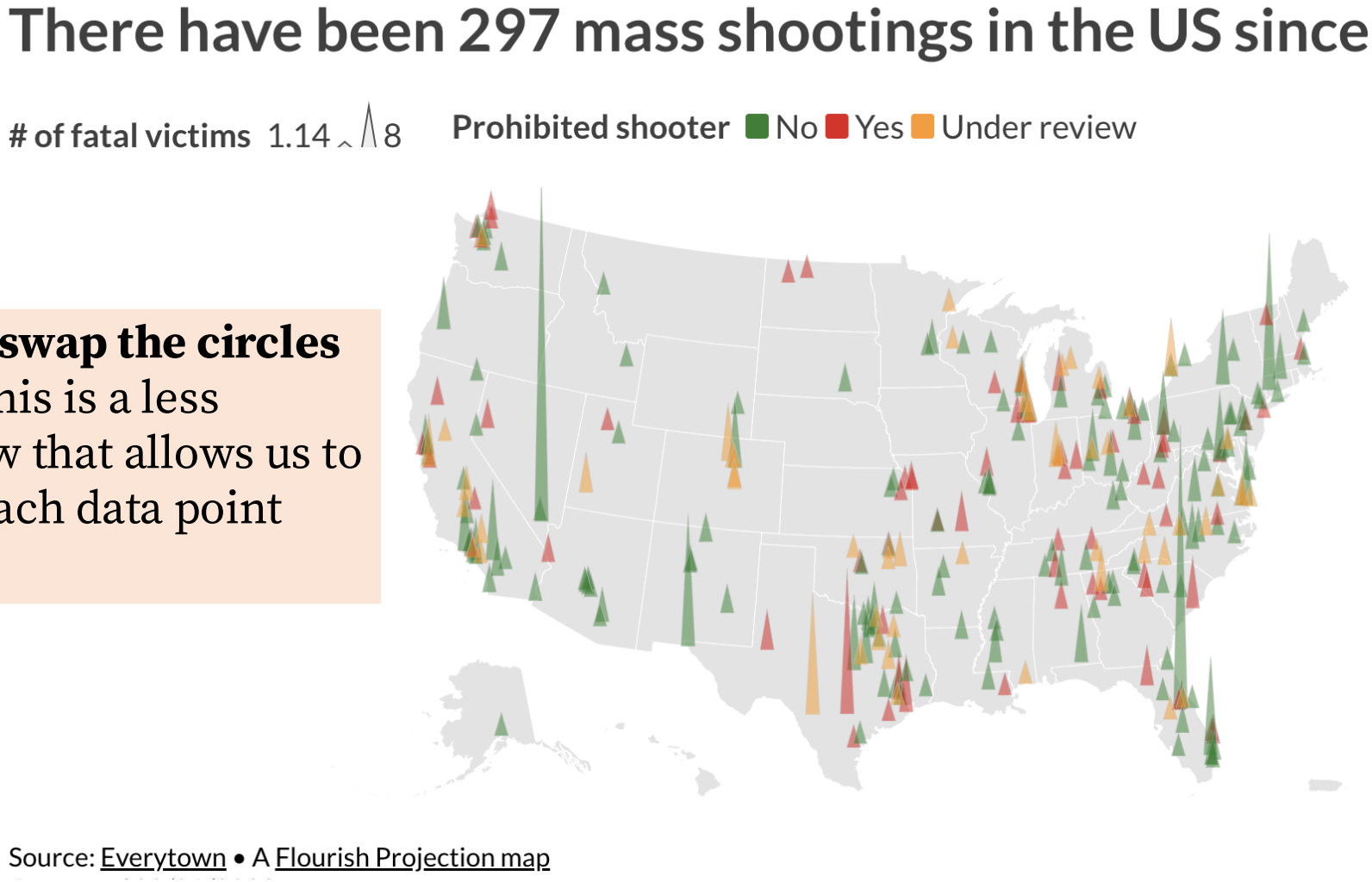
- Proportional symbol map builds on all the principles of the point map but adds a second dimension: size.
  - the symbols on top of the map are scaled based on a numeric value.



Source: Everytown • A Flourish Projection map

Proportional symbol map

You can also **swap the circles for spikes**. This is a less cluttered view that allows us to distinguish each data point more easily.

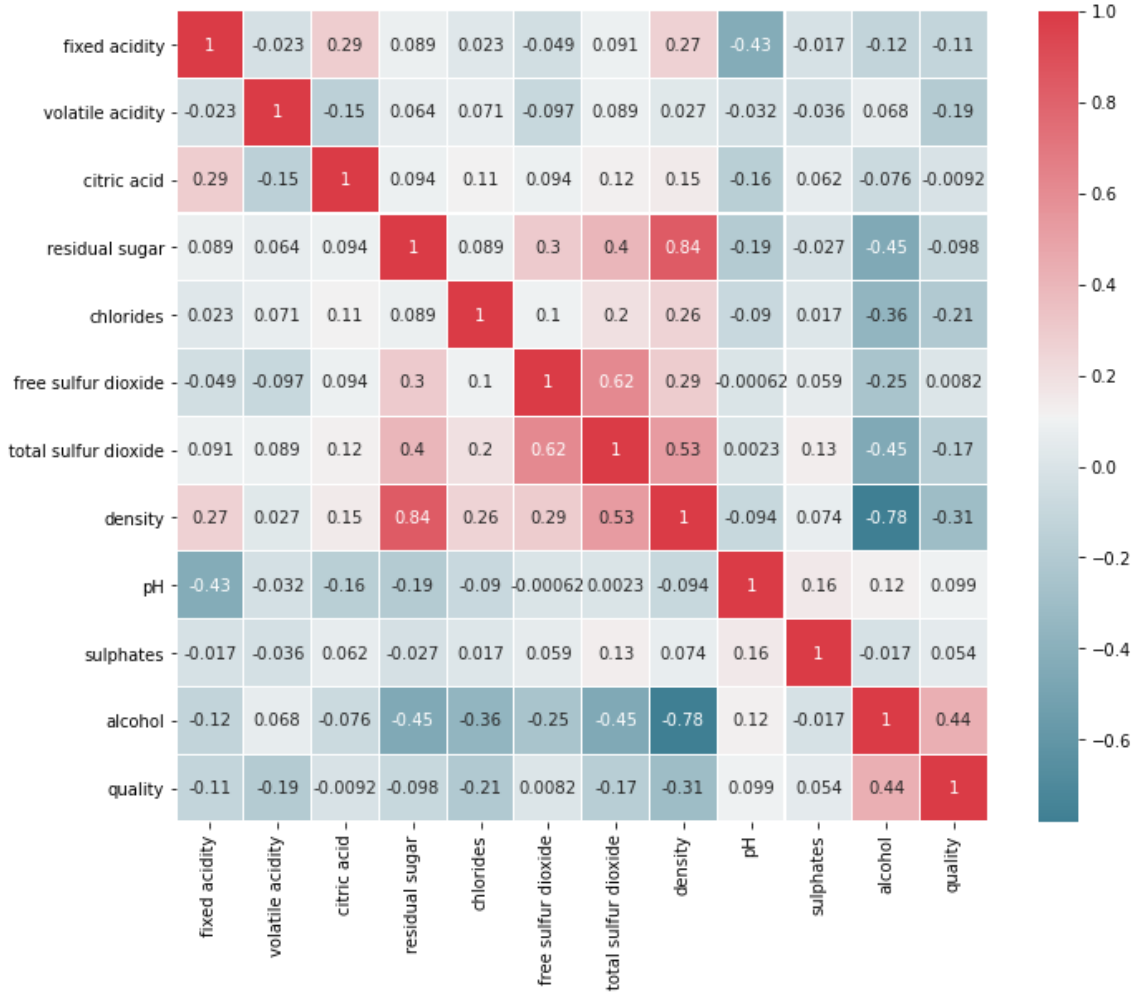


Heatmap

- Heatmaps are great tools to show density in geographical areas.
  - They are a good alternative to point maps when the occurrences are so close to each other that they create clusters that would affect legibility,
  - or when showing general trends rather than specific data points is the goal of the map.

Heatmap

- Table where entries are displayed as a color

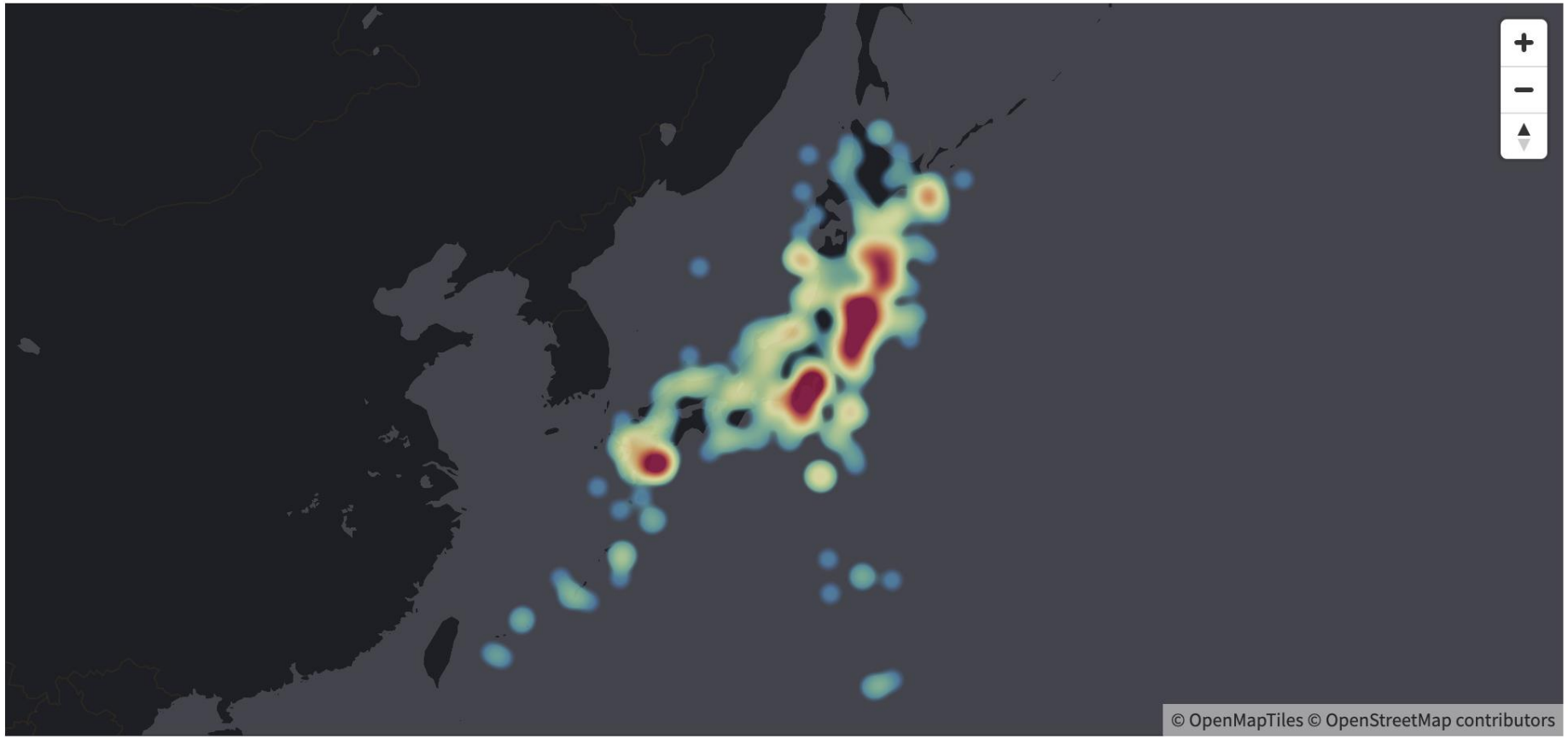


Microsoft Excel conditional formatting

Heatmap

A historical overview of Japan's earthquakes (1802-2021)

Most events have occurred in the east side of the island



Source: National Oceanic and Atmospheric Administration • A Flourish 3D map

Word Cloud

Word Cloud in Moodle

Find materials here!

- https://github.com/xiaoyifu777/data-jour-workshop

References

- Stuart Card, Jock Mackinlay, Ben Schneiderman: Readings in Information Visualization
  - Curated collection of research papers
- Colin Ware: Information Visualization: Perception for Design
  - Perceptual principles applied to data visualization
  - Focus on user interaction
- Chuck Hansen, Chris Johnson: The Visualization Handbook
  - Scientific Visualization
- Tamara Munzner: Visualization Analysis & Design
  - Covers both scientific and information visualization
  - Includes task analysis
- Randy Lao: A Beginner's Guide to the Data Science Pipeline

References

- Alberto Cairo. 2016. The Truthful Art: Data, Charts, and Maps for Communication (1st. ed.). New Riders Publishing, USA.
- https://www.kaggle.com/datasets/tunguz/data-on-covid19-coronavirus?resource=download
- Free Public Data Sets For Analysis: https://www.tableau.com/learn/articles/free-public-data-sets

Thank You