# Lecture #2: Data and Viz

## CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai and Chris Gumb

# Lecture Outline: Data, Summaries, and Visuals

- What (is)are Data?

- Exploratory Data Analysis (EDA)
  - Descriptive Statistics
  - Basic Visualizations

- Historical Interlude

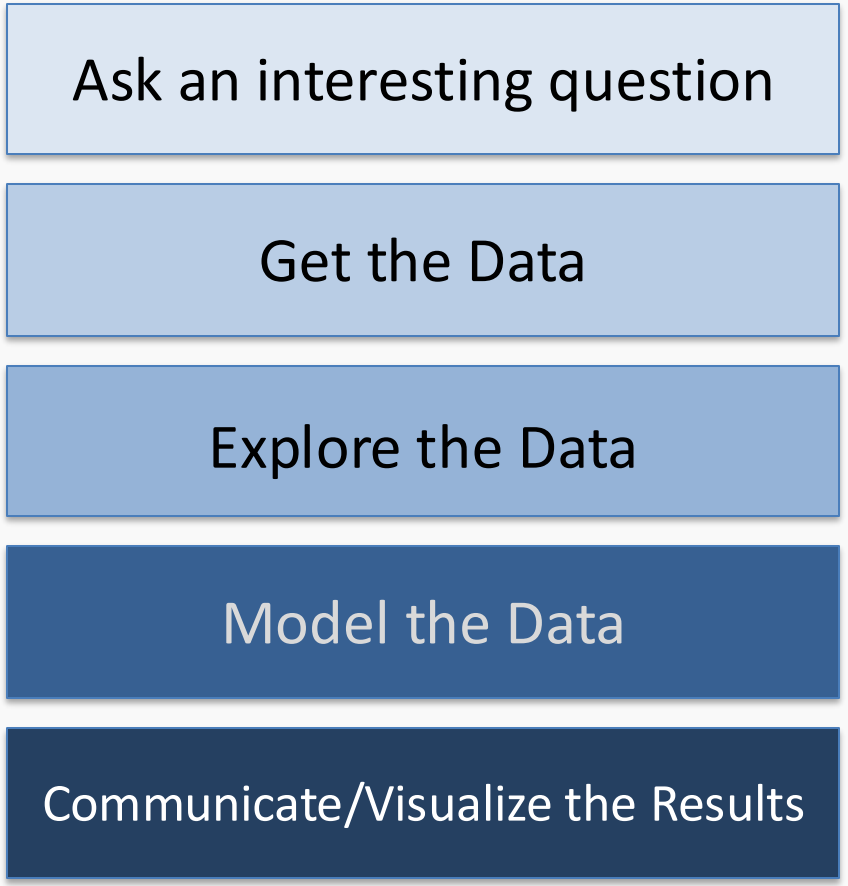
- Effective Visualizations
- Thanks to Kevin Rader for some of the material.

Reading: Ch. 1 in *An Introduction to Statistical Learning* (ISL)

# The Data Science Process

Recall the data science process:

Today we will begin introducing the data collection and data exploration steps.



# What are data?

"A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements."

Claim: everything is (can be) data!

# Where do data come from?

- **Internal sources:** already collected by or is part of the overall data collection of you organization.
  For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources:** available in ready to read format from an outside source for free or for a fee.
  For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference, Kaggle.
- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing.
  For example: data appearing only in print form, or data on websites.

# Ways to gather online data

How to get data generated, published or hosted online:

- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use. For example: Google Map API, Facebook API, Twitter API
- **RSS (Rich Site Summary):** summarizes frequently updated online content in standard format. Free to read if the site has one. For example: news-related sites, blogs
- **Web scraping:** using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file (often in tables).

# Web scraping

- **Why do it?** Older government or smaller news sites might not have APIs for accessing data, or publish RSS feeds or have databases for download. Or, you don't want to pay to use the API or the database.
- **How do you do it?** See HW1 (beautifulsoup)
- **Should you do it?**
  - You just want to explore: Are you violating their terms of service? Privacy concerns for website and their clients?
  - You want to publish your analysis or product: Do they have an API or fee that you are bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

# Types of data

What kind of values are in your data (data types)?

Simple or atomic:

- **Numeric:** integers, floats
- **Boolean:** binary or true false values
- **Strings:** sequence of symbols

# Data types

What kind of values are in your data (data types)? Compound, composed of a bunch of atomic types:

- **Date and time:** compound value with a specific structure
- **Lists:** a list is a sequence of values
- **Dictionaries:** A dictionary is a collection of key-value pairs, a pair of values *x* : *y* where *x* is usually a string called the key representing the “name” of the entry, and *y* is a value of any type.

Example: Student record: what are *x* and *y*?

- First: Natesh
- Last: Pillai
- Classes: [CS-1009A]

# Data storage

How is your data represented and stored (data format)?

- **Tabular Data:** a dataset that is a two-dimensional table, where each row typically represents a single data record, and each column represents one type of measurement (csv, dat, xlsx, etc.).
- **Structured Data:** each data record is presented in a form of a [possibly complex and multi-tiered] dictionary (json, xml, etc.)
- **Semistructured Data:** not all records are represented by the same set of keys or some data records are not represented using the key-value pair structure.

# Tabular Data

In tabular data, we expect each record or observation to represent a set of measurements of a single object or event. Here's an example:

In [4]:

Slide Type

```
imdb = pd.read_csv('imdb_top_1000.csv')
imdb.head()
```

Out[4]:

| | Poster_Link | Series_Title | Released_Year | Certificate | Runtime | Genre | IMDB_Rating | Overview | Meta_score | Director | Star1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | https://m.media-amazon.com/images/M/MV5BMDFkYT... | The Shawshank Redemption | 1994 | A | 142 min | Drama | 9.3 | Two imprisoned men bond over a number of years... | 80.0 | Frank Darabont | Tim Robbins |
| 1 | https://m.media-amazon.com/images/M/MV5BM2MyNj... | The Godfather | 1972 | A | 175 min | Crime, Drama | 9.2 | An organized crime dynasty's aging patriarch t... | 100.0 | Francis Ford Coppola | Marlon Brando |
| 2 | https://m.media-amazon.com/images/M/MV5BMTMxNT... | The Dark Knight | 2008 | UA | 152 min | Action, Crime, Drama | 9.0 | When the menace known as the Joker wreaks havo... | 84.0 | Christopher Nolan | Christian Bale |
| 3 | https://m.media-amazon.com/images/M/MV5BMWMwMG... | The Godfather: Part II | 1974 | A | 202 min | Crime, Drama | 9.0 | The early life and career of Vito Corleone | 90.0 | Francis Ford Coppola | Al Pacino |

# Tabular Data

In [4]:

```
imdb = pd.read_csv('imdb_top_1000.csv')
imdb.head()
```

Out[4]:

| | Poster_Link | Series_Title | Released_Year | Certificate | Runtime | Genre | IMDB_Rating | Overview | Meta_score | Director | Star1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | https://m.media-amazon.com/images/M/MV5BMDFkYT... | The Shawshank Redemption | 1994 | A | 142 min | Drama | 9.3 | Two imprisoned men bond over a number of years... | 80.0 | Frank Darabont | Tim Robbins |
| 1 | https://m.media-amazon.com/images/M/MV5BM2MyNj... | The Godfather | 1972 | A | 175 min | Crime, Drama | 9.2 | An organized crime dynasty's aging patriarch t... | 100.0 | Francis Ford Coppola | Marlon Brando |
| 2 | https://m.media-amazon.com/images/M/MV5BMTMxNT... | The Dark Knight | 2008 | UA | 152 min | Action, Crime, Drama | 9.0 | When the menace known as the Joker wreaks havo... | 84.0 | Christopher Nolan | Christian Bale |
| 3 | https://m.media-amazon.com/images/M/MV5BMWMwMG... | The Godfather: Part II | 1974 | A | 202 min | Crime, Drama | 9.0 | The early life and career of Vito Corleone | 90.0 | Francis Ford Coppola | Al Pacino |

Each type of measurement is called a **variable** or an **attribute** of the data (e.g. `seq_id`, `status` and `duration` are variables or attributes). The number of attributes is called the **dimension**. These are often called **features**.

We expect each table to contain a set of **records** or **observations** of the same kind of object or event.

# Types of Data

We'll see later that it's important to distinguish between classes of variables or attributes based on the type of values they can take on.

- **Quantitative variable:** is numerical and can be either:
  - **discrete -** a finite number of values are possible in any bounded interval. For example: "Number of siblings" is a discrete variable
  - **continuous -** an infinite number of values are possible in any bounded interval. For example: "Height" is a continuous variable
- **Categorical variable:** no inherent order among the values For example: "What kind of pet do you have?" is a categorical variable

# Common Issues

Common issues with data:

- Missing values: how do we fill in?
- Wrong values: how can we detect and correct?
- Messy format
- Not usable: the data cannot answer the question posed

# Messy Data

The following is a table accounting for the number of produce deliveries over a weekend.

What are the variables in this dataset? What object or event are we measuring?

| | Friday | Saturday | Sunday |
|---|---|---|---|
| Morning | 15 | 158 | 10 |
| Afternoon | 2 | 90 | 20 |
| Evening | 55 | 12 | 45 |

What's the issue? How do we fix it?

| ID | Time | Day | Number |
|---|---|---|---|
| 1 | Morning | Friday | 15 |
| 2 | Morning | Saturday | 158 |
| 3 | Morning | Sunday | 10 |
| 4 | Afternoon | Friday | 2 |
| 5 | Afternoon | Saturday | 9 |
| 6 | Afternoon | Sunday | 20 |
| 7 | Evening | Friday | 55 |
| 8 | Evening | Saturday | 12 |
| 9 | Evening | Sunday | 45 |

# Tabular = Happy Pavlos

Common causes of messiness are:

- Column headers are values, not variable names
- Variables are stored in both rows and columns
- Multiple variables are stored in one column/entry
- Multiple types of experimental units stored in same table

In general, we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation.

We want to **tabularize** the data. This makes Python happy.

# Lecture Outline: Data, Summaries, and Visuals

- What are Data?
- **Exploratory Data Analysis (EDA)**
  - **Descriptive Statistics**
  - Basic Visualizations
- Historical Interlude
- Effective Visualizations

Reading: Ch. 1 in *An Introduction to Statistical Learning* (ISL)

# Basics of Sampling

Population versus sample:

- A **population** is the entire set of objects or events under study. Population can be hypothetical “all students” or all students in this class.
- A **sample** is a “representative” subset of the objects or events under study. Needed because it’s impossible or intractable to obtain or compute with population data.

Biases in samples:

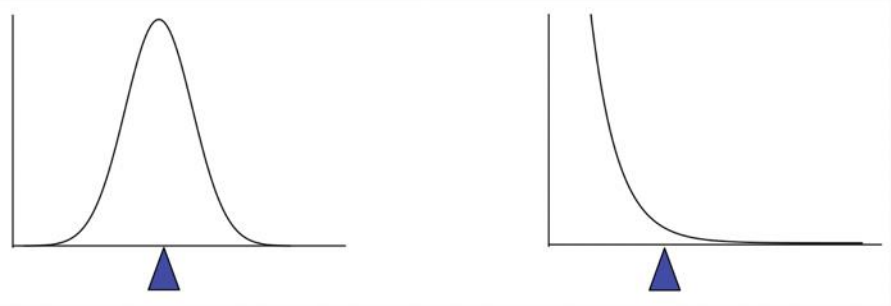
- **Selection bias**: some subjects or records are more likely to be selected
- Volunteer/**nonresponse bias**: subjects or records who are not easily available are not represented

Examples?

# Sample mean

The **mean** of a set of *n* observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.

Key theme: there is always uncertainty involved when calculating a sample mean to estimate a population mean.

# Sample median

The **median** of a set of n number of observations in a sample, ordered by value, of a variable is is defined by

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example (already in order):

Ages: 17, 19, 21, 22, 23, 23, 23, 38

Median = (22+23)/2 = 22.5

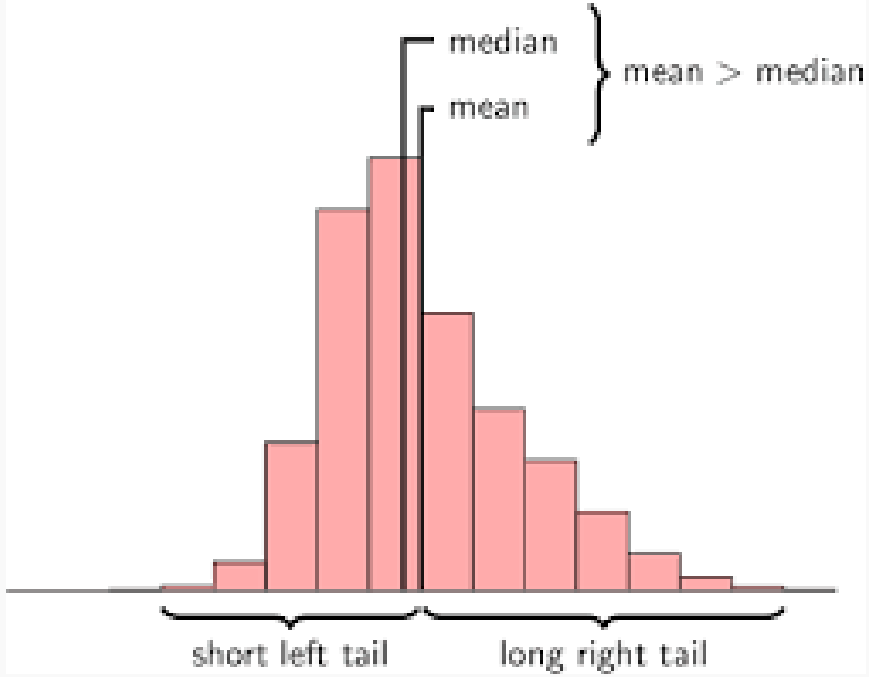
The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

# Mean vs. Median

The mean is sensitive to extreme values (**outliers**)

# Mean, median, and skewness

The mean is sensitive to outliers:




The above distribution is called **right-skewed** since the mean is greater than the median. Note: **skewness** often "follows the longer tail".

# Computational time

How hard (in terms of algorithmic complexity) is it to calculate

- the mean?

  at most $O(n)$

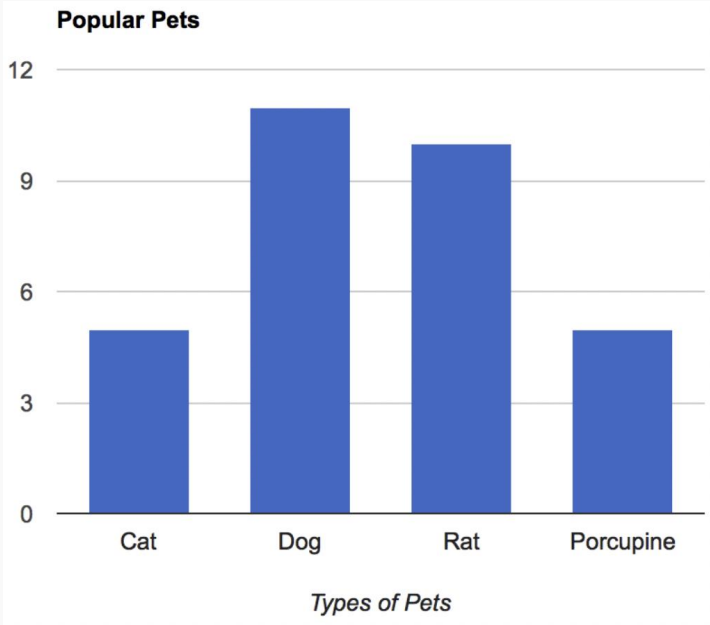
- the median?

  at most $O(n \log n)$ or possibly $O(n)$

Note: Practicality of implementation should be considered!

# Regarding Categorical Variables…

For categorical variables, neither mean or median make sense. Why?




The mode might be a better way to find the most "representative" value.

# Measures of Spread: Range

The spread of a sample of observations measures how well the mean or median describes the sample.

One way to measure spread of a sample of observations is via the **range**.

Range = Maximum Value - Minimum Value

# Measures of Spread: Variance

The (sample) **variance**, denoted $s^2$, measures how much on average the sample values deviate from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2$$

Note: the term $|x_i - \bar{x}|$ measures the amount by which each $x_i$ deviates from the mean $\bar{x}$. Squaring these deviations means that $s^2$ is sensitive to extreme values (outliers).

Note: $s^2$ doesn't have the same units as the $x_i$ :(

What does a variance of 1,008 mean? Or 0.0001?

# Measures of Spread: Standard Deviation

The (sample) **standard deviation**, denoted $s$, is the square root of

the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Note: $s$ does have the same units as the $x_i$. Phew!

# Lecture Outline: Data, Summaries, and Visuals

- What are Data?
- **Exploratory Data Analysis (EDA)**
  - Descriptive Statistics
  - **Basic Visualizations**
- Historical Interlude
- Effective Visualizations

Reading: Ch. 1 in *An Introduction to Statistical Learning* (ISL)

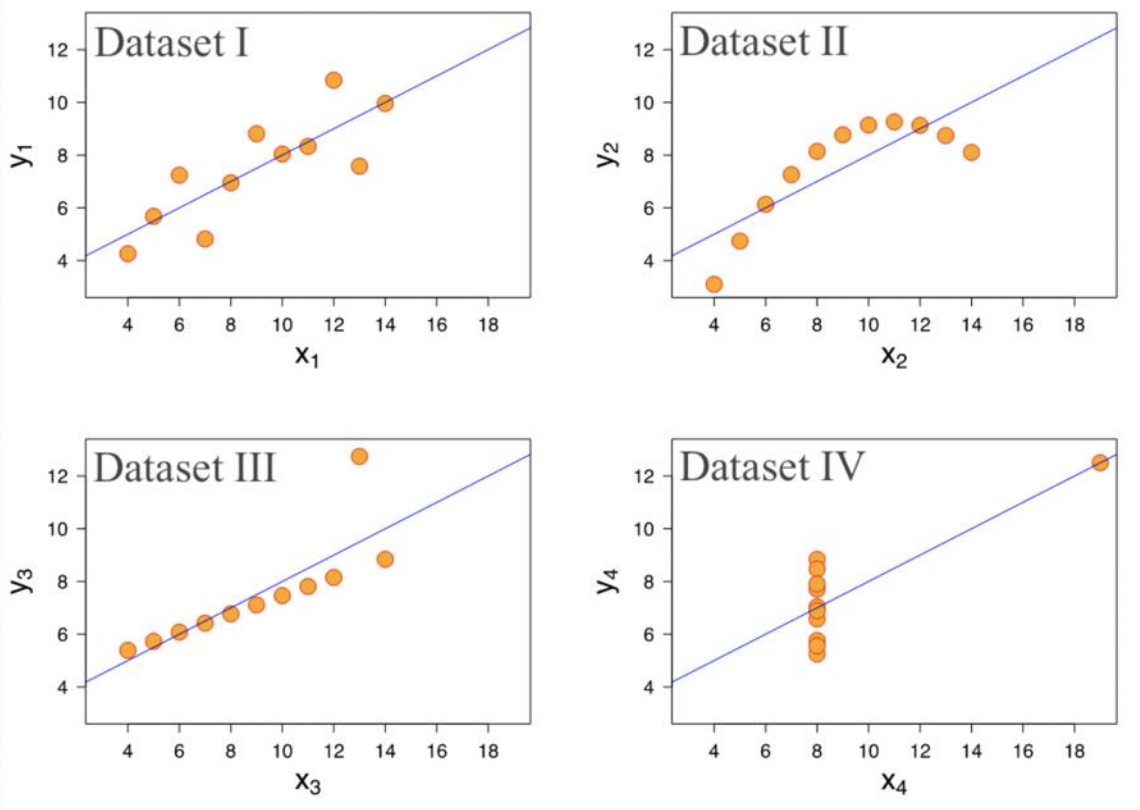
# Anscombe's Data

The following four data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

| | Dataset I | | Dataset II | | Dataset III | | Dataset IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| | 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| | 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |
| **Sum:** | 99.00 | 82.51 | 99.00 | 82.51 | 99.00 | 82.51 | 99.00 | 82.51 |
| **Avg:** | 9.00 | 7.50 | 9.00 | 7.50 | 9.00 | 7.50 | 9.00 | 7.50 |
| **Std:** | 3.32 | 2.03 | 3.32 | 2.03 | 3.32 | 2.03 | 3.32 | 2.03 |

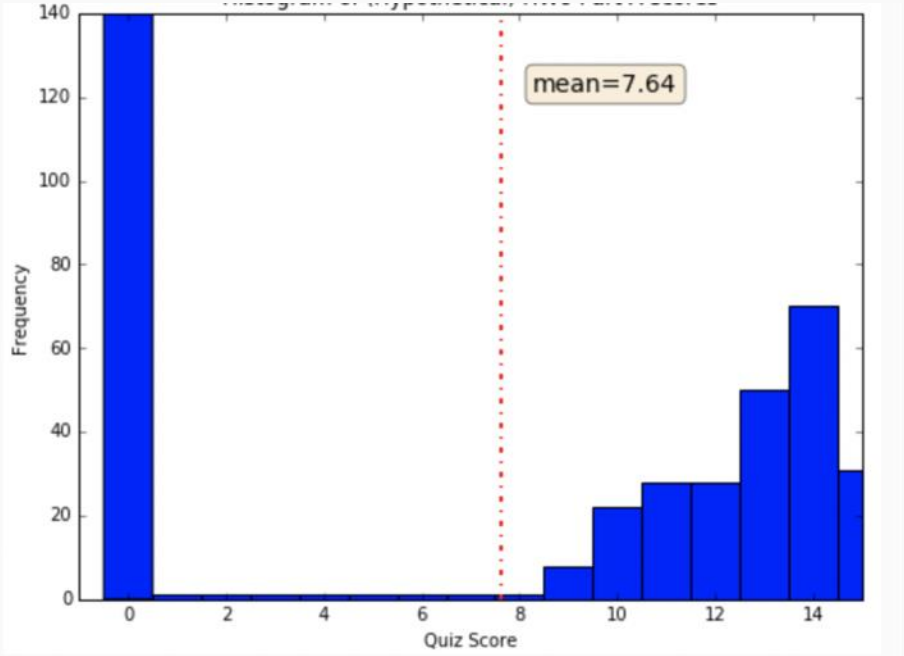
# Anscombe's Data (cont.)

Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words:



# More Visualization Motivation

If I tell you that the average score for Homework 0 in my other class was: 7.64/15 = 50.9%, what does that suggest?




And what does the graph suggest?

# More Visualization Motivation

Visualizations help us to analyze and explore the data. They help to:

- Identify hidden patterns and trends
- Formulate/test hypotheses
- Communicate any modeling results
  - Present information and ideas succinctly
  - Provide evidence and support
  - Influence and persuade
- Determine the next step in analysis/modeling

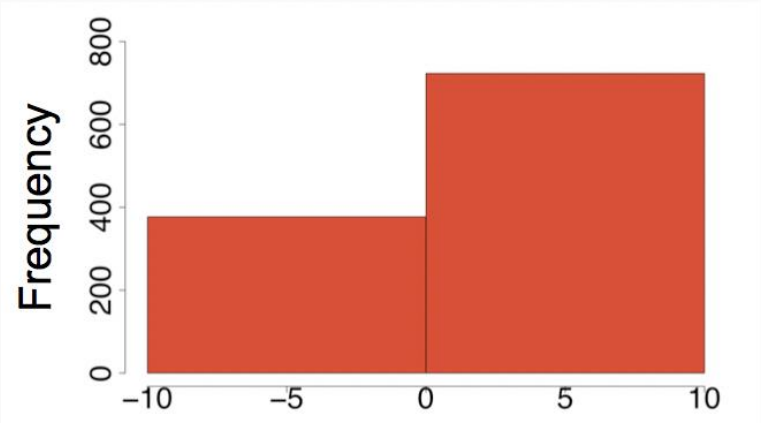
# Types of Visualizations

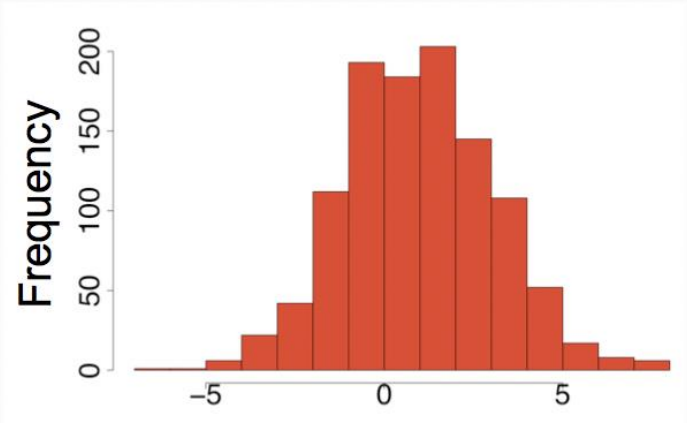
What do you want your visualization to show about your data?

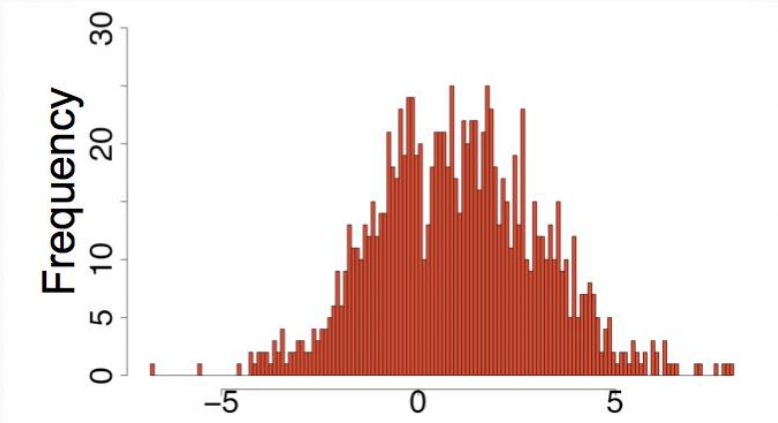
- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how the dataset breaks down into subgroups
- **Comparison:** how trends in multiple variable or datasets compare

# Histograms to visualize distribution

A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.



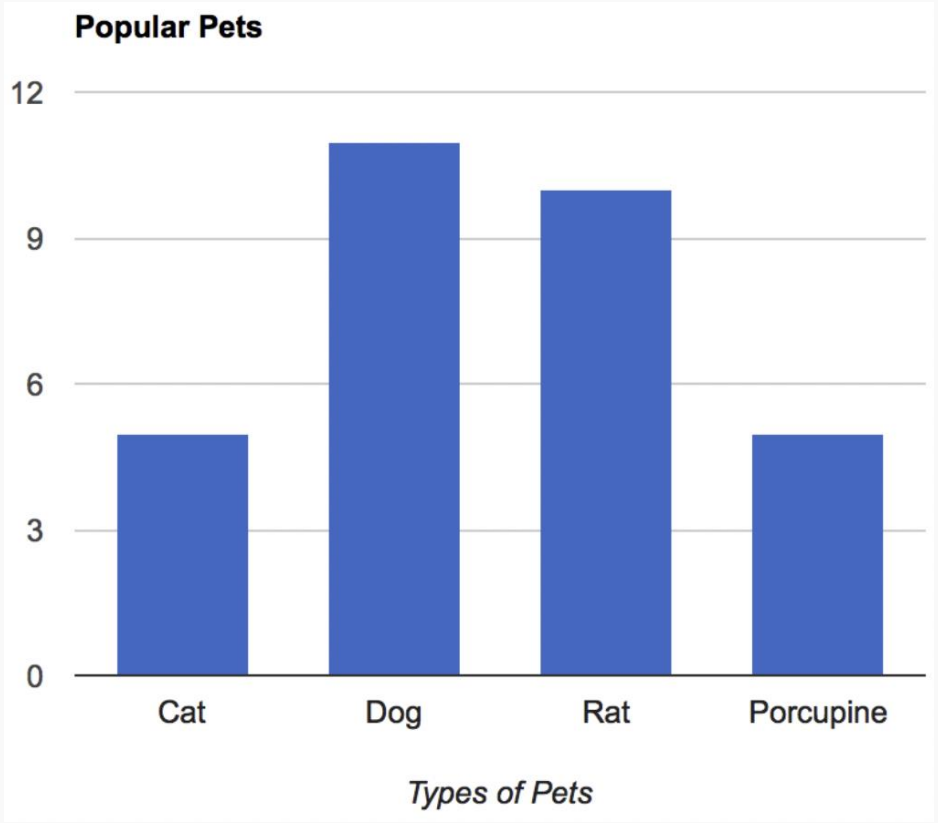




Note: Trends in histograms are sensitive to number of bins.

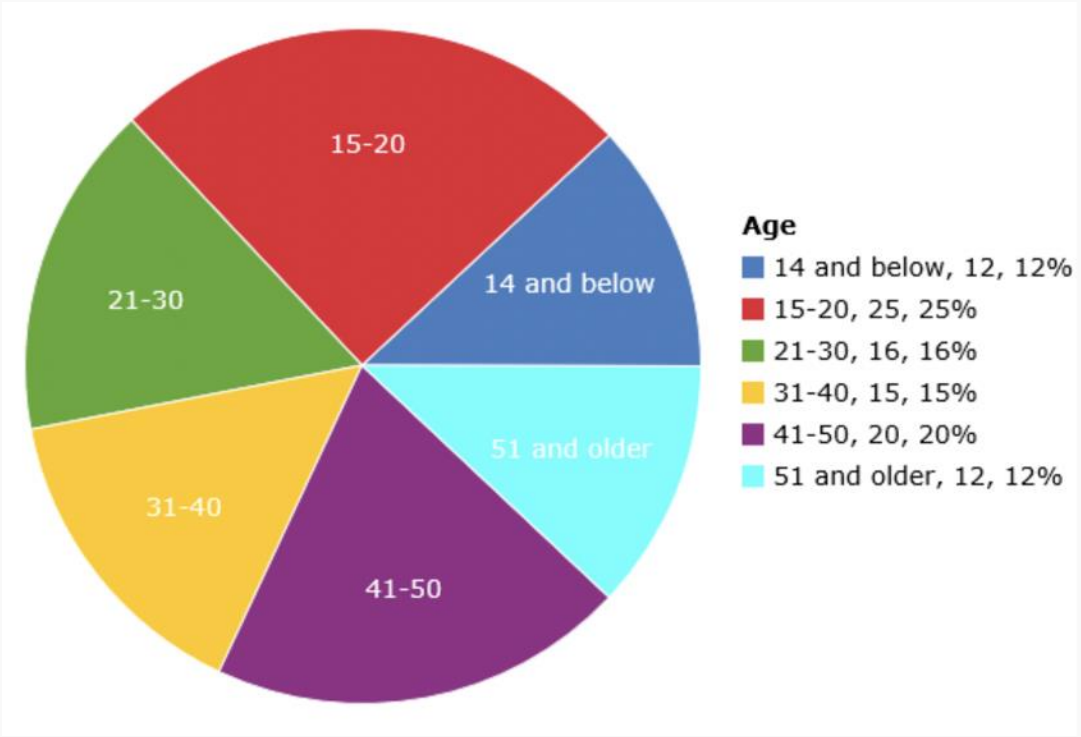
# Bar plot for a categorical variable

A **bar plot** is a way to visualize the composition (aka, distribution) of a categorical variable.



# Pie chart for a categorical variable?

A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).



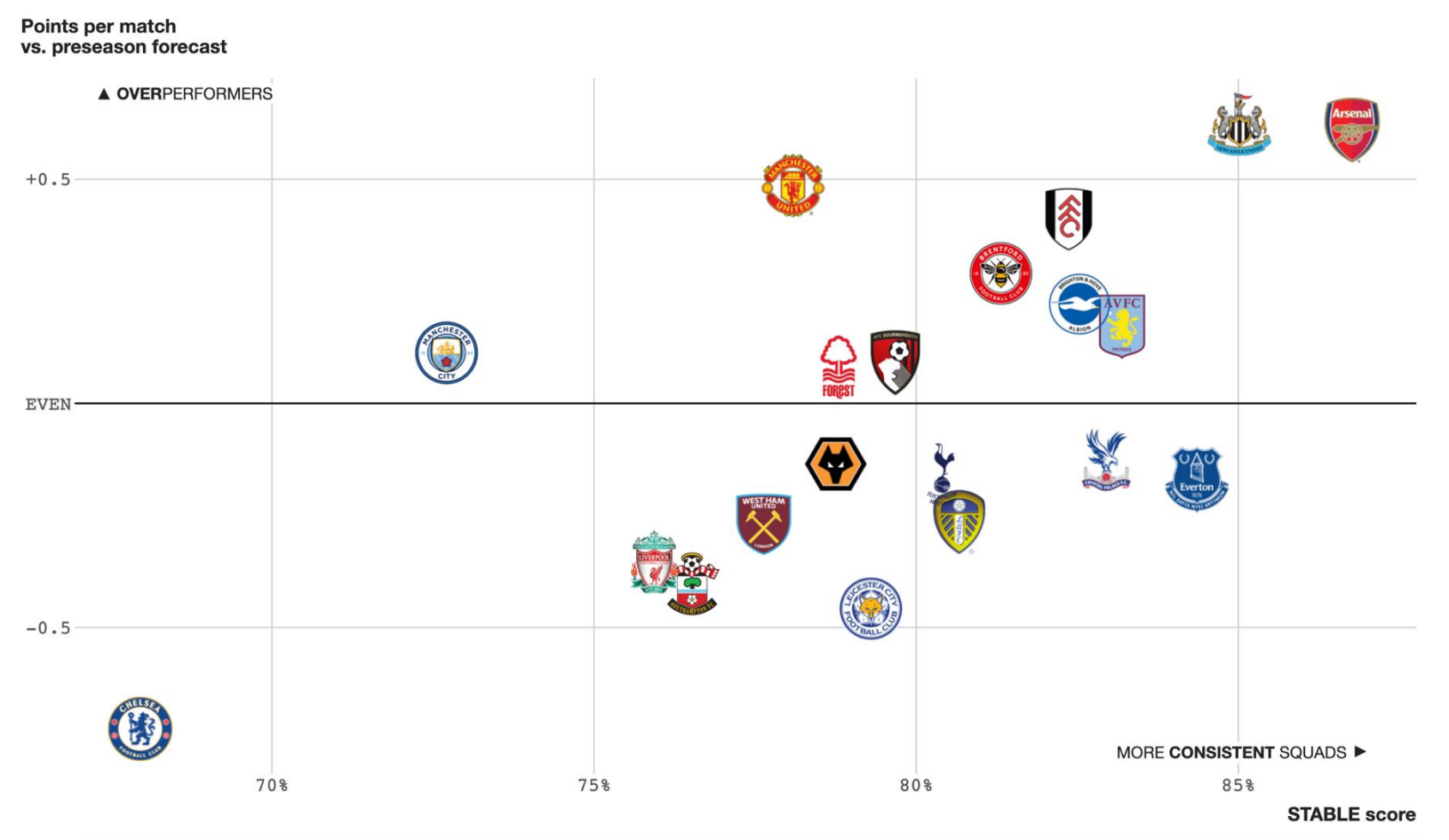

Pie charts are often frowned upon (and bar charts are used instead). Why?

# Pies vs. Bars


MY FAVORITE BARS
(In Percentage of Awesomeness)
McGee's
MacLaren's
28%
30%
12%
27%
White Horse Tavern
3%
King Cole Bar
P & G
MY FAVORITE PIES
(In Percentage of Tastiness)

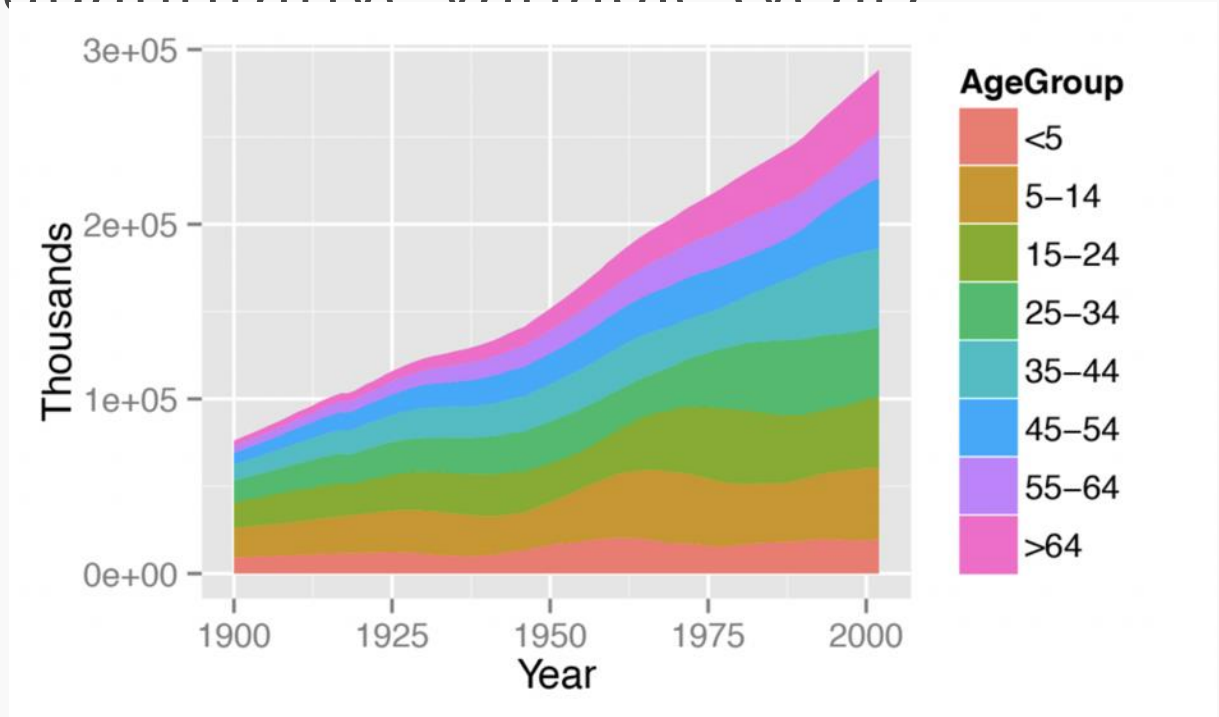
# Scatter plots to visualize relationships

A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.



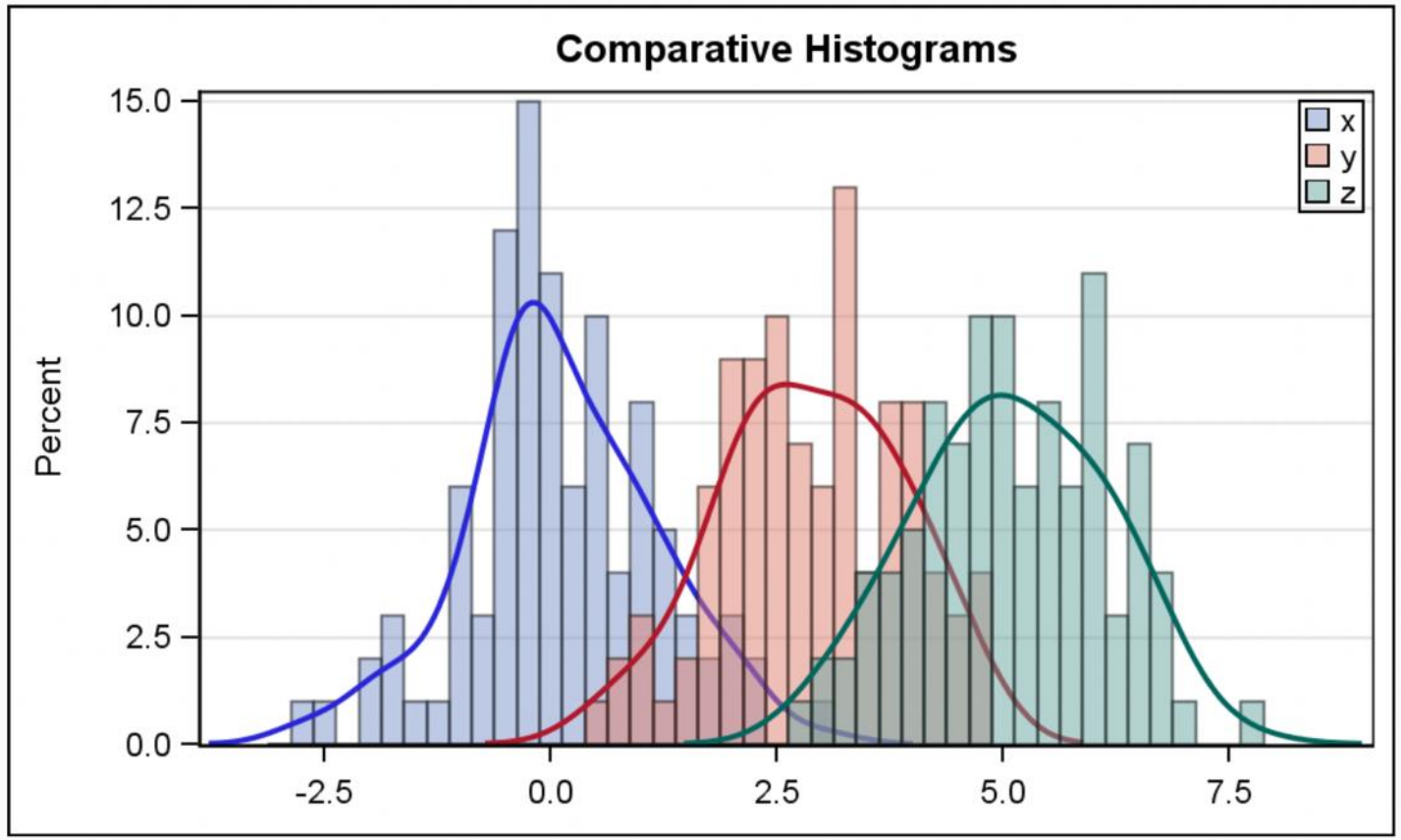
# Stacked area graph to show trend over time

A **stacked area graph** is a way to visualize the composition of a group as it changes over time (or some other quantitative variable). This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year)



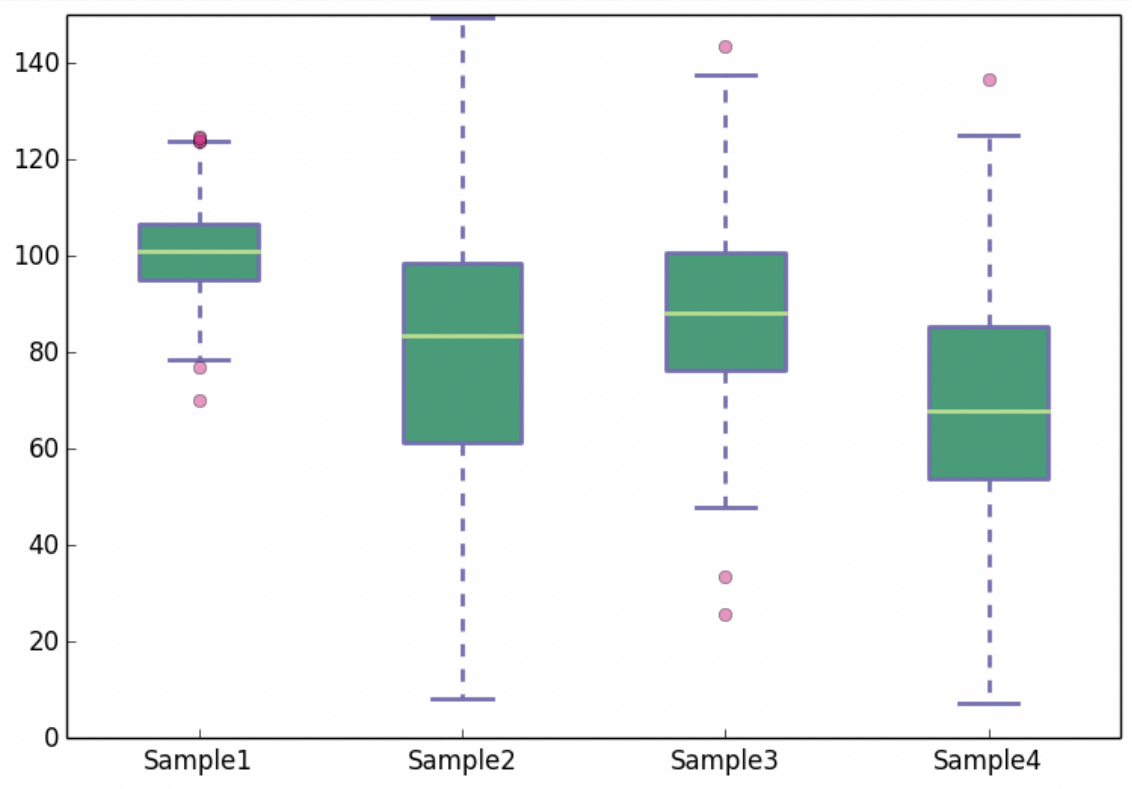
# Multiple histograms

Plotting **multiple histograms (**and **kernel density estimates** of the distribution, here) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).



# Boxplots

A **boxplot** is a simplified visualization to compare a quantitative variable across groups. It highlights the range, quartiles, median and any outliers present in a data set.



# [Not] Anything is possible!

Often your dataset seem too complex to visualize:

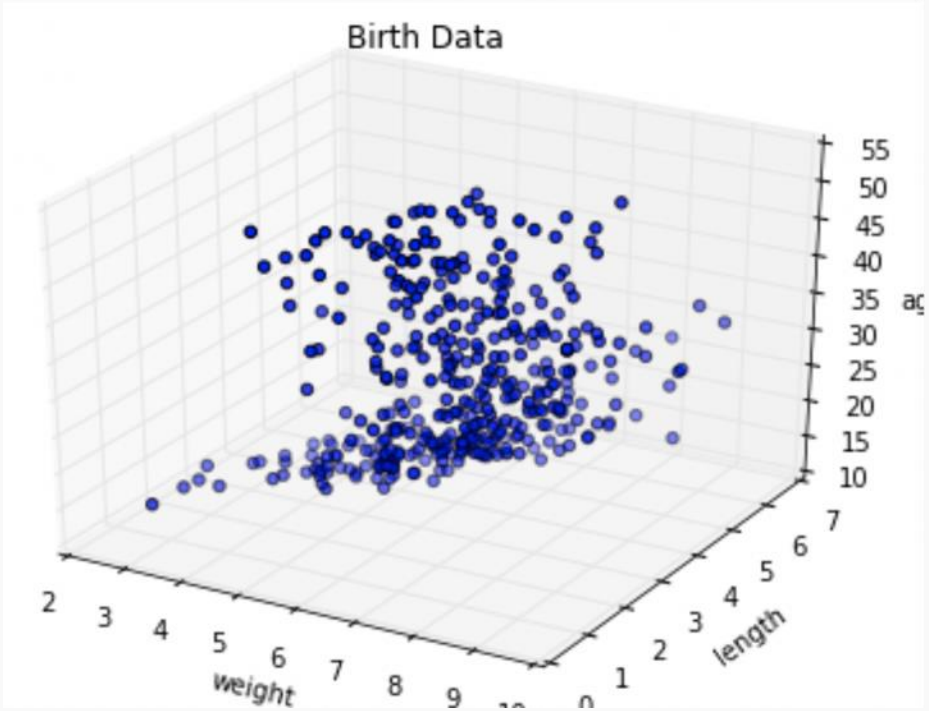
- Data is too high dimensional (how do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like Cat or No?)

# 3+ variables

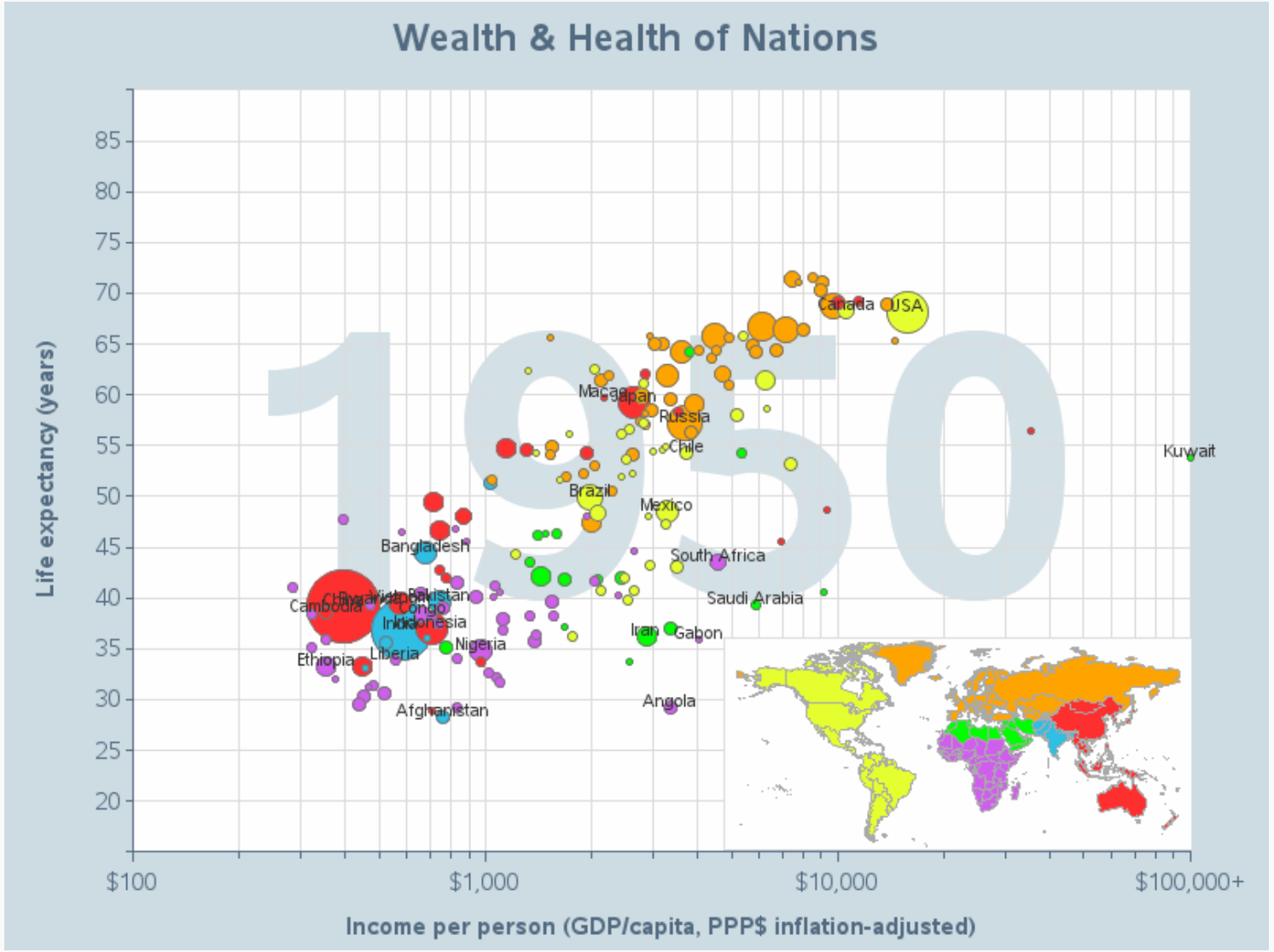
- Visualizing the distribution of 1 variable or the relationship between two variables is often straightforward.
- But how can we visualize how 3, 4, or more variables relate?
- It depends on the type of variables (categorical or numeric) that are involved.
- What is the best way to visualize relationships when…
  - all 3 are numeric?
  - 1 or more are categorical?

# More dimensions not always better

When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful



# Visualizing 3+ variables

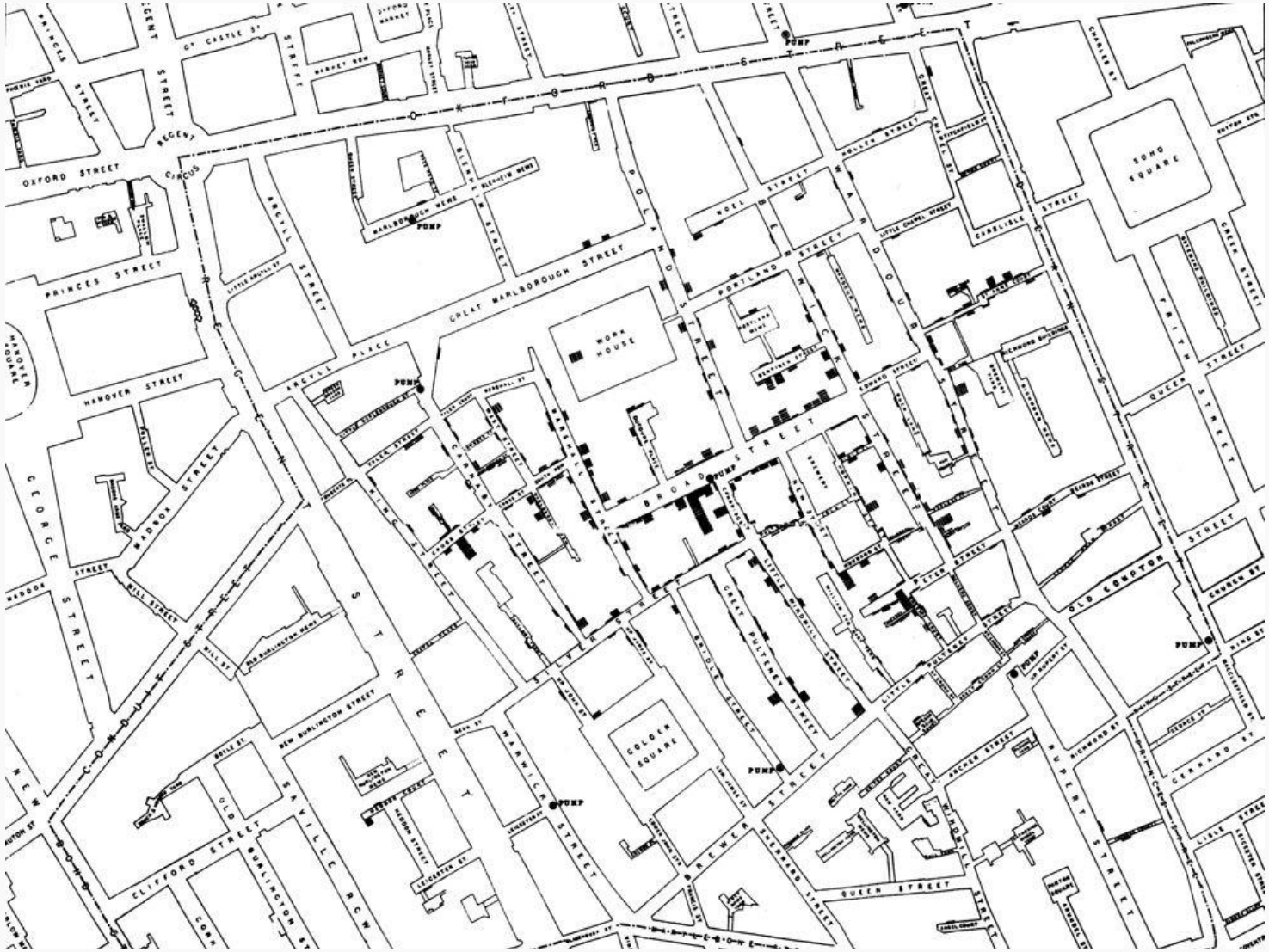


# Lecture Outline: Data, Summaries, and Visuals

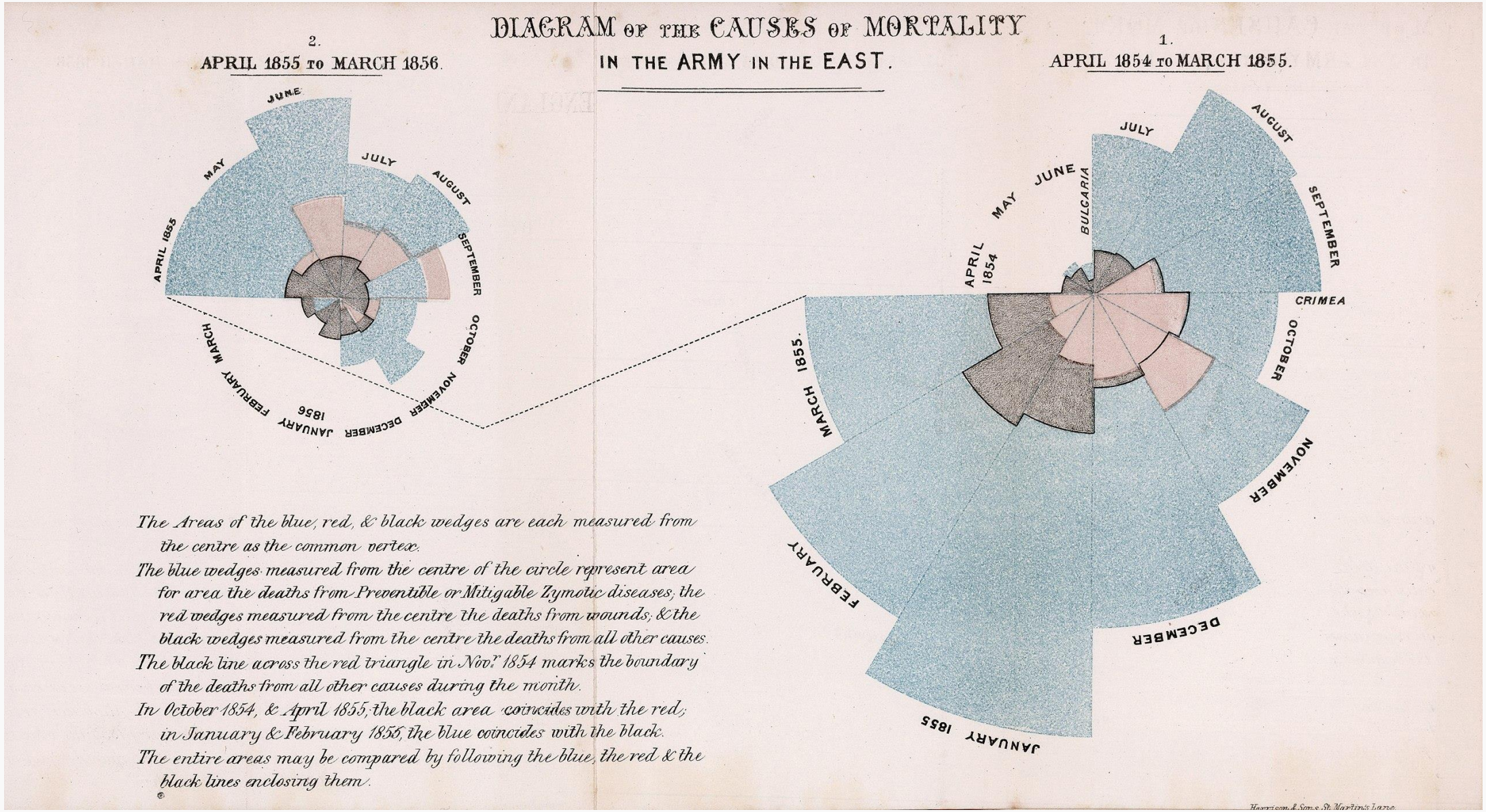
- What are Data?
- Exploratory Data Analysis (EDA)
  - Descriptive Statistics
  - Basic Visualizations
- **Historical Interlude**
- Effective Visualizations

Reading: Ch. 1 in *An Introduction to Statistical Learning* (ISL)

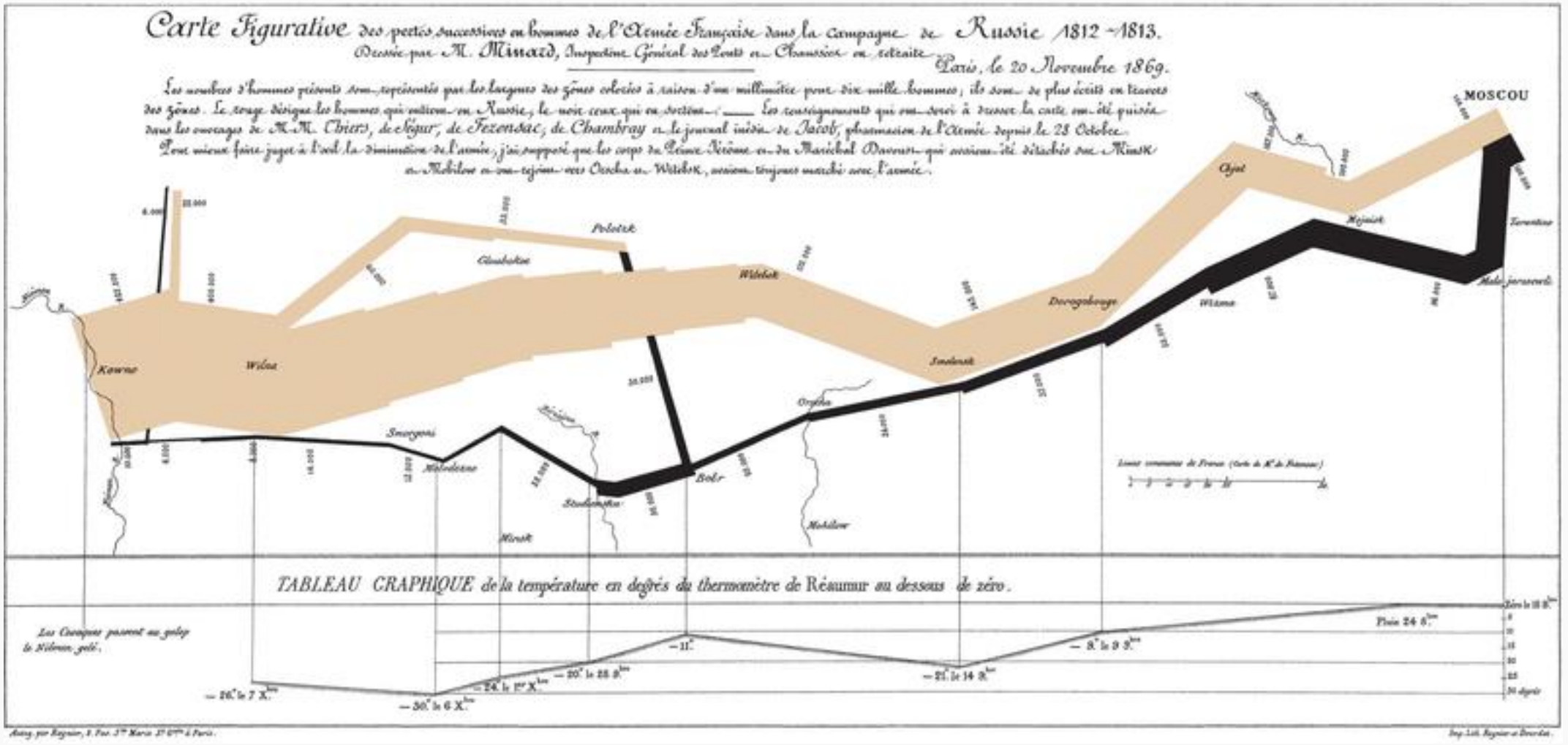
# John Snow’s Cholera Outbreak (1854)



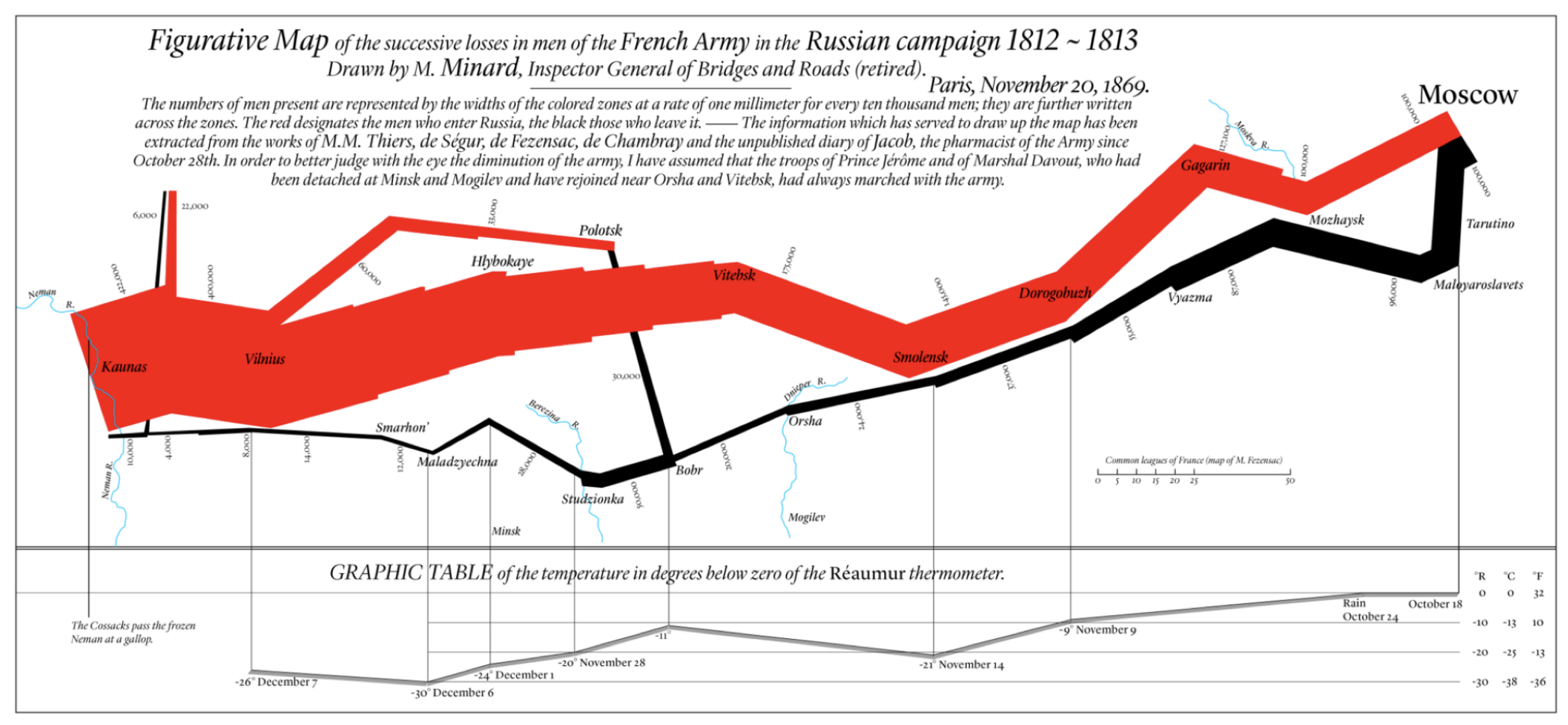
# Florence Nightingale's Rose Chart of the Crimean War (1858)



# Minard's Visual of Napoleon's March through Russia (1869)



# Minard's Visual of Napoleon's March through Russia (1869)



# Lecture Outline: Data, Summaries, and Visuals

- What are Data?
- Exploratory Data Analysis (EDA)
  - Descriptive Statistics
  - Basic Visualizations
- Historical Interlude
- **Effective Visualizations**

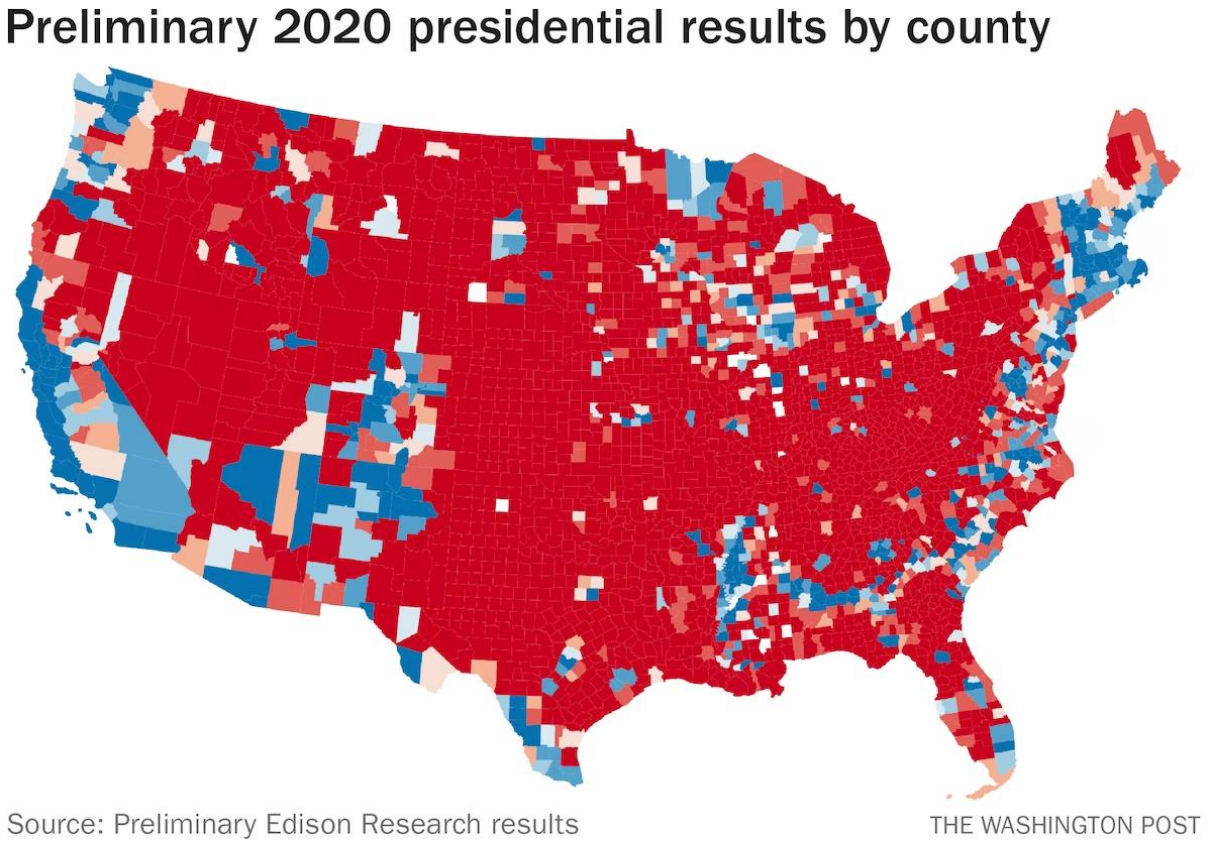
Reading: Ch. 1 in *An Introduction to Statistical Learning* (ISL)

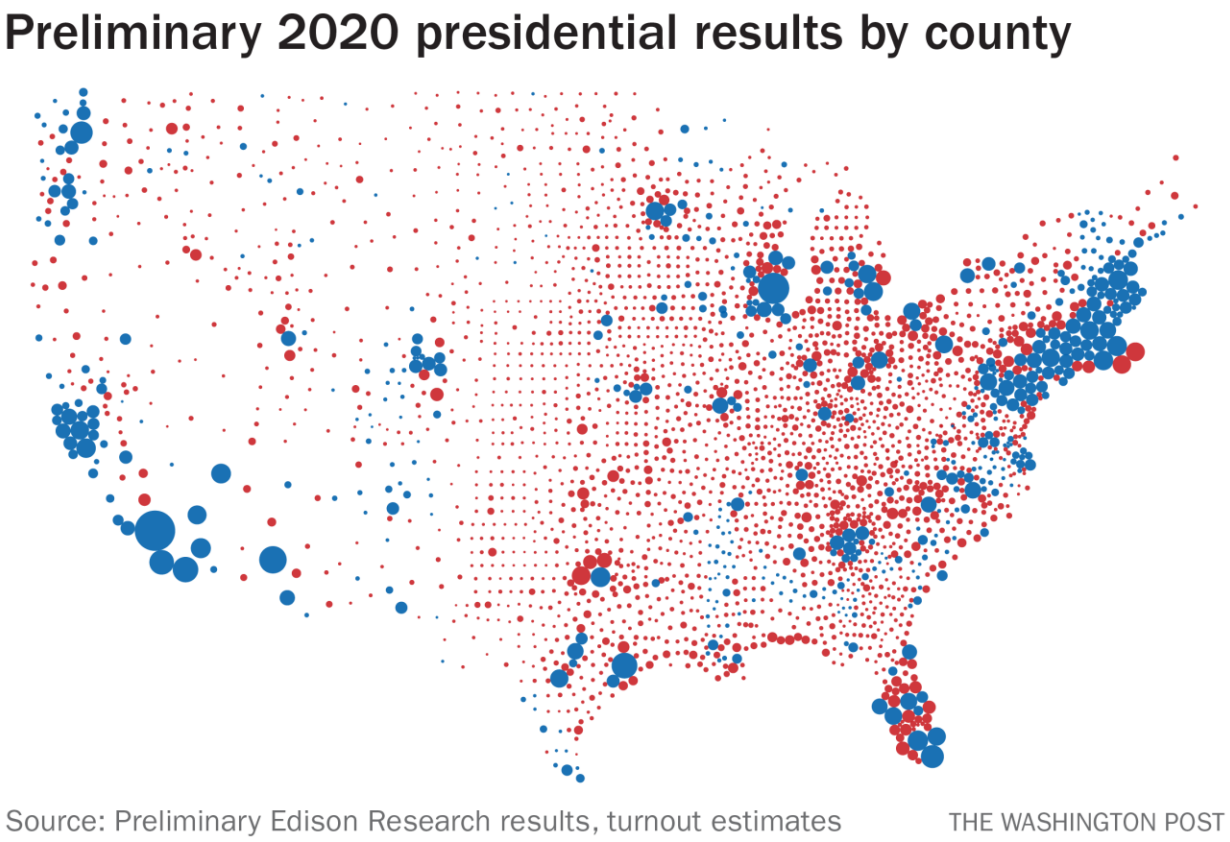
# Pointers for Effective Visualizations

- Have graphical integrity
- Keep it simple
- Use the right display
- Use color strategically
- Know your audience

# Graphical Integrity

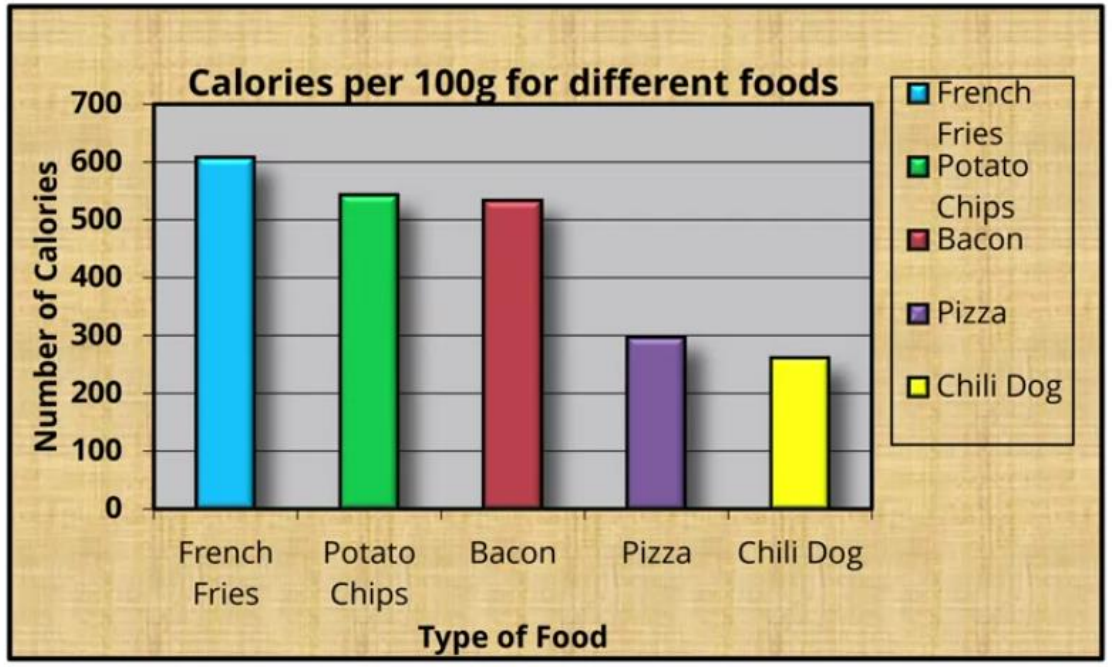
What does this map suggest about the 2020 election?

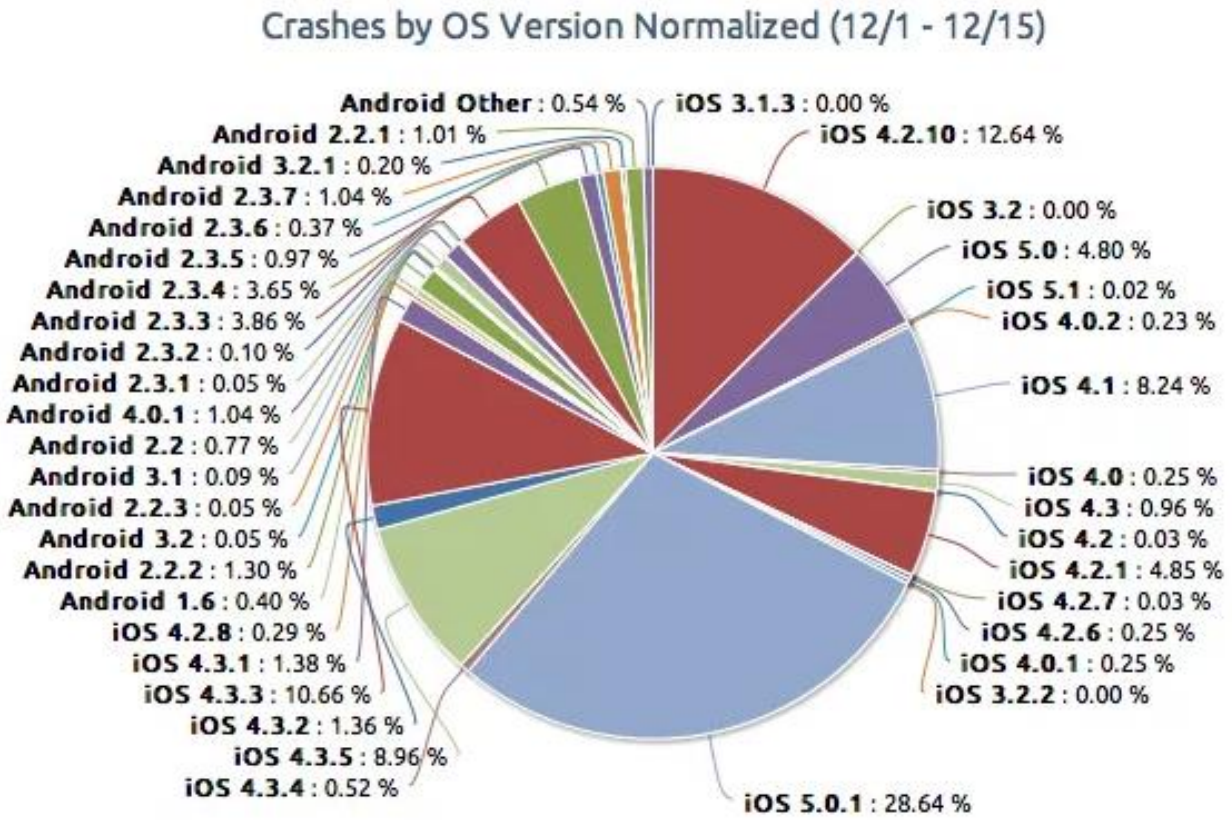


# Keep it simple

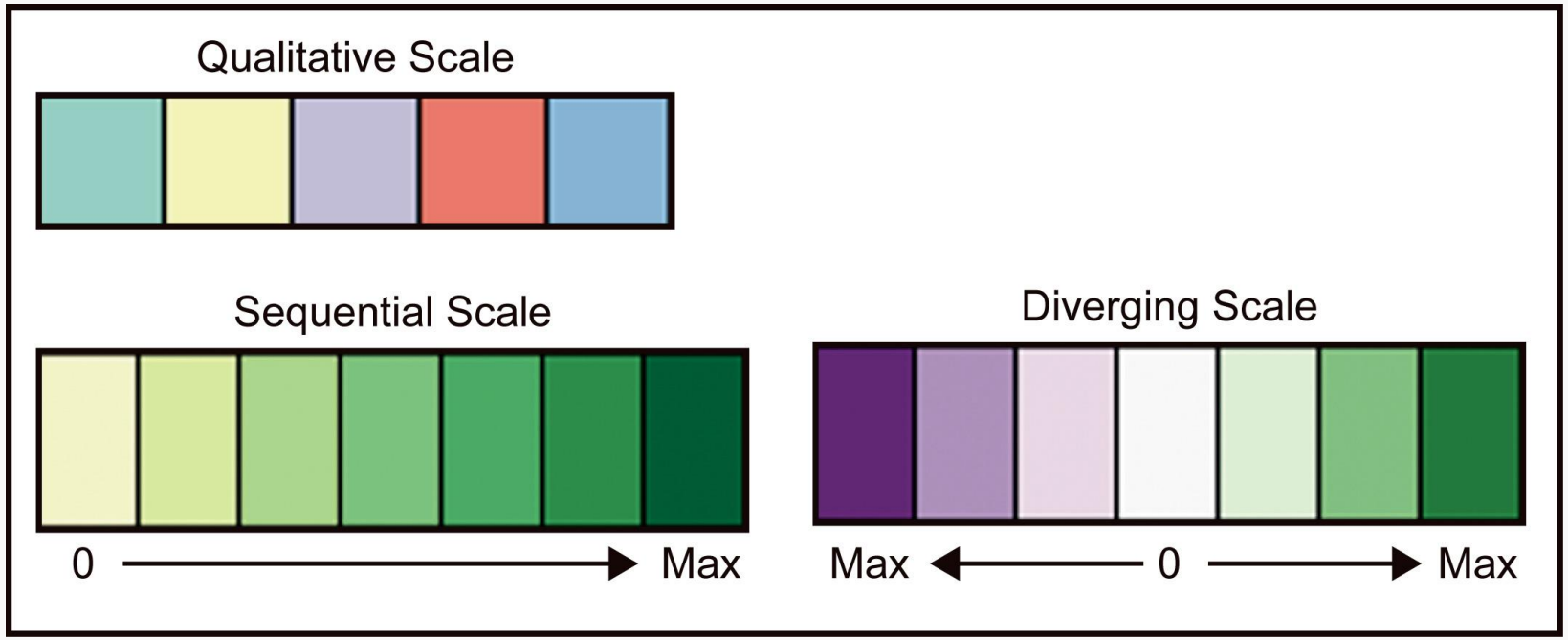
## Avoid Chart Junk



# Use color strategically

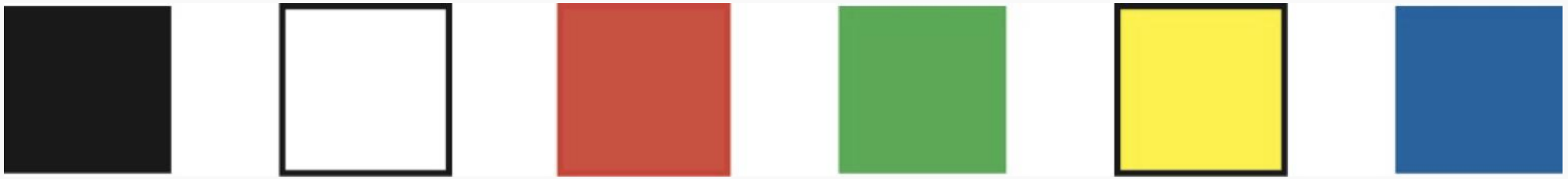
Type of scale matters:




What are example variables for each?

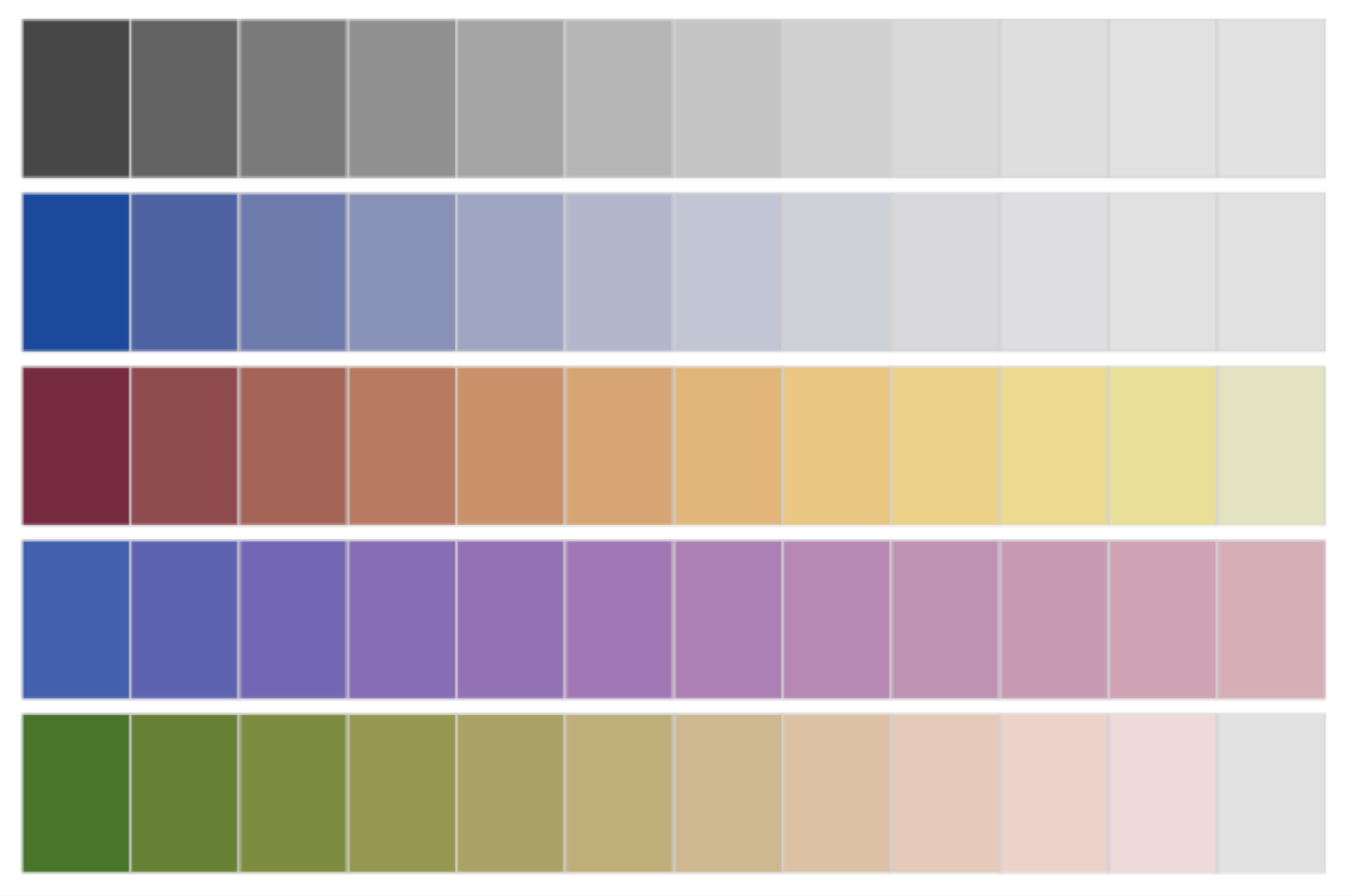
# Use color strategically

Colors for categories (Do not use more than 5-8 colors at once)



# Use color strategically

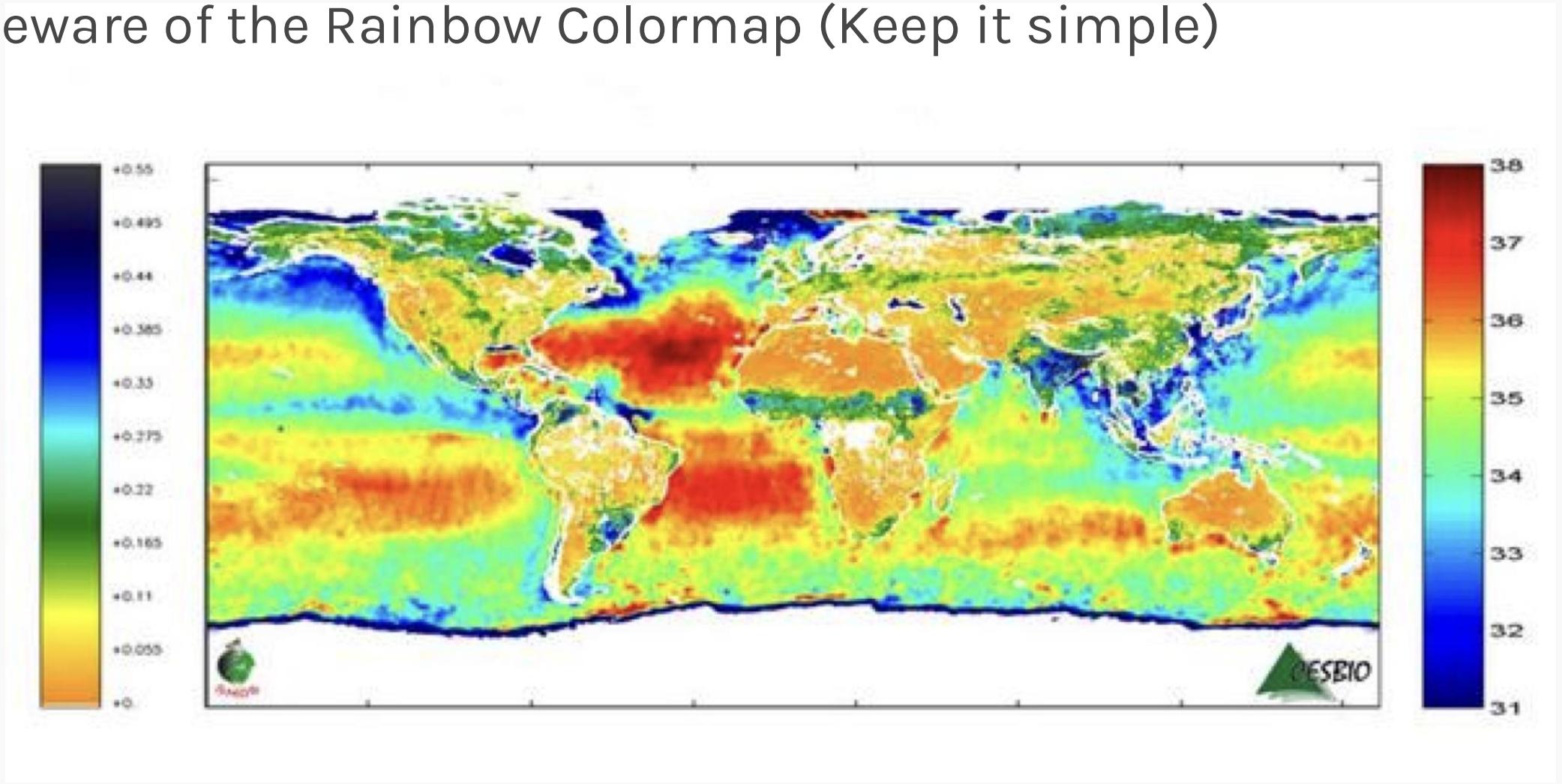
## Colors for Ordinal Data -- Vary luminance and saturation



Zeilis et al, 2009, "Escaping RGBland: Selecting Colors for Statistical Graphics"

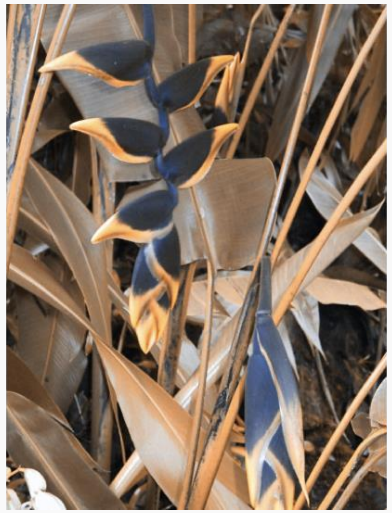
# Use color strategically

## Beware of the Rainbow Colormap (Keep it simple)

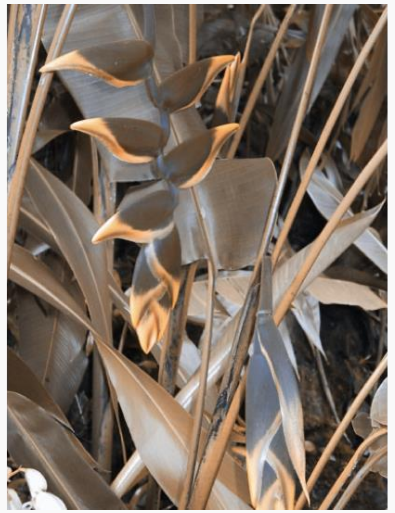


# Use color strategically

## Be cognizant of color blindness



Protanope



Deuteranope

Red / green deficiencies

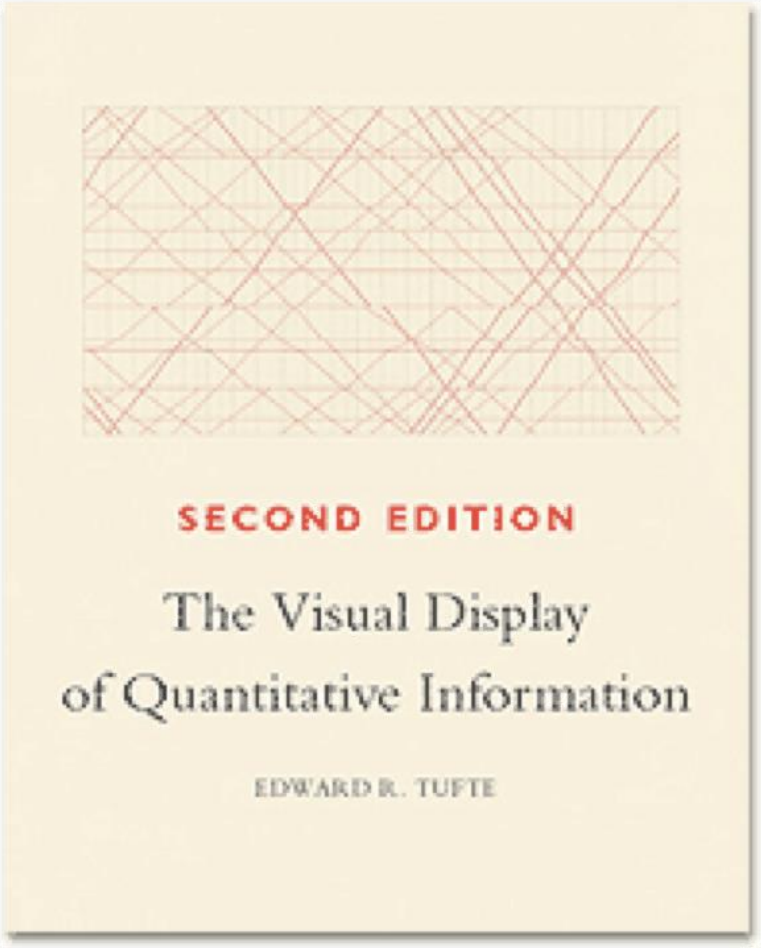
Tritanope

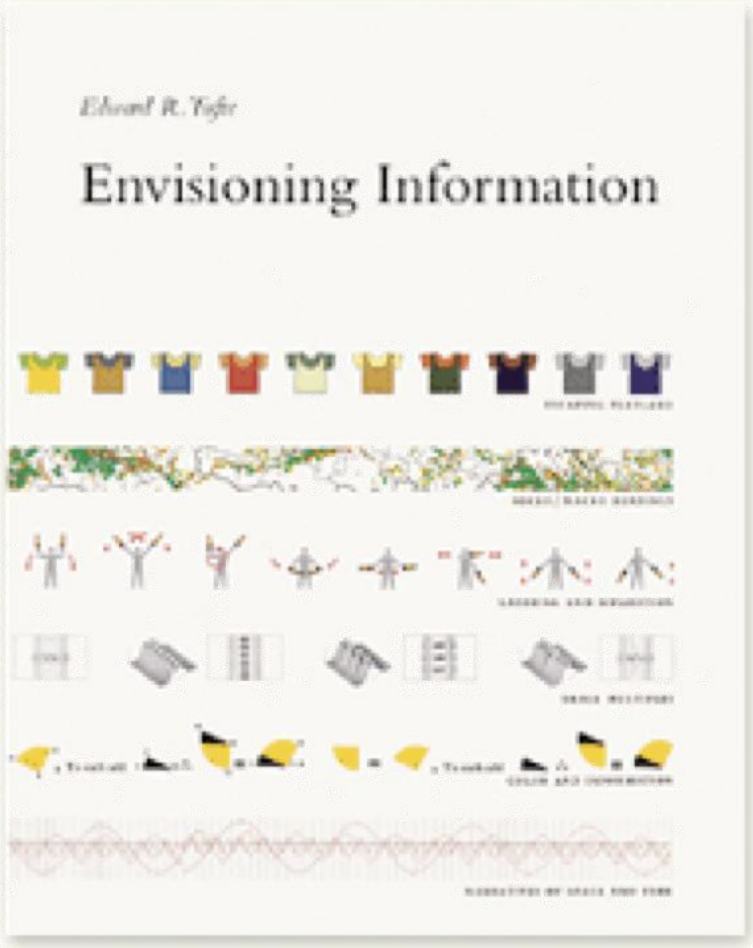
Blue / Yellow deficiency

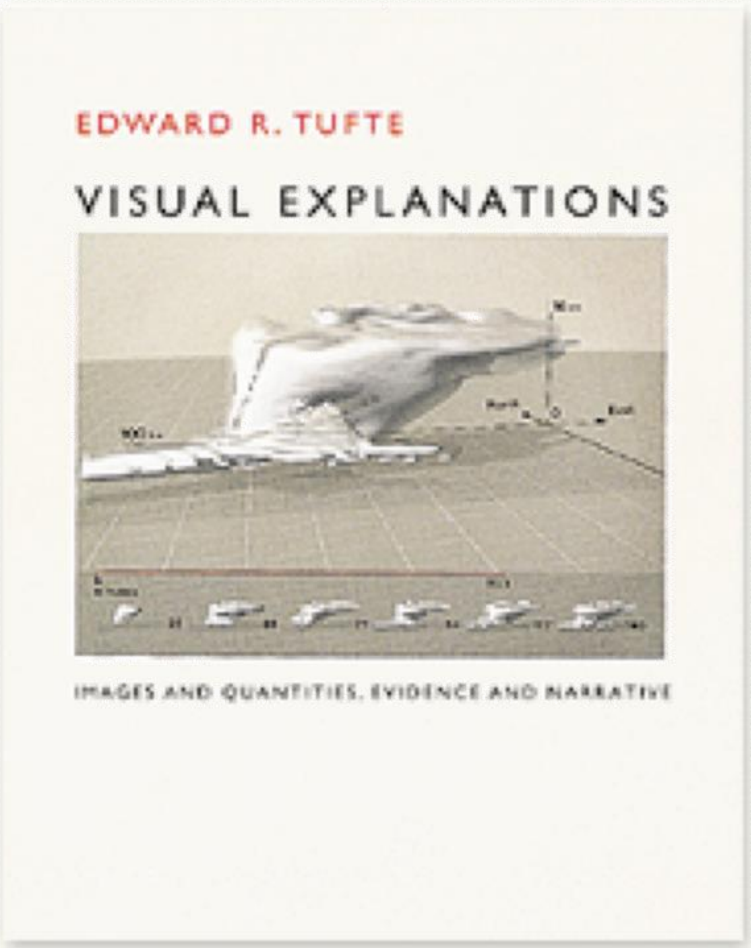
# Know your audience

- What do they know?
- What motivates them? What do they desire?
- What experiences do you share? What are common goals?
- What insights can you provide? What tools and “magical gifts”?

# Edward Tufte



# Edward Tufte's Principles of Graphical Excellence

Graphical excellence …

- is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design.
- consists of complex ideas communicated with clarity, precision and efficiency.
- is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- is nearly always multivariate.
- requires telling the truth.