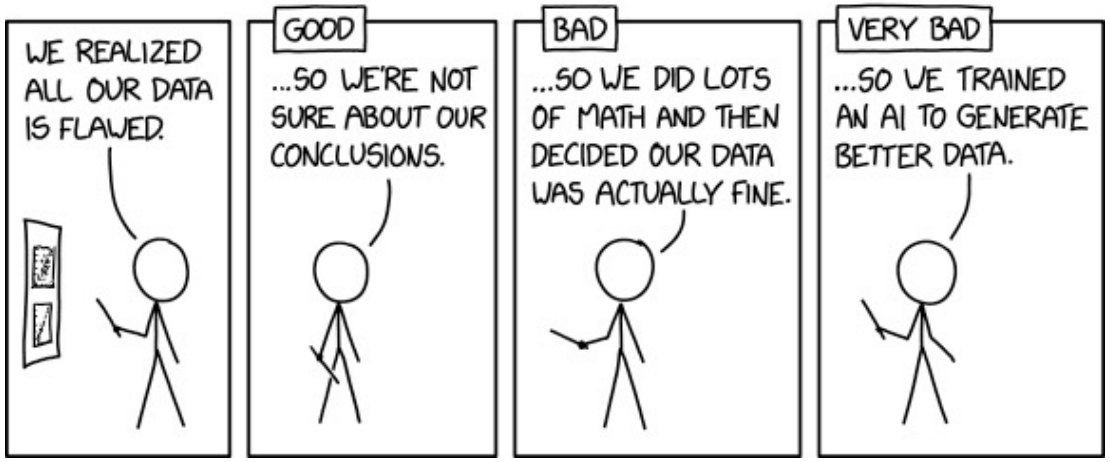


xkcd #2494

Exploratory Data Analysis

CS 4821 – CS 5831 – s24

some slides adapted from: G. Piatetsky-Shapiro; Han, Kamber, & Pei; P. Smyth; C. Volinsky; Tan, Steinbach, & Kumar; J. Taylor; G. Dong; A. Mueller, J. DeNero, and others

Outline

- Exploratory Data Analysis
  - Summary Statistics
- Visualization
  - Graphs Elements
  - Visual Properties
    - Pre-attentive Properties
    - Visual Illusions
  - Graphical Basics
  - Minimalist Principles

EDA and Visualization

- Exploratory Data Analysis (EDA) and Visualization are important steps in any data mining task
- Get to know your data!
  - Distributions (symmetric, skewed, type)
  - Data quality issues
  - Outliers
  - Correlations and inter-relationships
- EDA or a visualization may be the goal of the analysis

Summary Statistics

Sample statistics for a variable, *X*

Measures of Central Tendency

- Mean: "center of the data"

$$\bar{X} = \hat{\mu} = \frac{\sum_i X_i}{n}$$

- Median: 50% of values below *m*, 50% above *m*

$$P(X \leq m) \geq \frac{1}{2} \; and \; P(X \geq m) \geq \frac{1}{2}$$

- Mode: most common value

Summary Statistics

Sample statistics for a variable, *X*

Measures of Spread

- Variance: "spread of data"

$$\hat{\sigma}^2 = \frac{\sum_i (X_i - \overline{X}_i)^2}{n}$$

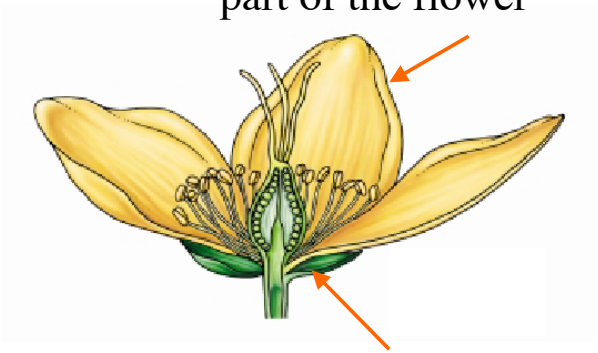
- Skewness: indication of non-symmetry
- Range: max – min value
- Quantiles/Quartiles/Percentiles:
- Five number summary:
  - Min, Q1, median (Q2), Q3, max

Example of Summary Statistics

- A running example using the Iris data
  - Data set available at the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/
  - Data set created by statistician R.A. Fisher

Data Properties $X_{150\times5}$

- 4 attributes
  - Sepal width and length
  - Petal width and length
- 1 class (3 flower types)
  - Setosa
  - Virginica
  - Versicolous



Iris Data

Table 1: Iris Data

sepal.length	sepal.width	petal.length	petal.width	iris.type
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa

Summary Statistics: Iris Data

Five number summary:

	Min.	1st Quartile	Median	3rd Quartile	Max.
sepal.length	4.3	5.1	5.80	6.4	7.9
sepal.width	2.0	2.8	3.00	3.3	4.4
petal.length	1.0	1.6	4.35	5.1	6.9
petal.width	0.1	0.3	1.30	1.8	2.5

Relationships between 2 Variables

- Covariance and Correlation measure **linear** dependence
- Covariance depends on ranges of *X* and *Y*, standardize variables (divide by st. dev.)
- For two variables *X* and *Y* over *n* samples with values: *x*(1), …, *x*(n) and *y*(1), …, *y*(n), then

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \bar{x})(y(i) - \bar{y})$$

- Correlation = scaled covariance [-1, 1]

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i) - \bar{x})(y(i) - \bar{y})}{\left(\sum_{i=1}^{n}(x(i) - \bar{x})^2 \sum_{i=1}^{n}(y(i) - \bar{y})^2\right)^{\frac{1}{2}}}$$

Example: Anscombe Data

Four Data Sets with Identical Linear Models (Anscombe's Quartet)

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

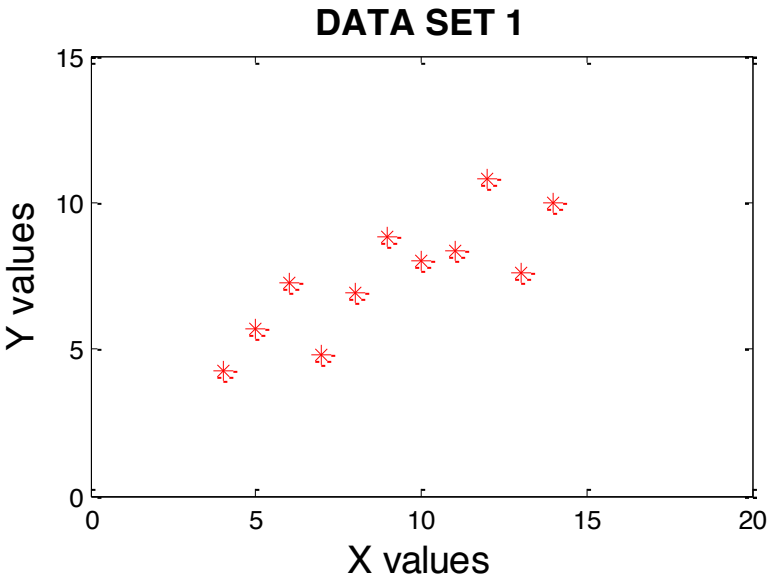
Tufte, Edward (1983) Visual Display Of Quantitative Information

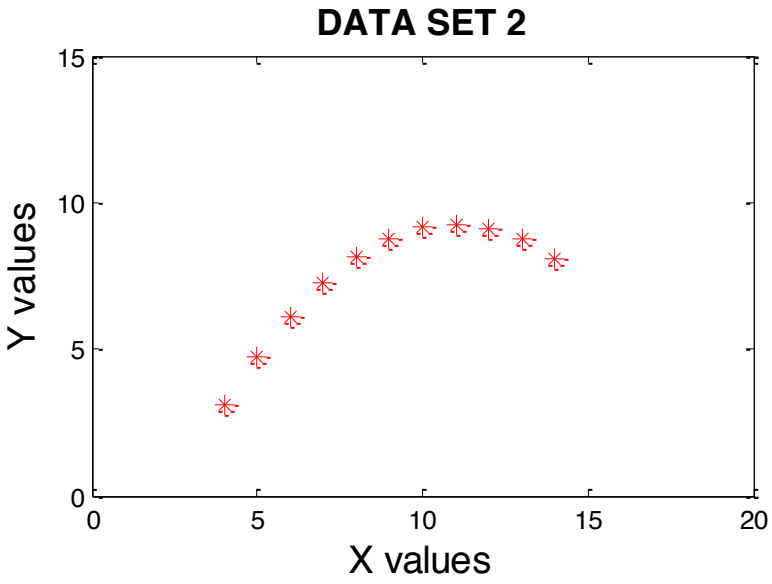
Slide adapted from: C. Aragon

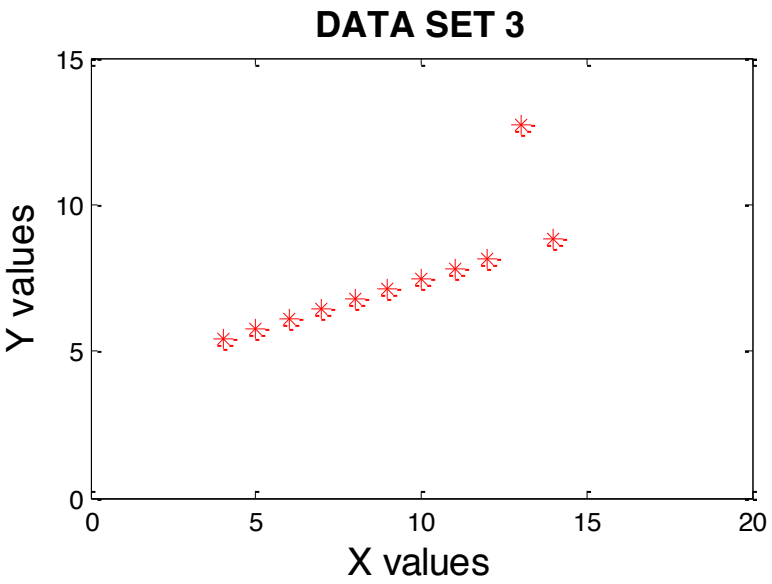
Example: Anscombe Data

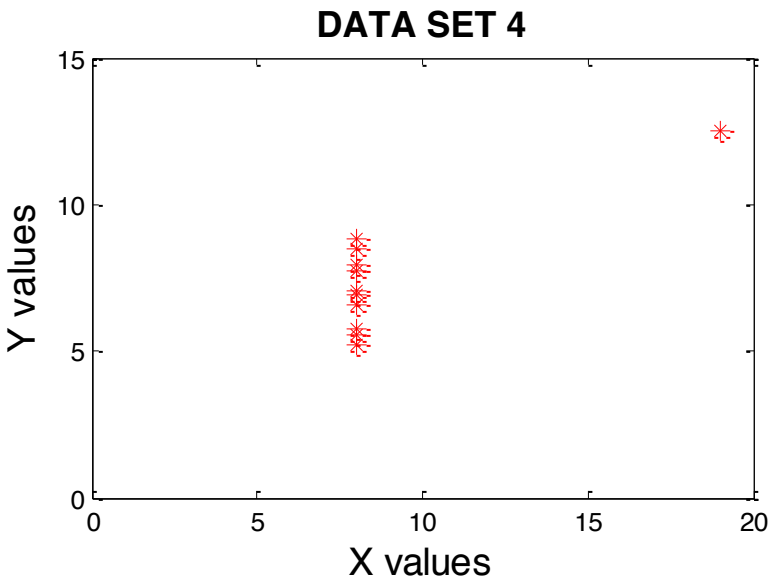
- Summary Statistics of Data 1
  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Correlation = 0.82
- Summary Statistics of Data 2
  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Correlation = 0.82
- Summary Statistics of Data 3
  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Correlation = 0.82
- Summary Statistics of Data 4
  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Correlation = 0.82

Example: Anscombe Data









Examples of Correlation

- Linear Dependence

- Non-Linear Dependence

Visualizations

- In addition to summary statistics, different visualizations may be helpful for understanding your data.
- Humans have a well developed ability to analyze large amounts of information when presented visually to detect patterns, anomalies, etc.

- We will look at:
  - Principles of Visualization
  - Types of Graphs

What is Information Visualization?

“Transformation of the symbolic into the geometric”

(McCormick et al., 1987)

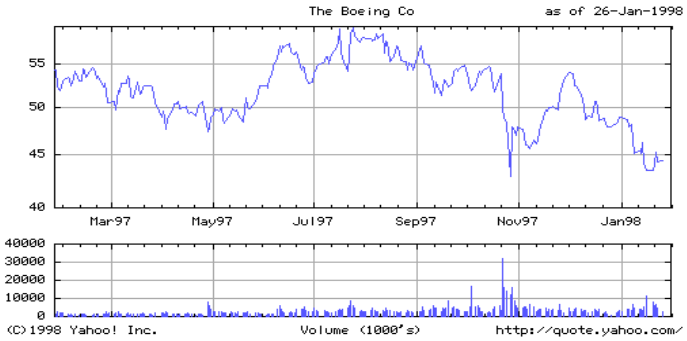
“... finding the artificial memory that best supports our natural means of perception."

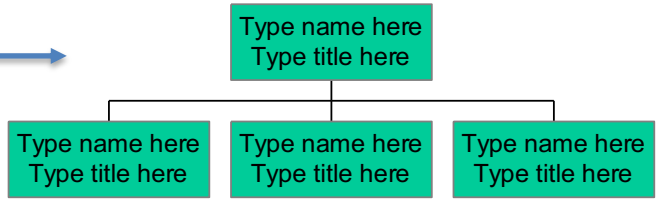
(Bertin, 1983)

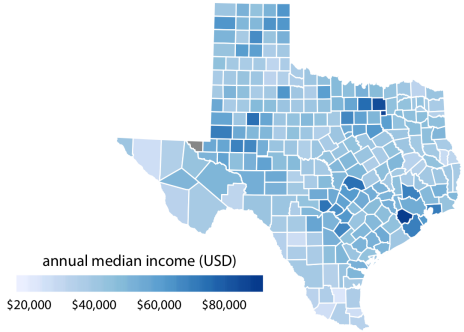
The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system.

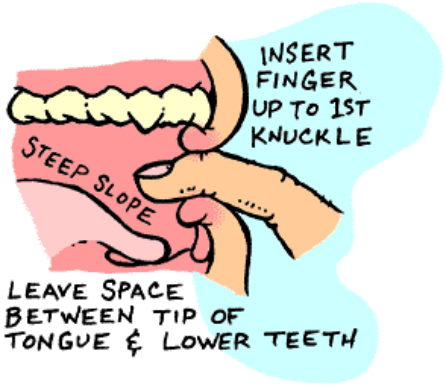
Types of Symbolic Displays

- Graphs
- Charts
- Maps
- Diagrams









Goals of Viz

Visualizations should:

- Make large datasets coherent
  Present huge amounts of information compactly
- Present information from various viewpoints
- Present information at several levels of detail
  from overviews to fine structure
- Support visual comparisons
- Tell stories about the data

VISUALIZATIONS

Elements of Graphs

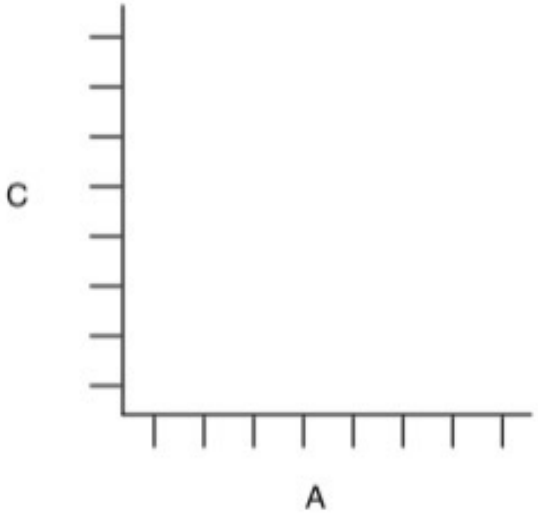
Anatomy of a Graph

- Framework
  - Sets the stage
  - Kinds of measurements, scale, …
- Content
  - Marks
  - Point symbols, lines, areas, bars, …
- Labels
  - Title, axes, tic marks, …

Elements of a Plot

Geometric Objects

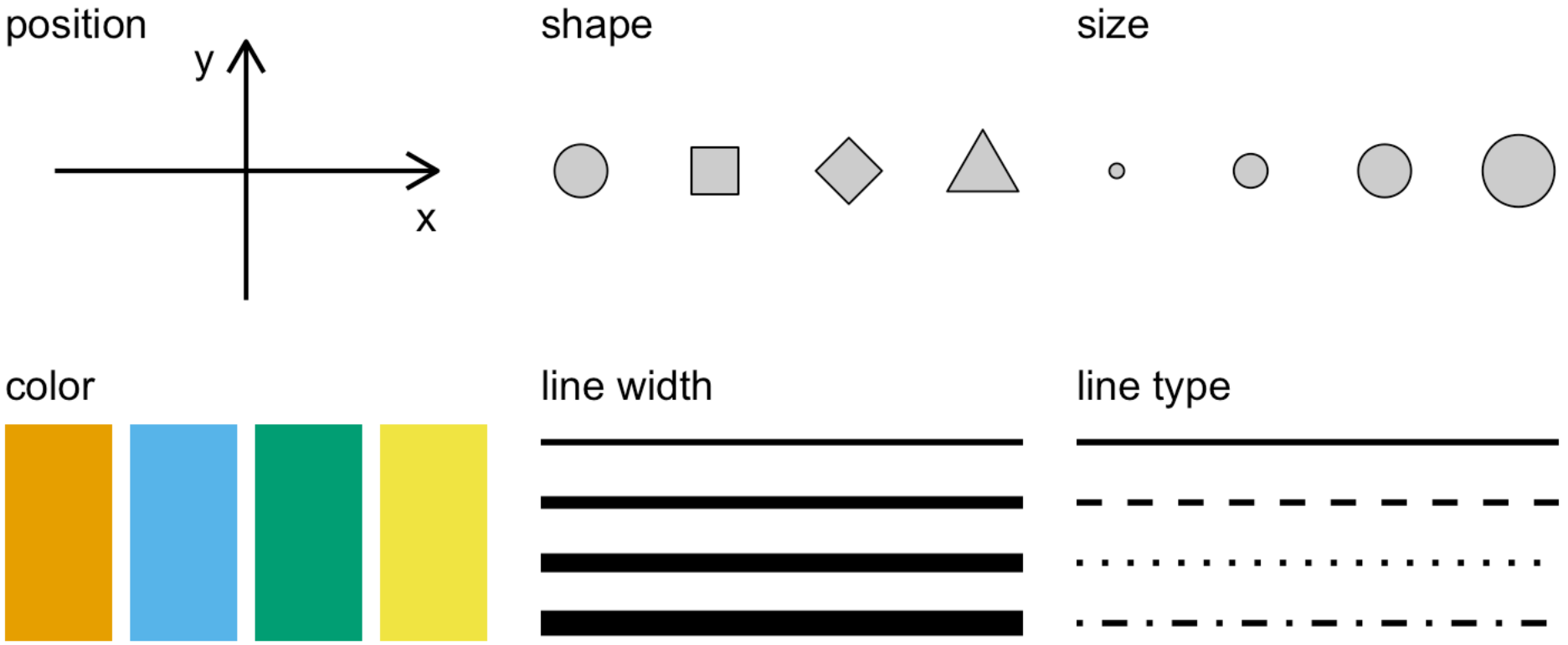
Scales & Coordinates



Annotations

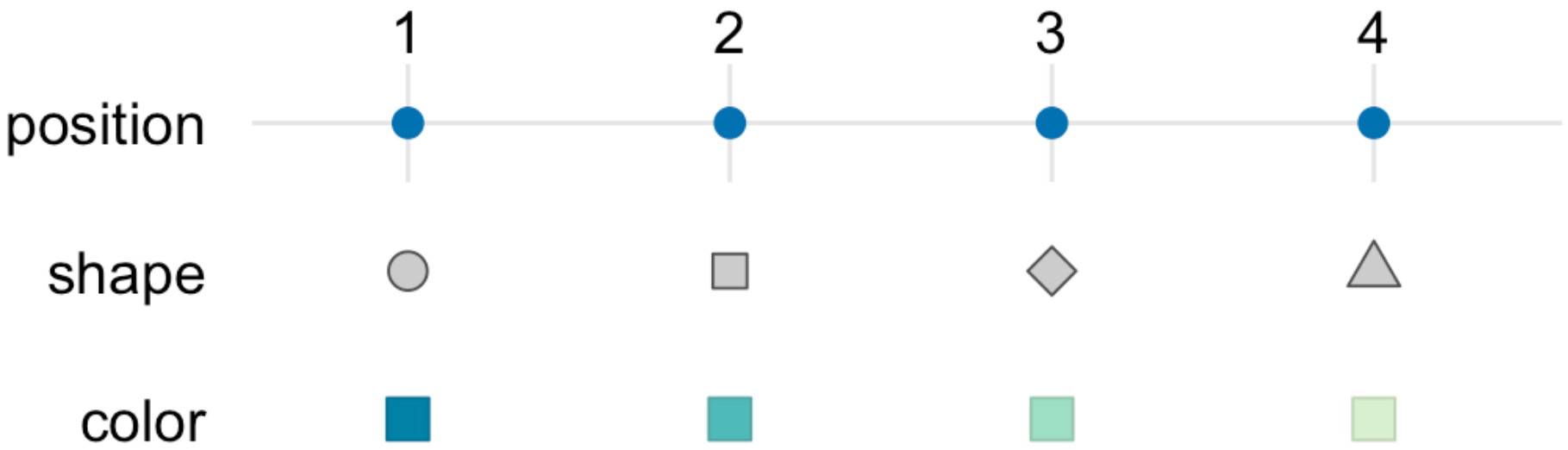
Wickham 2010

Aesthetics of a Plot



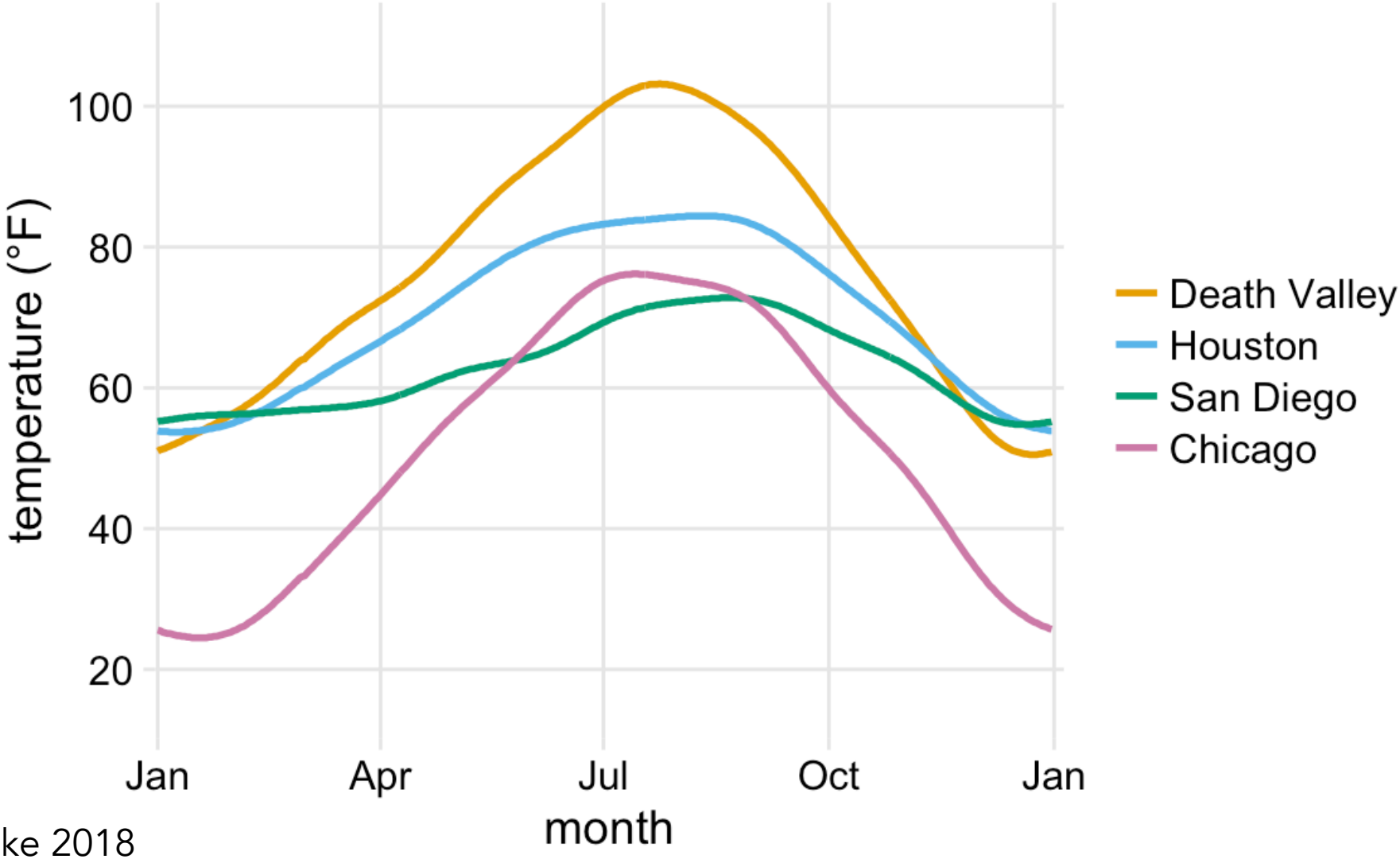
Wilke 2018

Aesthetics Map Data to Visual Representation



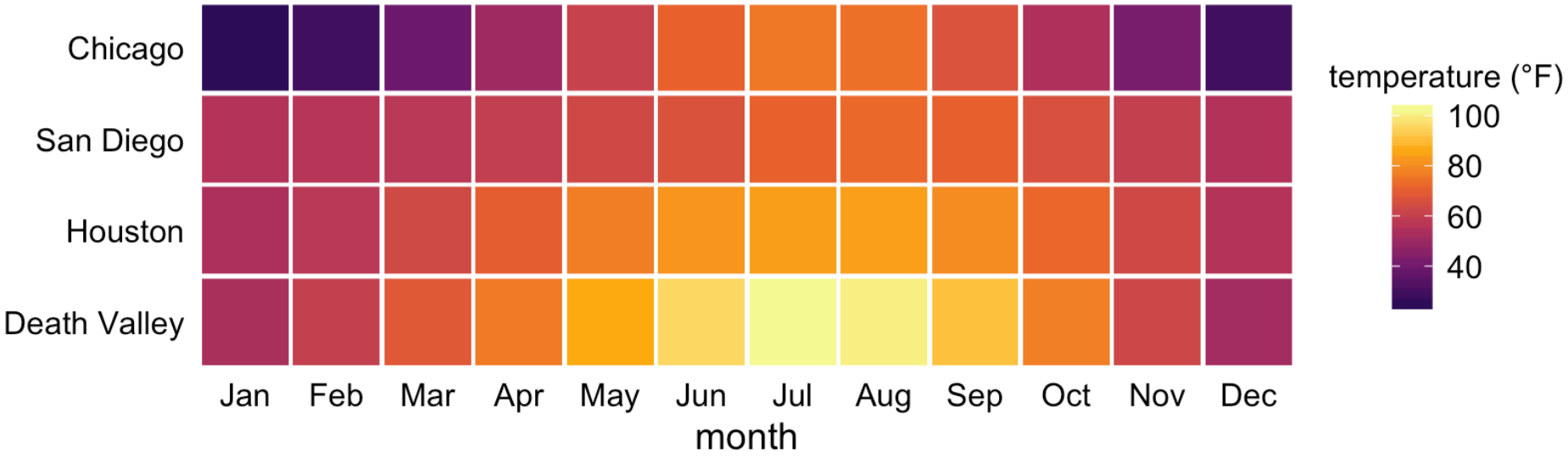
Wilke 2018

Example: Daily Temp.



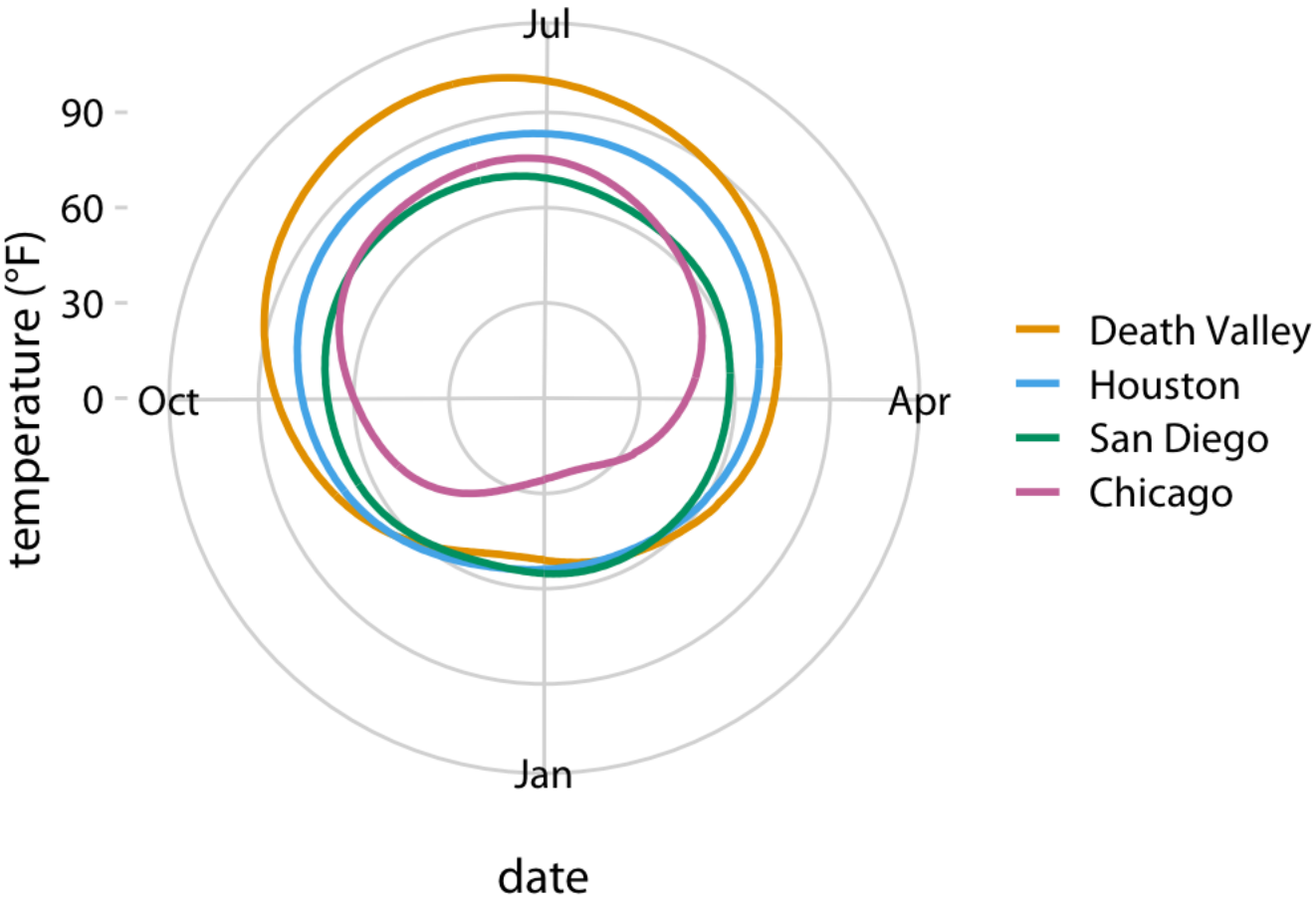
Wilke 2018

Example: Daily Temp.



Wilke 2018

Example: Daily Temp.



Wilke 2018

VISUAL PROPERTIES

Preattentive Processing

Visual Illusions

Preattentive Processing

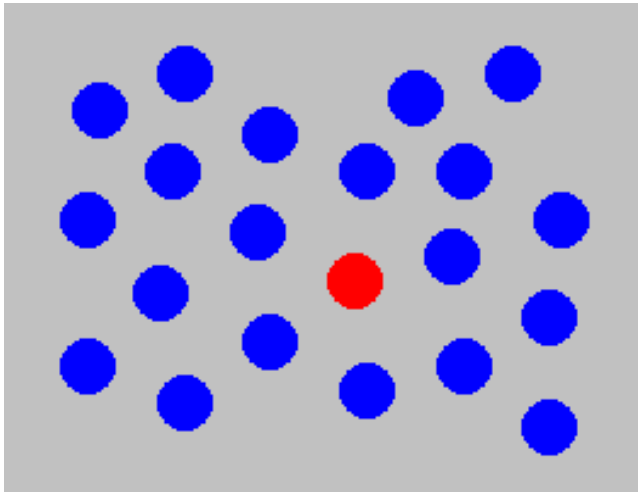
- A limited set of visual properties are processed preattentively
  - without need for focusing attention
- This is important for design of visualizations
  - what can be perceived immediately
  - what properties are good discriminators
  - what can mislead viewers

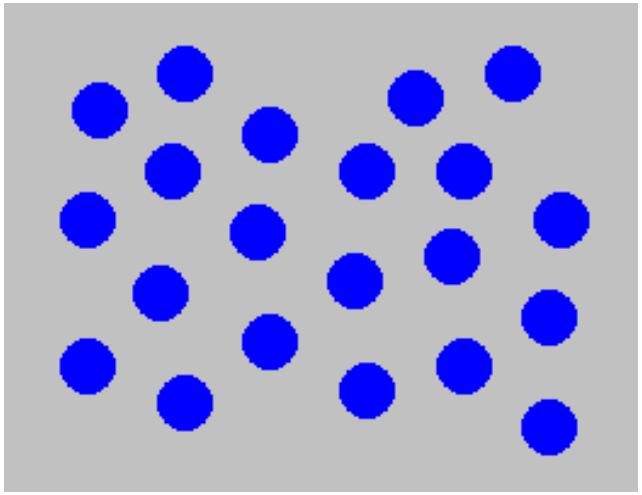
Preattentive Processing

- < 200 - 250ms qualifies as pre-attentive
  - eye movements take at least 200ms
  - yet certain processing can be done very quickly, implying low-level processing in parallel

- If a decision takes a fixed amount of time regardless of the number of distractors, it is preattentive

Example: Color Selection

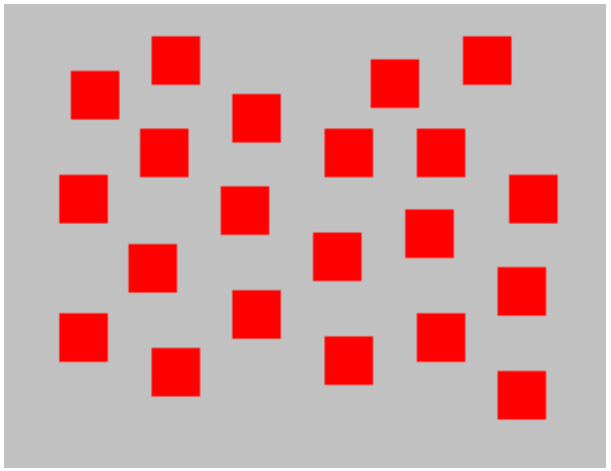


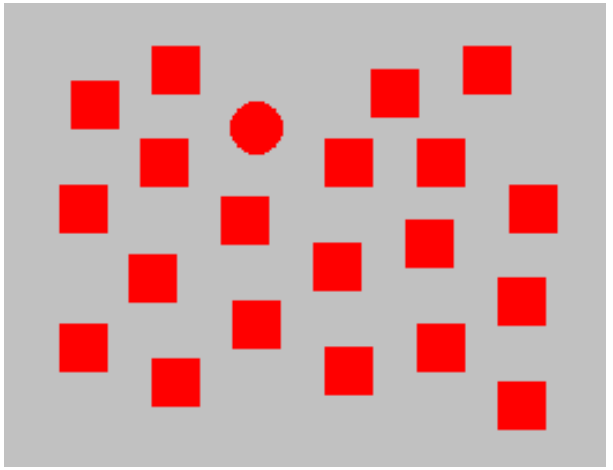


Viewer can rapidly and accurately determine whether the target (red circle) is present or absent.

Difference detected in color.

Example: Shape Selection



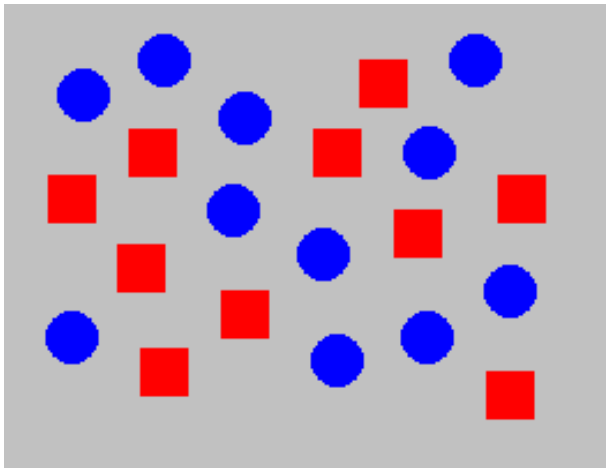


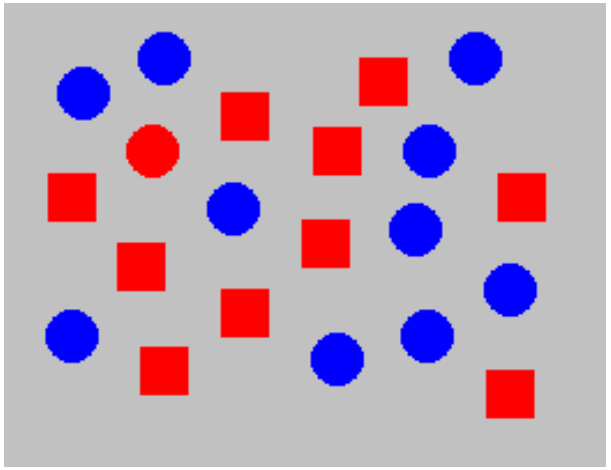
Viewer can rapidly and accurately determine whether the target (red circle) is present or absent.

Difference detected in form (curvature)

Healey 97

Example: Conjunction of Features





Viewer ***cannot*** rapidly and accurately determine whether the target (red circle) is present or absent when target has two or more features, each of which are present in the distractors. Viewer must search sequentially.

Healey 97

Gestalt Properties/Principles

- Law of Perceptual Organization
- Law of Proximity
- Law of Similarity
- Law of Common Fate
- Connectedness
- Continuity

Example: Continuity

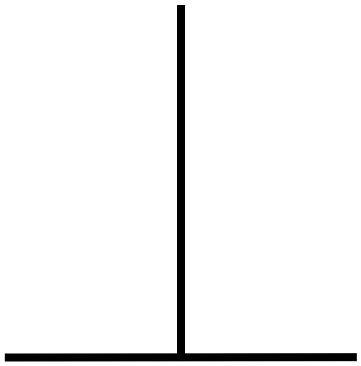
Visual Illusions

- People don't perceive length, area, angle, brightness they way they "should"
- Some illusions have been reclassified as systematic perceptual errors
  - e.g., brightness contrasts (grey square on white background vs. on black background)
- Nevertheless, the visual system does some unexpected things

Illusions of Linear Extent

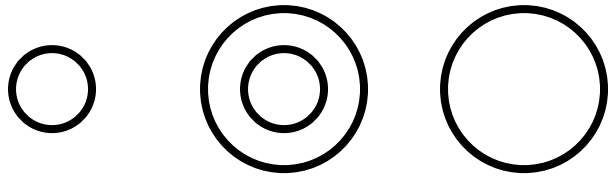
- Mueller-Lyon (off by 25-30%)

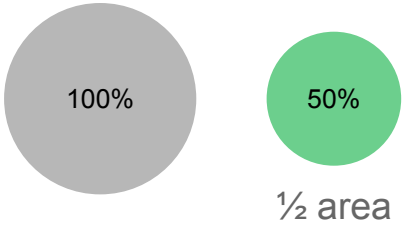
- Horizontal-Vertical

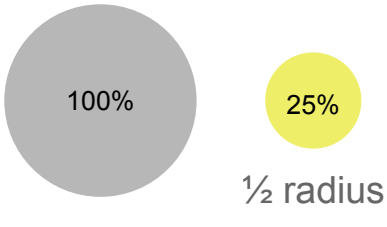


Illusions of Area

- Delboeuf Illusion



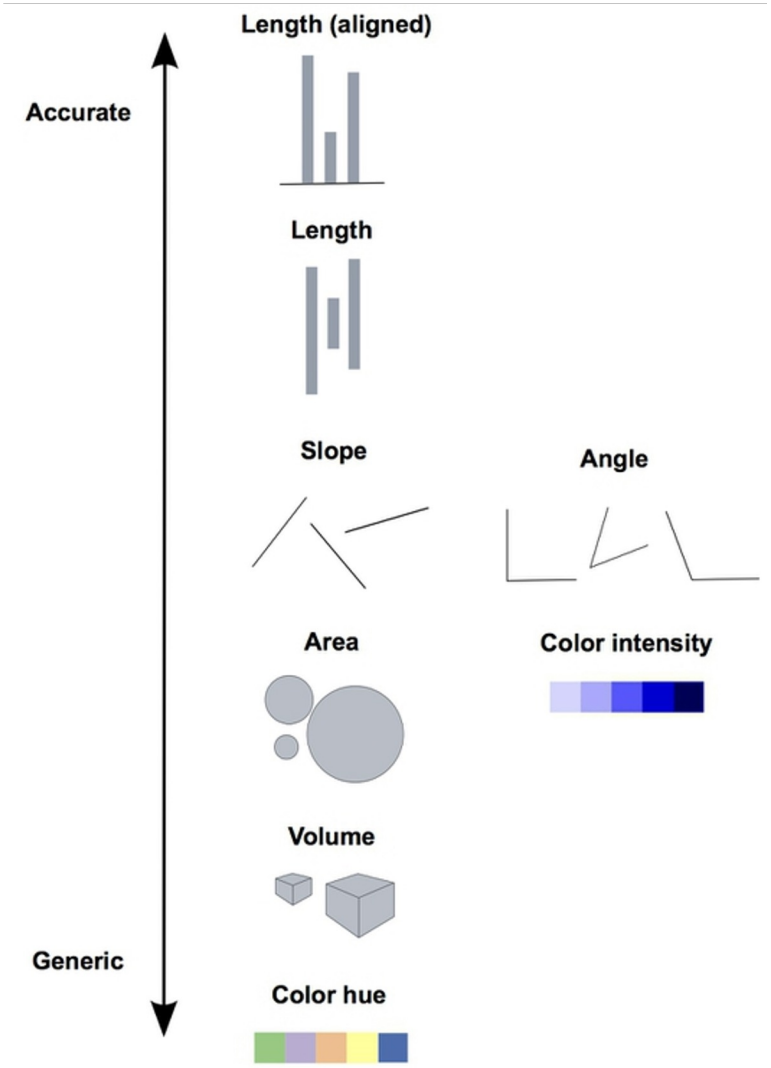




Visual Channels

Ranking of Quantitative Perceptual Tasks

- Accuracy of judgements depend on type of mark.
- Aligned lengths most accurate
- Color least accurate

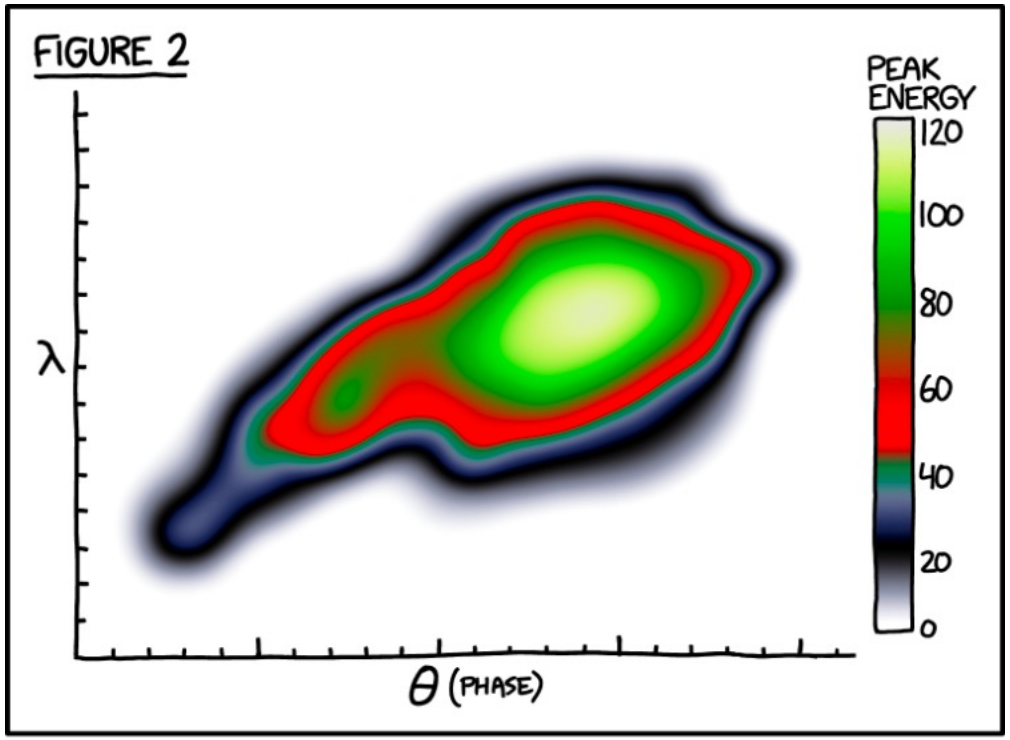


Mackinlay 88 from Cleveland & McGill

Ranking of Properties by Data Type

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density	Color Hue
Angle	Color Saturation	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color Saturation
Color Saturation	Length	Shape
Color Hue	Angle	Length

(Mackinlay 88, Not Empirically Verified)



EVERY YEAR, DISGRUNTLED SCIENTISTS COMPETE FOR THE PAINBOW AWARD FOR WORST COLOR SCALE.

xkcd #2537

GRAPH BASICS

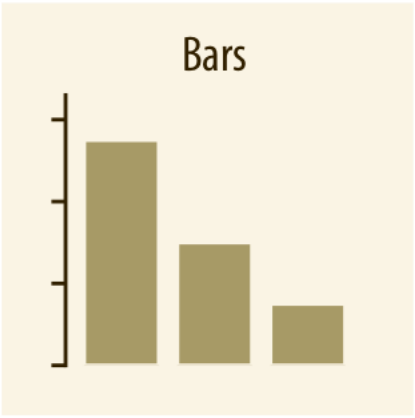
Overview of Visualizations

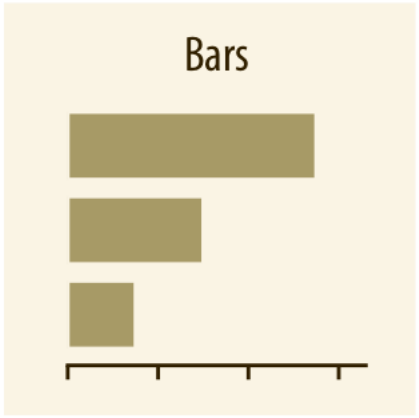
- Amounts
- Distributions
- Proportions
- x-y relationships
- Geospatial data
- Uncertainty

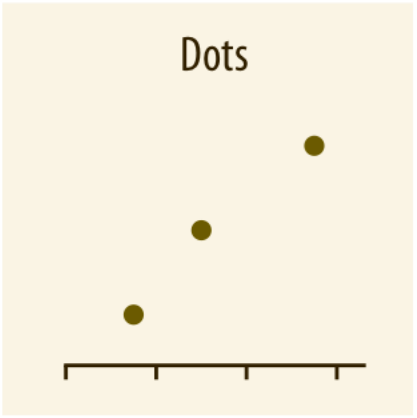
Examine these types of graphs for

- single attribute
- multiple attributes

Amounts





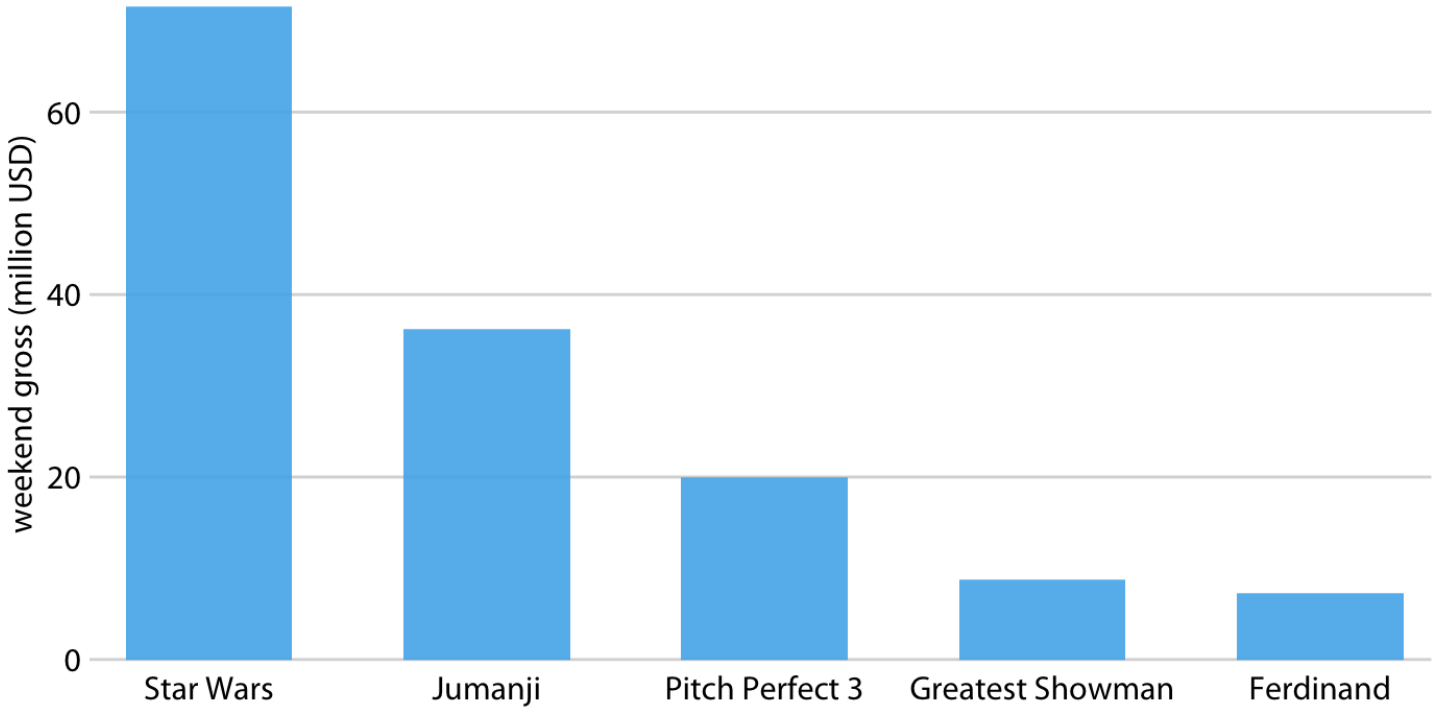


Wilke 2018

Bar Plots

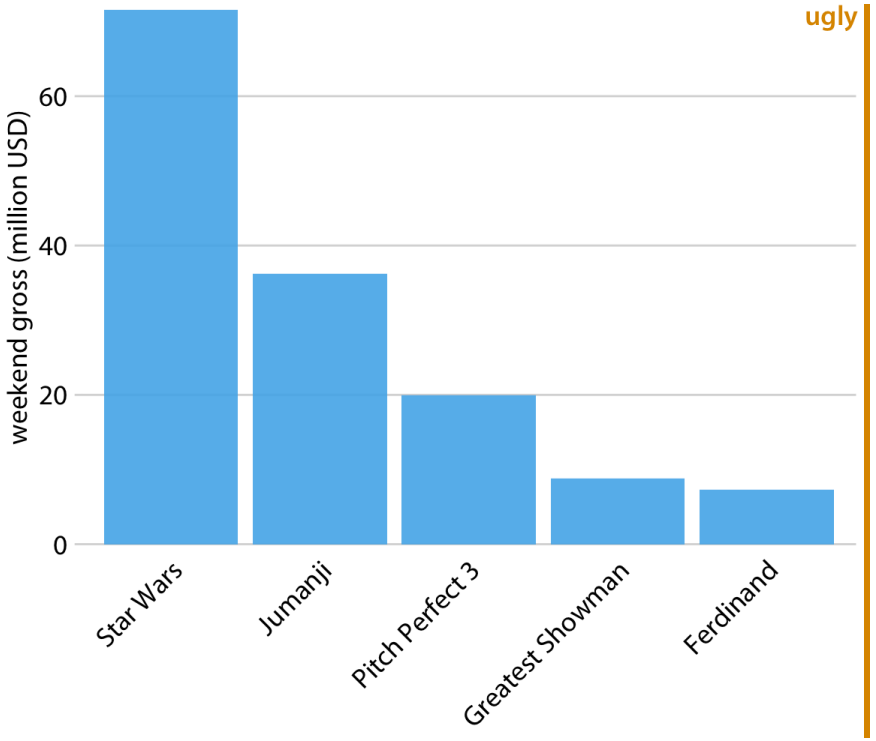
Highest grossing movies, Dec. 22-24, 2017

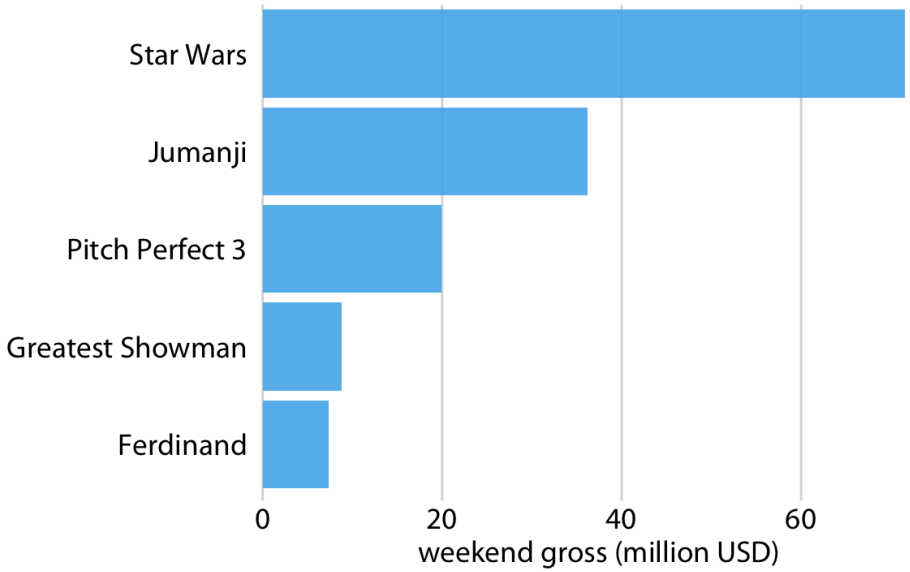
Title	Gross ($)
Star Wars: The Last Jedi	71,565,498
Jumanji: Welcome to the Jungle	36,169,328
Pitch Perfect 3	19,928,525
The Greatest Showman	8,805,843
Ferdinand	7,316,746



Wilke 2018

Issues with Bar Plots

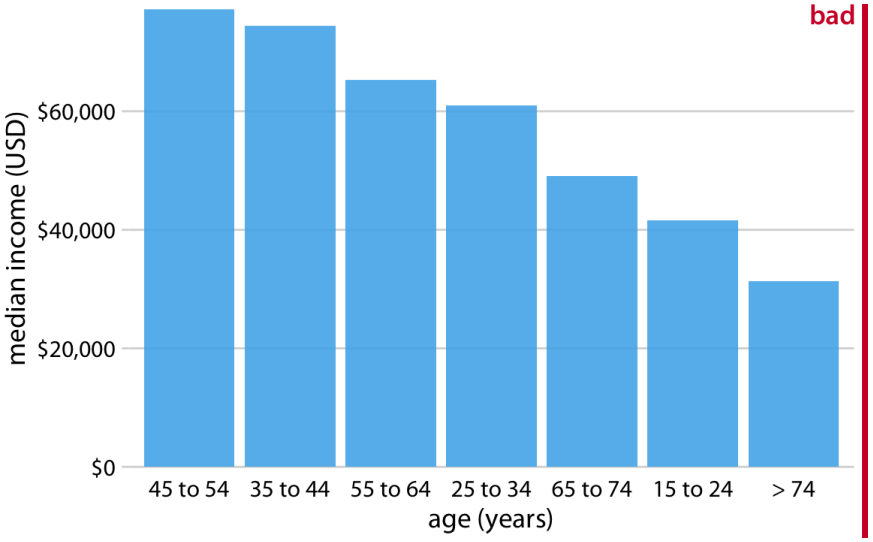




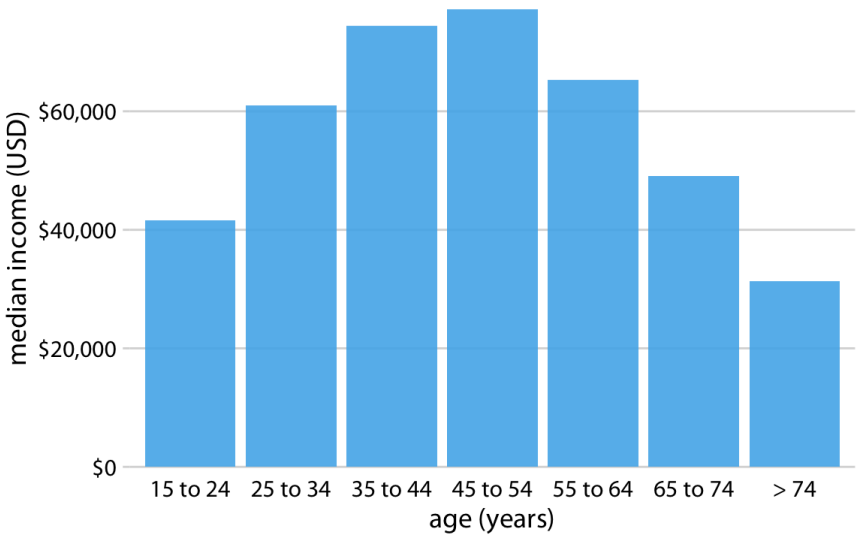
Wilke 2018

Issues with Bar Plots

Median U.S. annual household income versus age group



Median U.S. annual household income versus age group



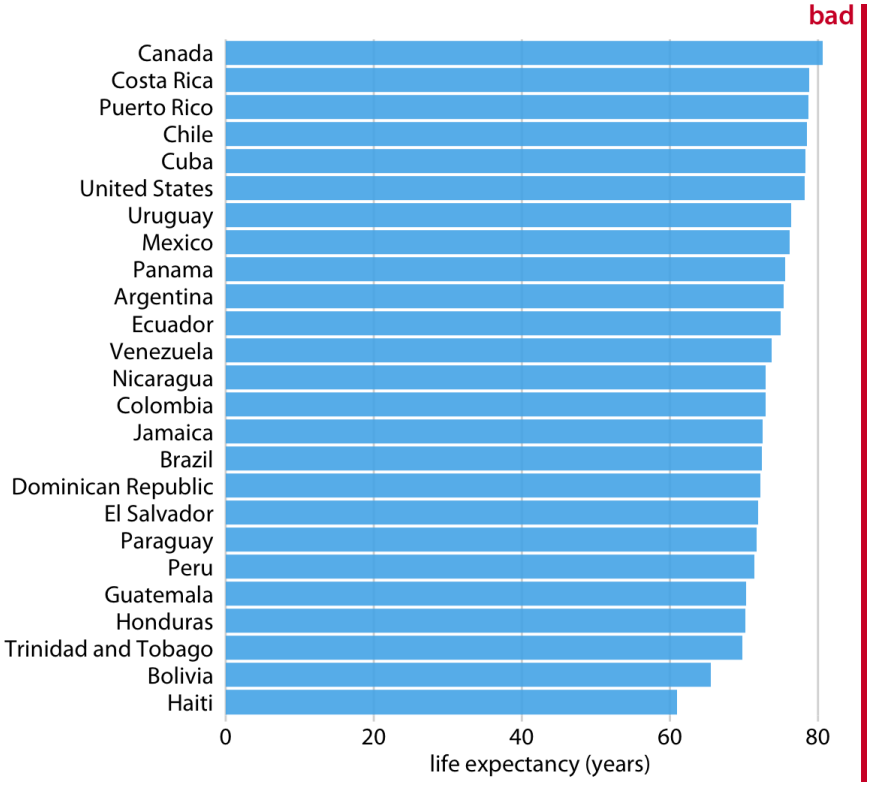
Wilke 2018

Dot Plots

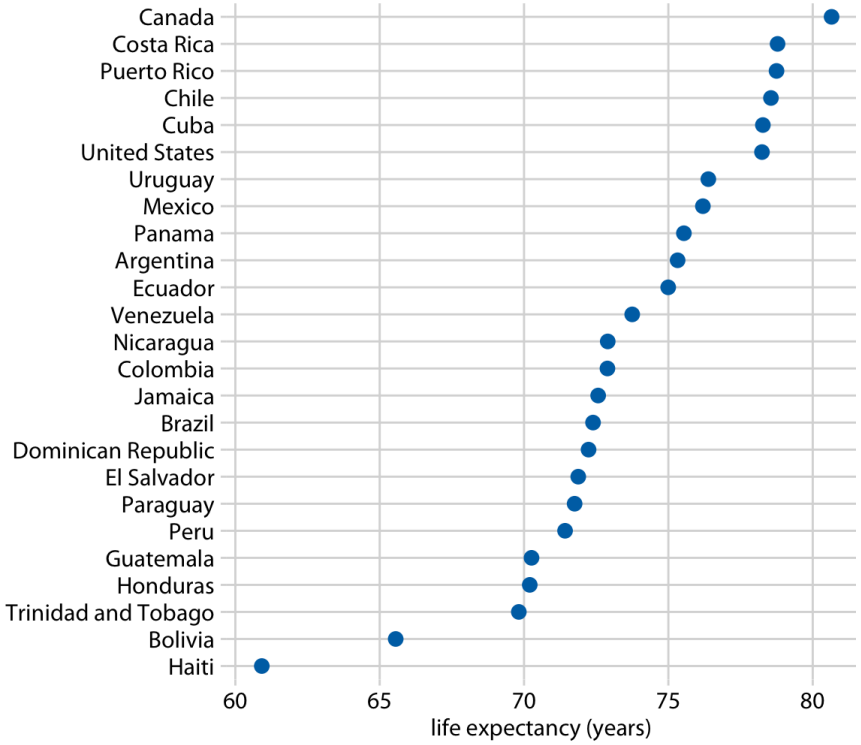
- Bar plots should have the bars start at 0, so that the bar length is proportional to the amount shown
- Dot plots instead place a dot along the axis which may have limited range (not start at 0)
  - They may allow features of the data to become easier to observe

Dot Plots

Life expectancies of countries in the Americas, 2007

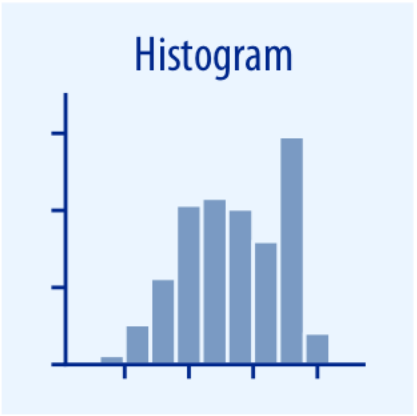


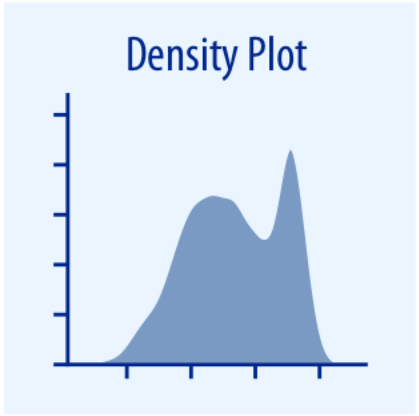
Life expectancies of countries in the Americas, 2007

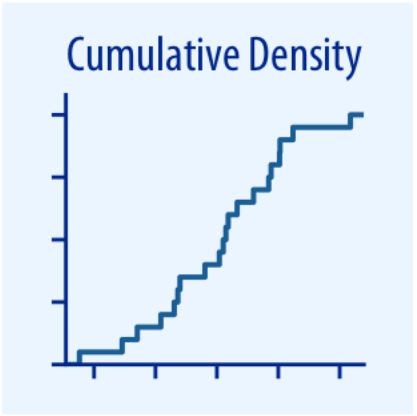


Wilke 2018

Distributions

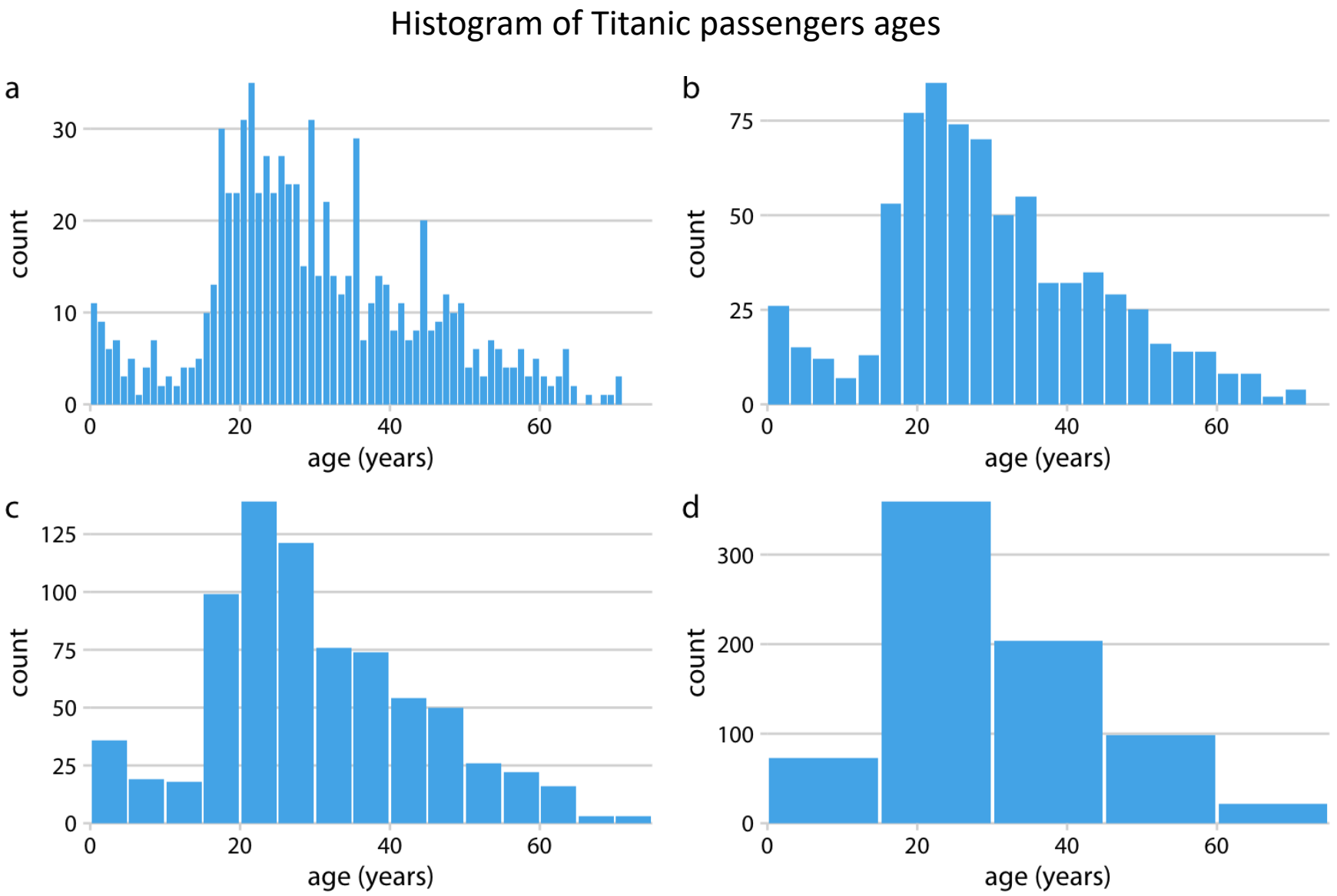






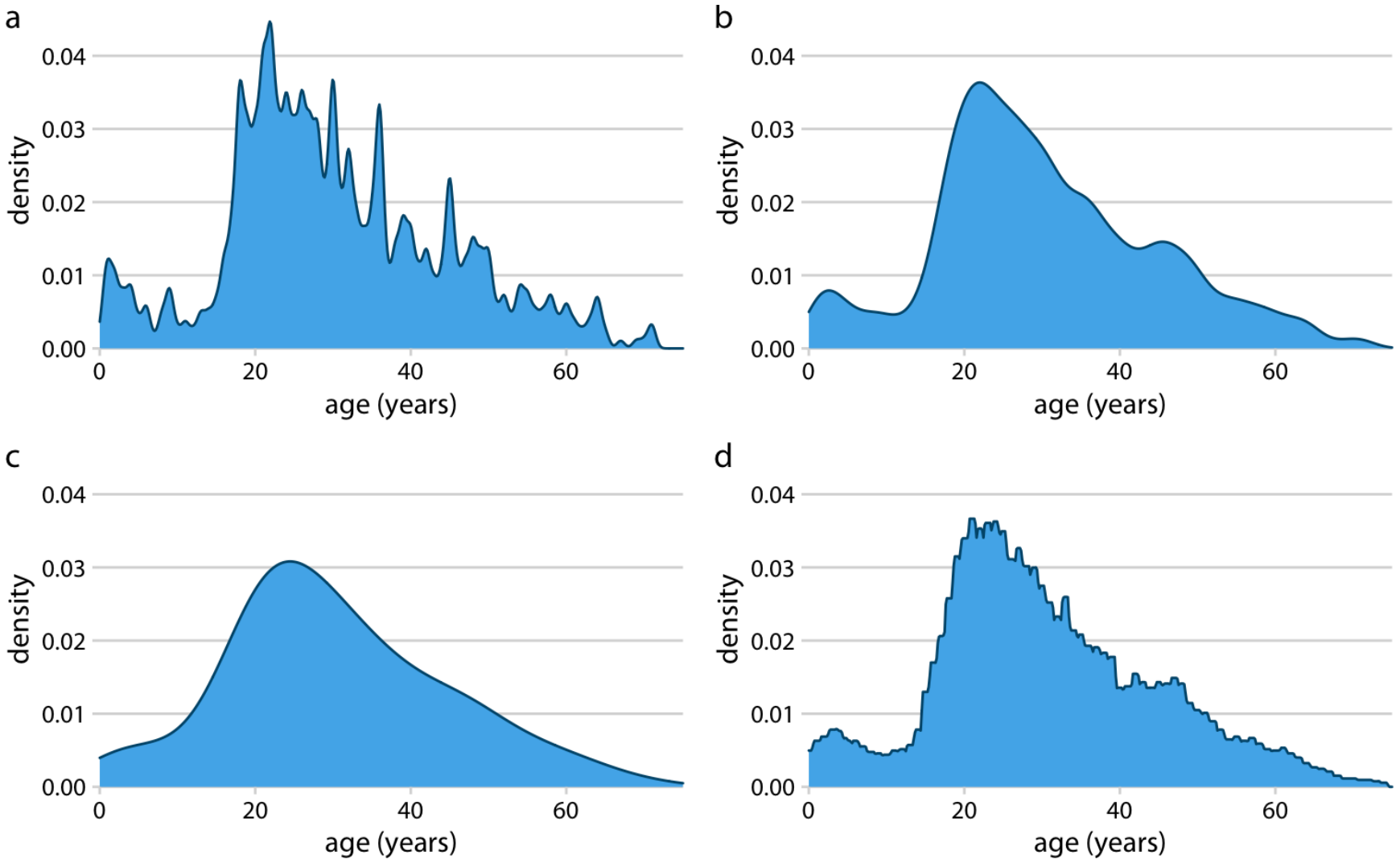
Wilke 2018

Histogram

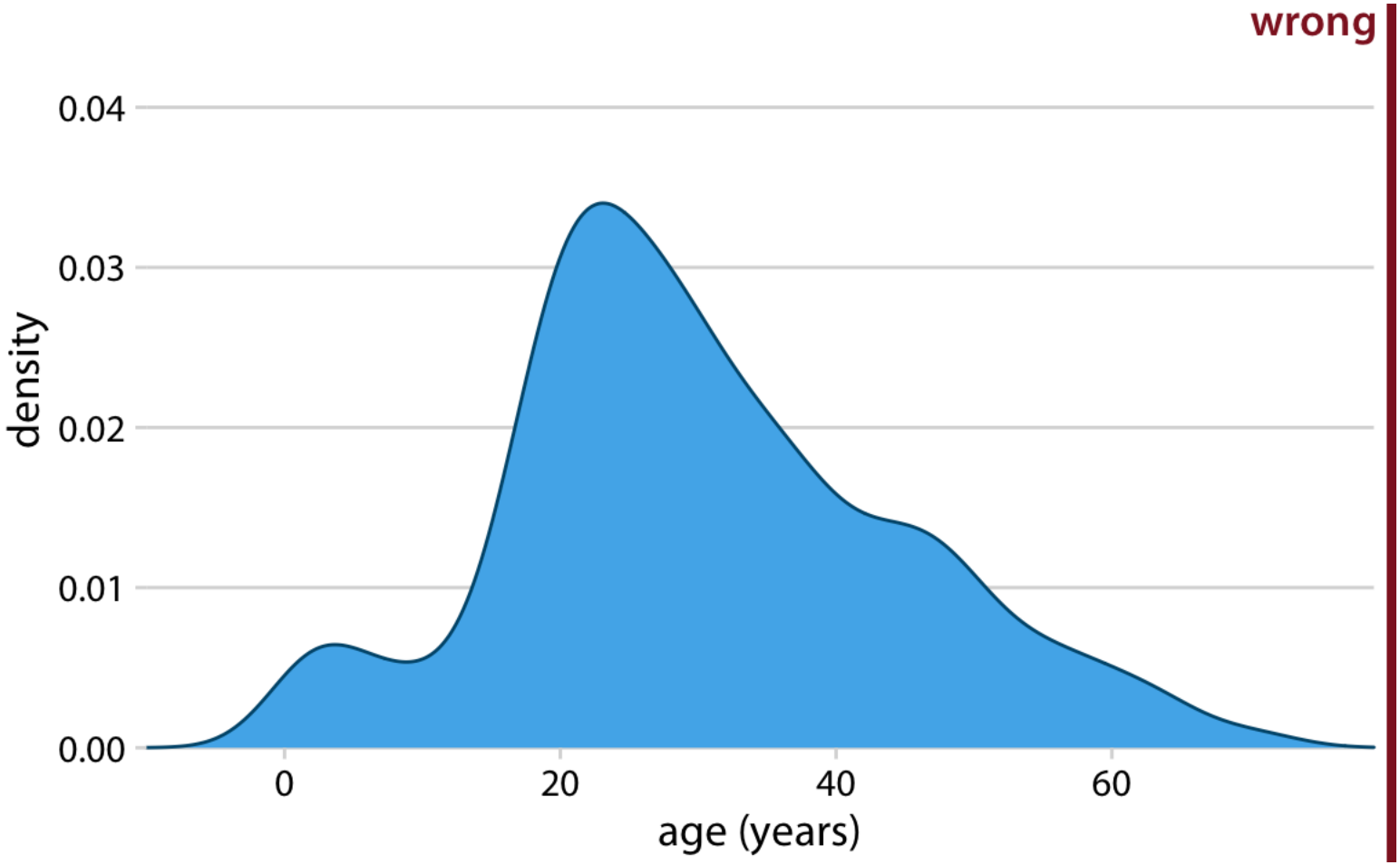

Histogram of Titanic passengers ages
a
count
30
20
10
0
0
20
40
60
age (years)
b
count
75
50
25
0
0
20
40
60
age (years)
c
count
125
100
75
50
25
0
0
20
40
60
age (years)
d
count
300
200
100
0
0
20
40
60
age (years)

Density Plot

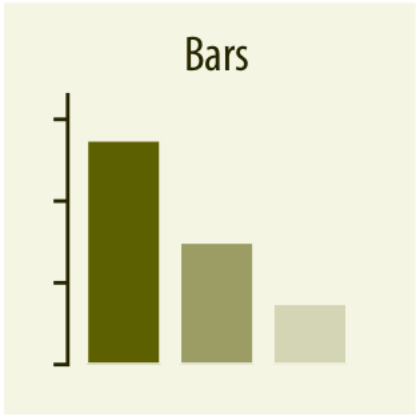
Kernel density estimates of Titanic passengers ages

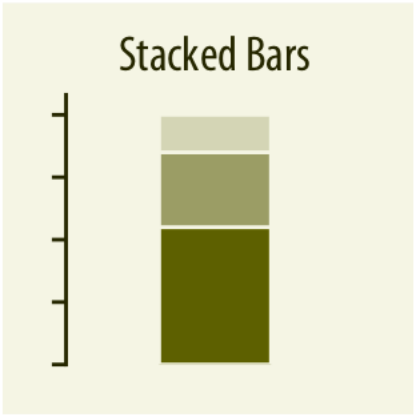


Issues with Density Plots



Proportions

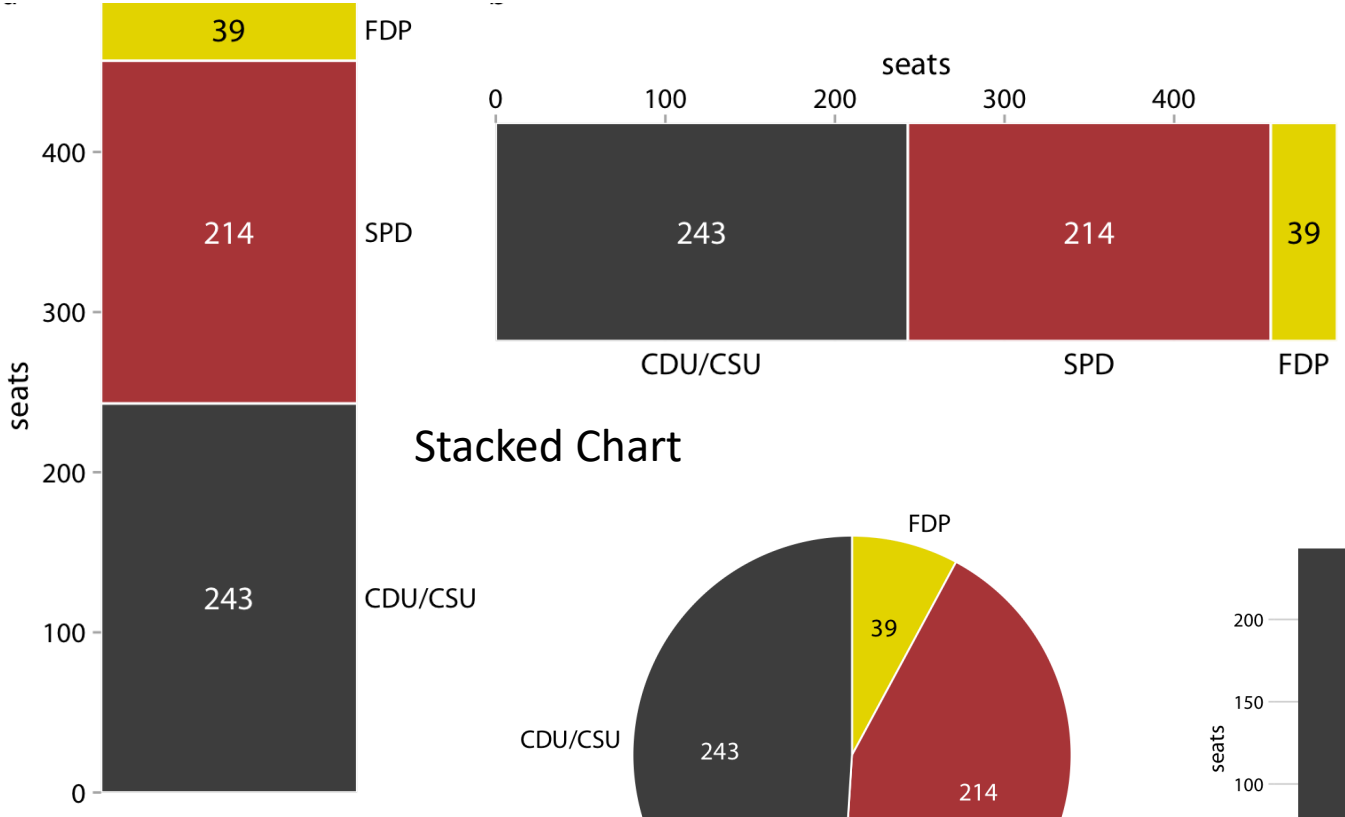




Wilke 2018

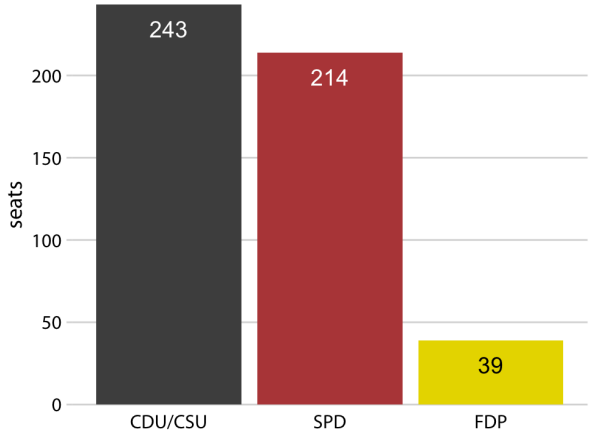
Proportion Examples

Party composition of the 8th German Bundestag, 1976–1980



Stacked Chart

Pie Chart



Bar Chart

Geospatial Data

Map

Choropleth

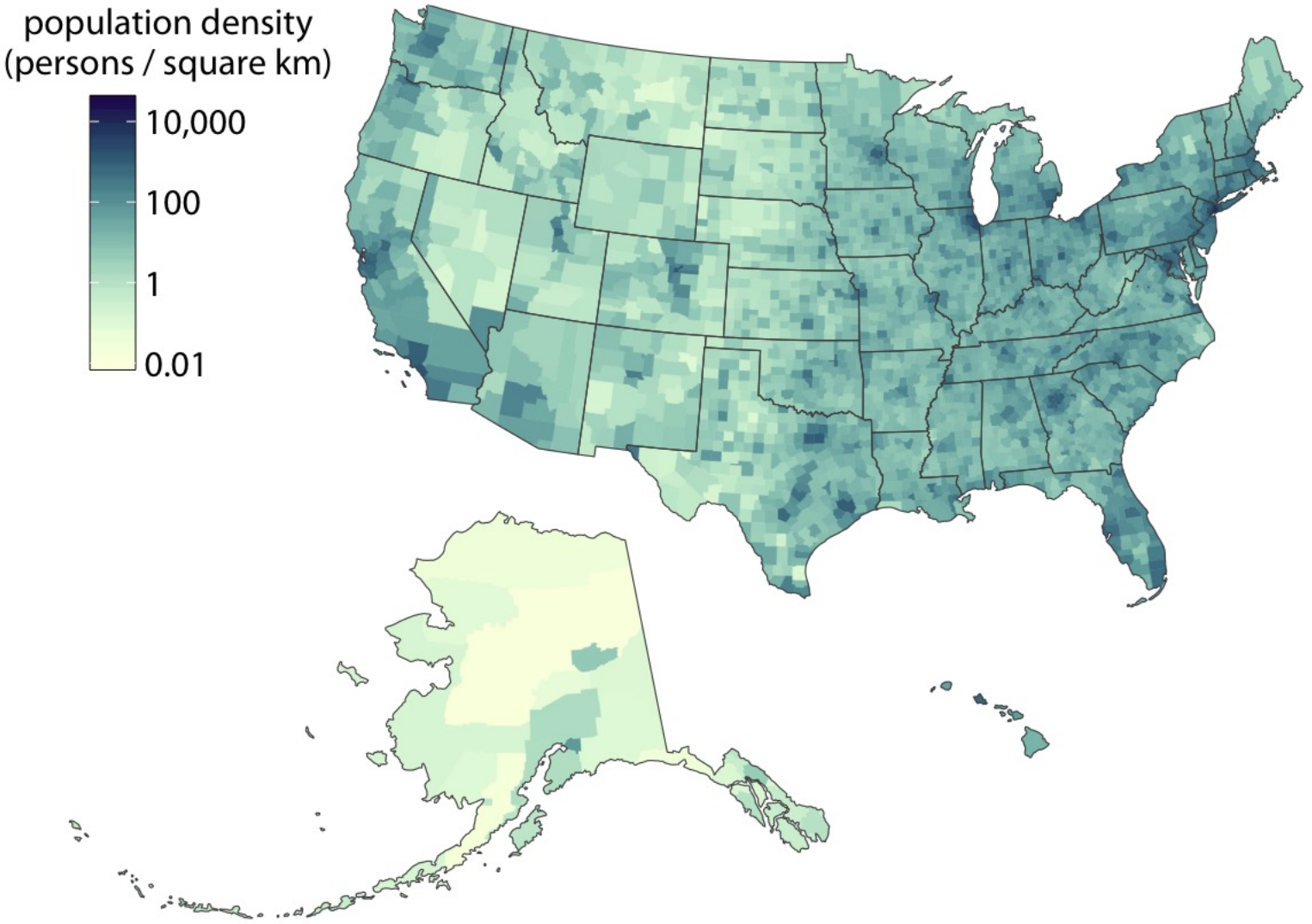
Cartogram

Cartogram Heatmap

Wilke 2018

Choropleth Maps

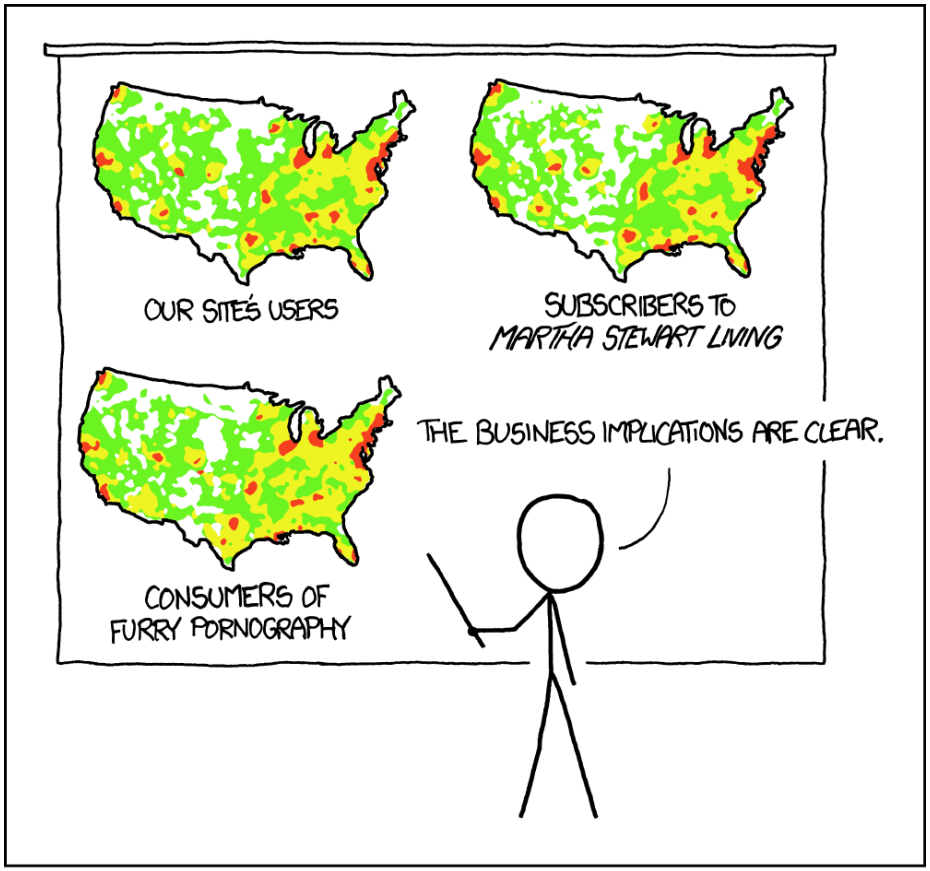
Population density in every U.S. county, 2015



Example: Dialect

Personal Dialect Map interactive quiz and visualization was published on the NYTimes in Dec. 2013.

The questions come from the Harvard Dialect Survey (Vaux, Bert and Scott Golder, 2003)

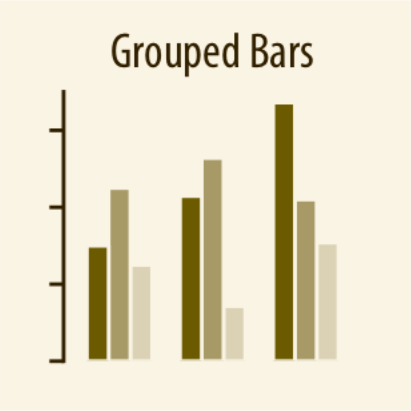


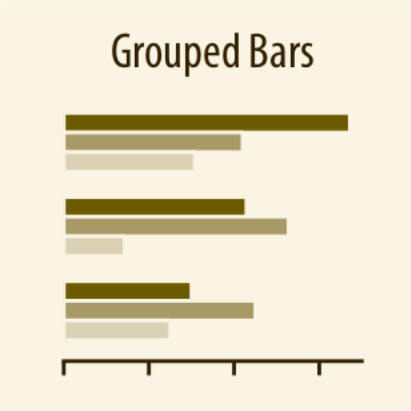
xkcd #1138

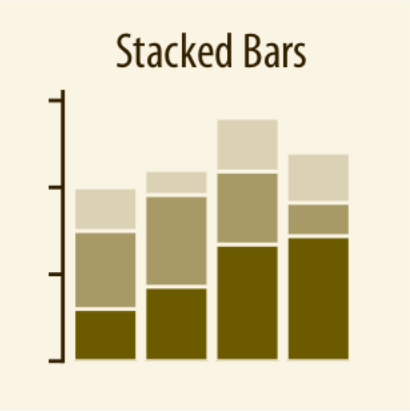
GRAPH BASICS

Multiple Attributes

Amounts – Two or More

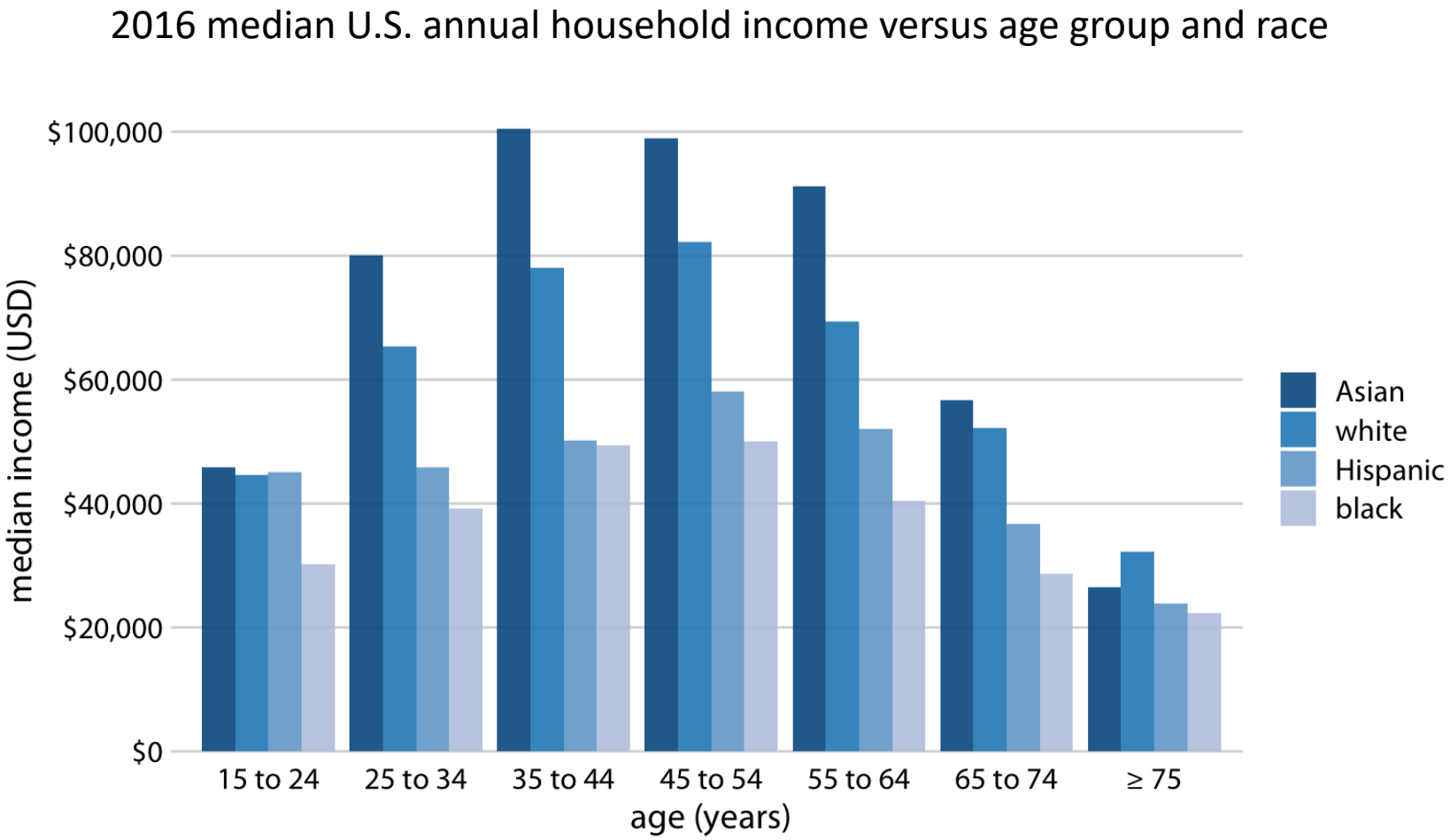




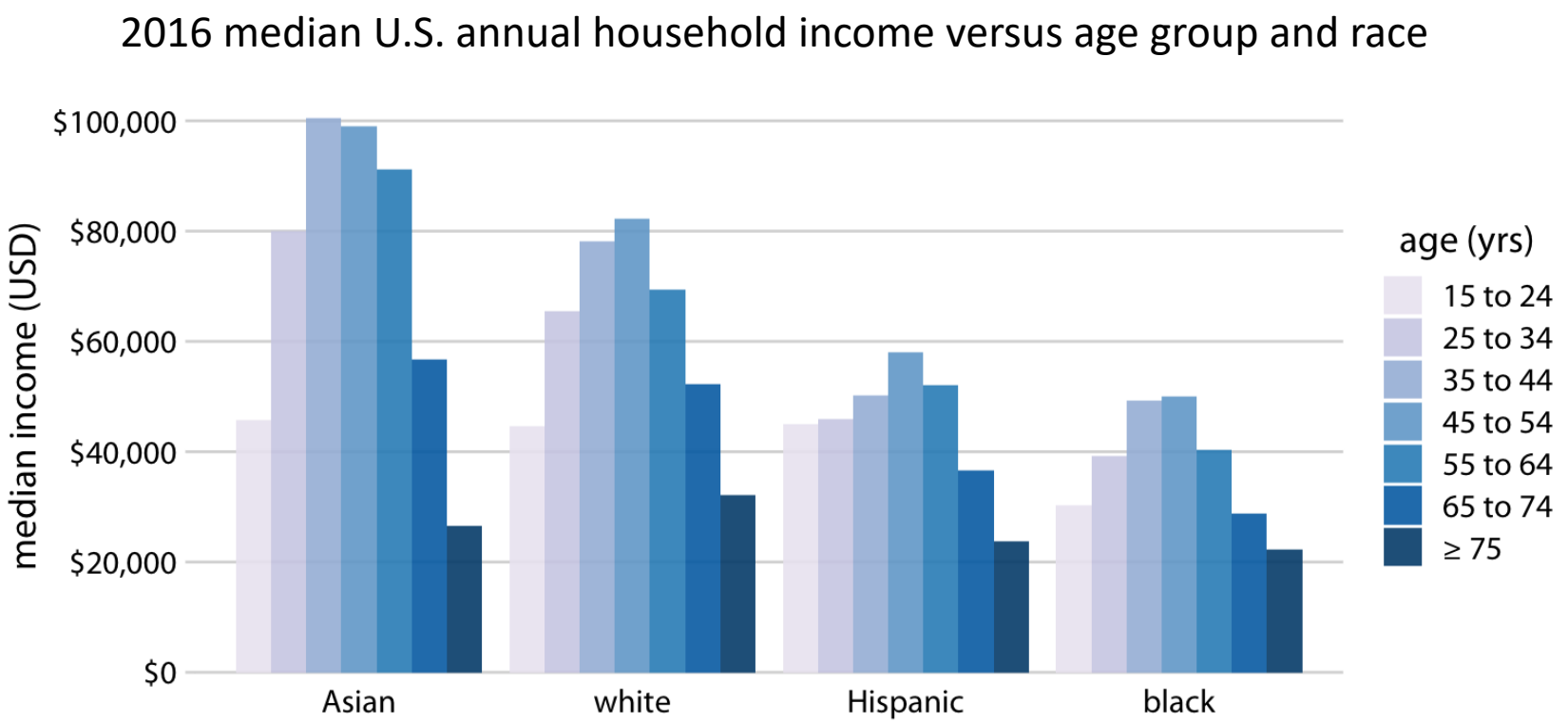


Wilke 2018

Grouped Bar Plots

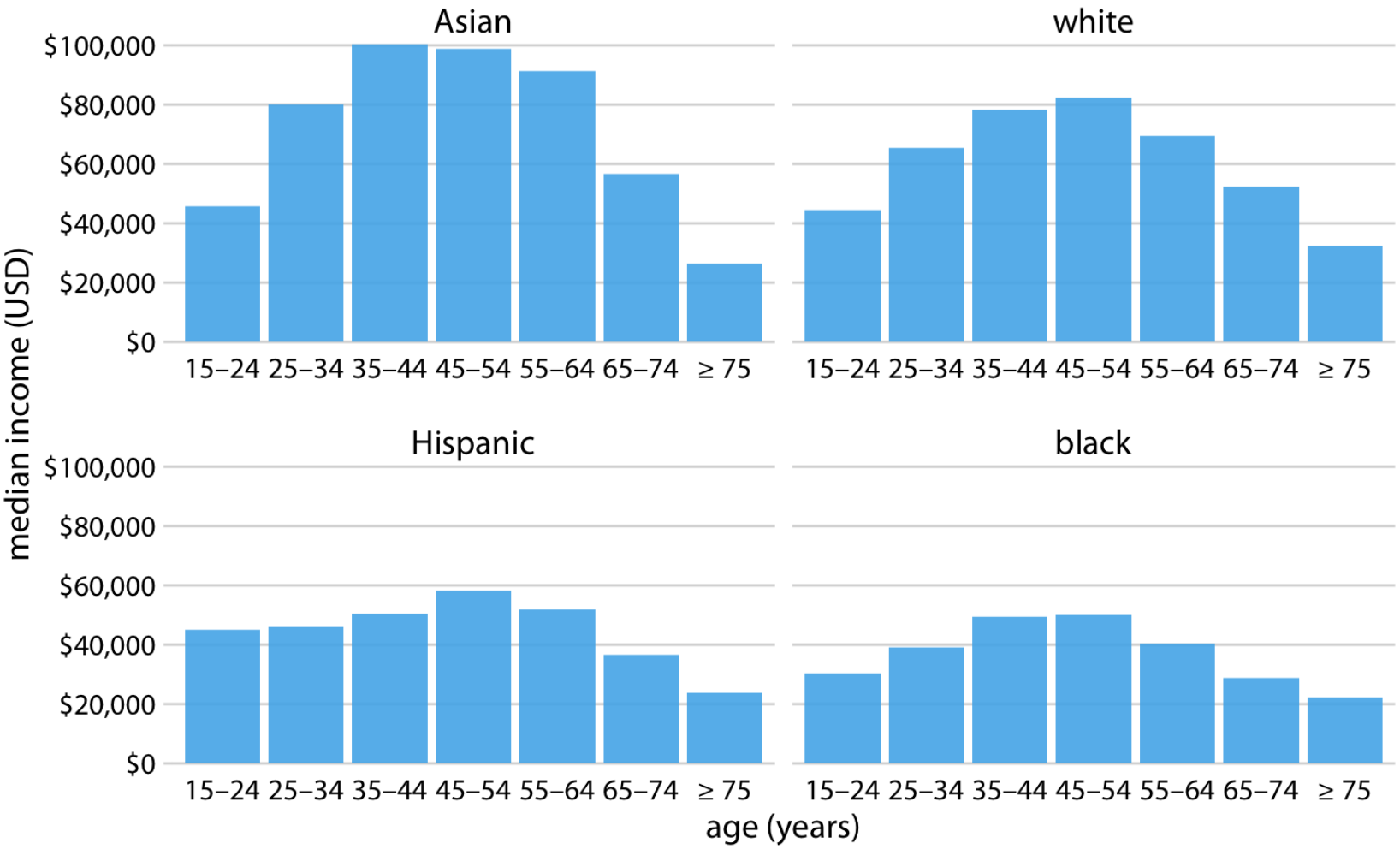


Grouped Bar Plots



Small Multiples Bar Plots

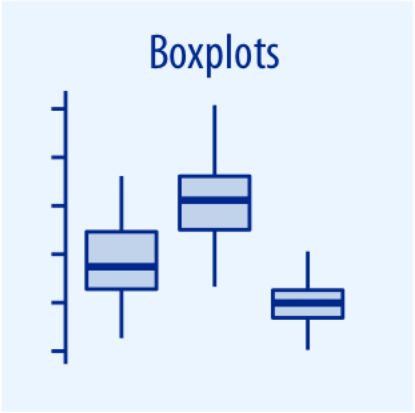
2016 median U.S. annual household income versus age group and race

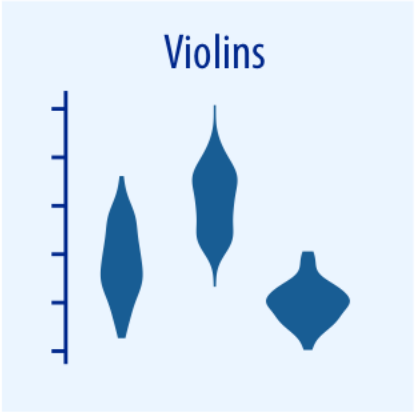


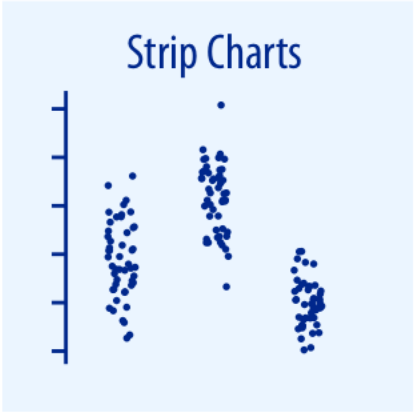
Stacked Bar Plots

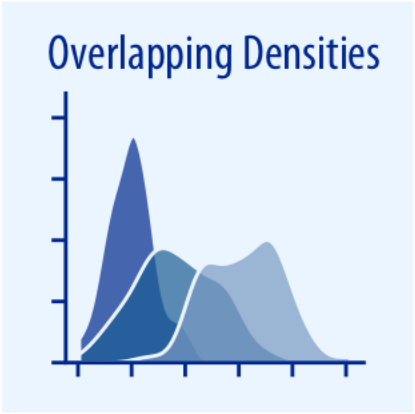
Numbers of female and male passengers on the Titanic traveling in 1st, 2nd, and 3rd class

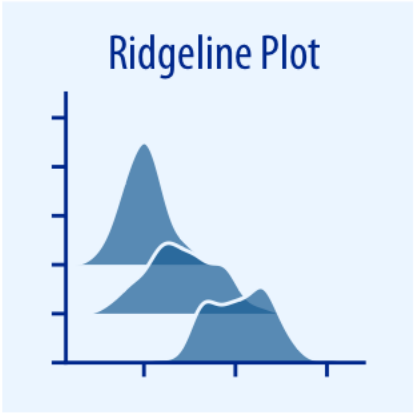
Distributions: Multiple







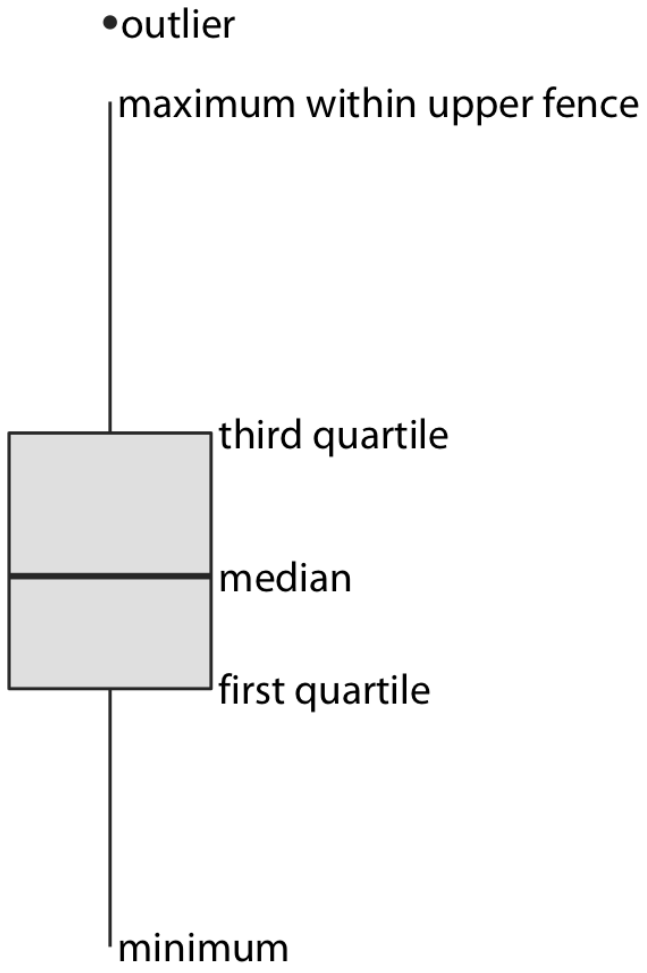




Wilke 2018

Boxplots

- Shows a lot of information:
  - Median
  - Quartiles
  - IQR – interquartile range
  - Range
  - Outliers



Violin Plots

- Combination of box plot with density plot
- Advantage:
  - Illustrates entire distribution

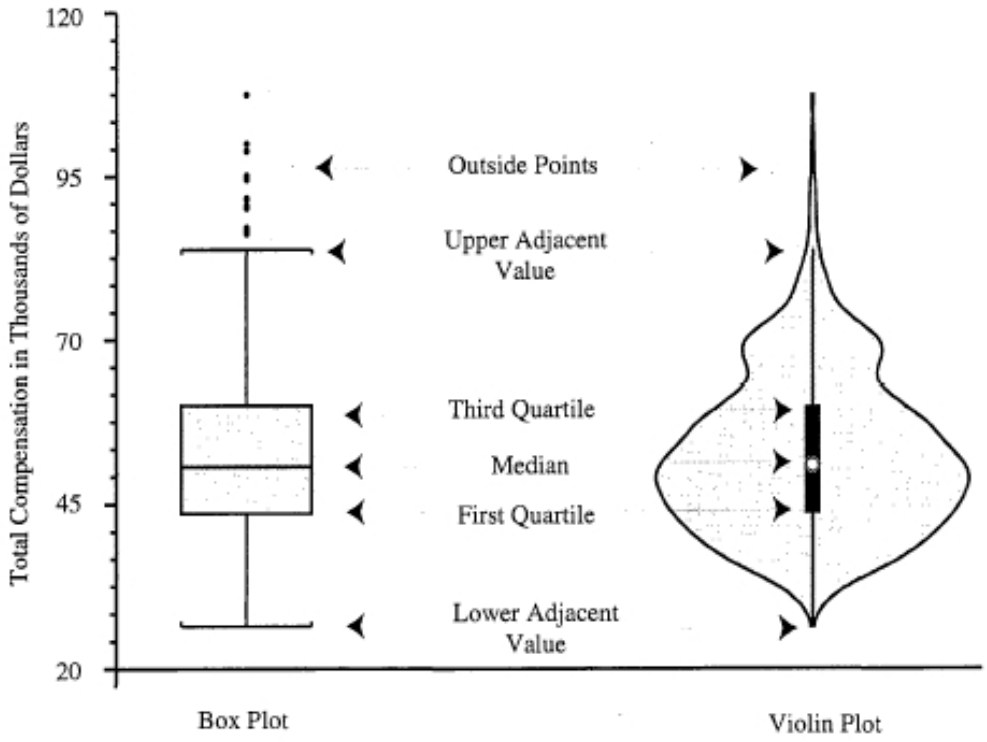
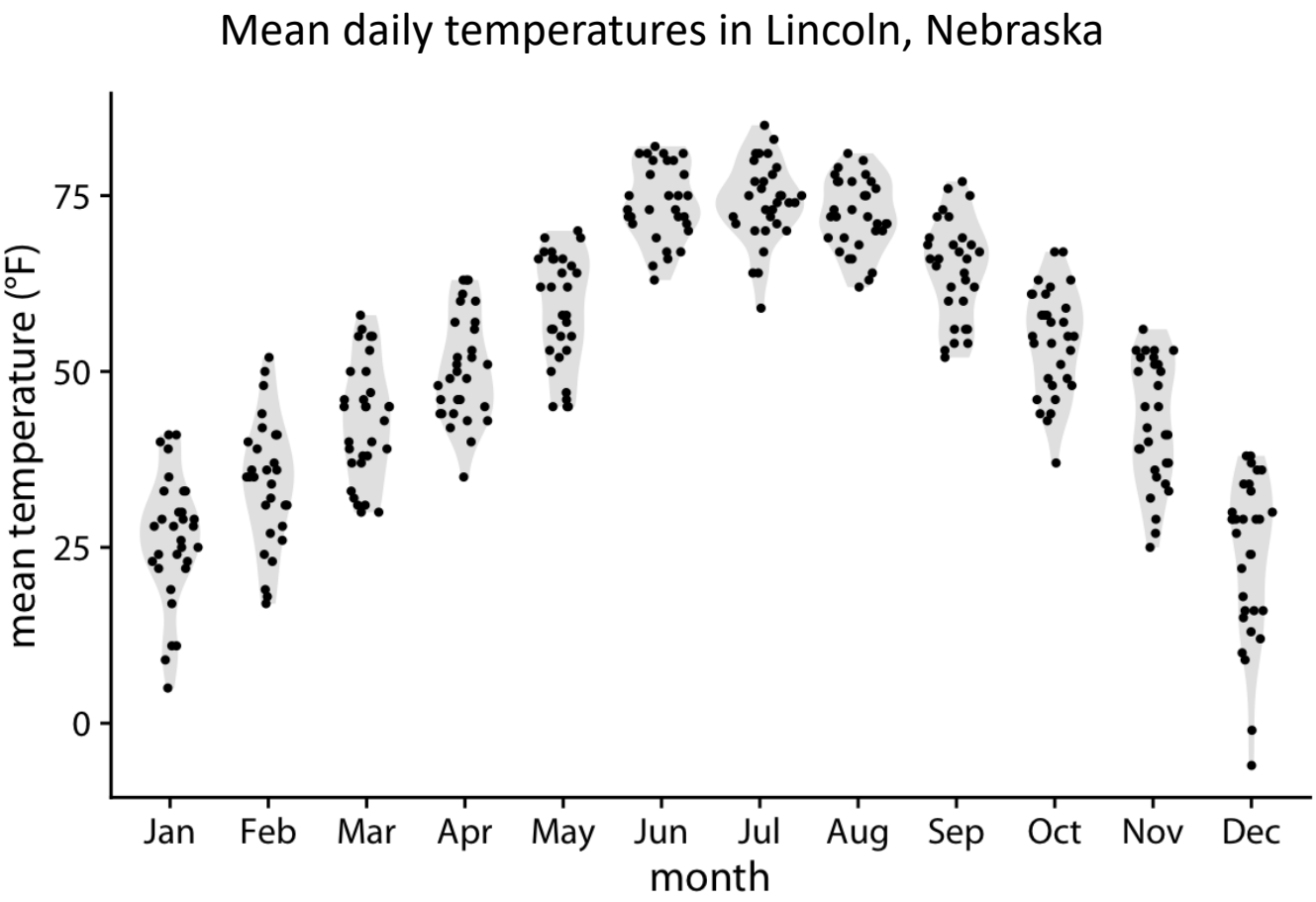


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

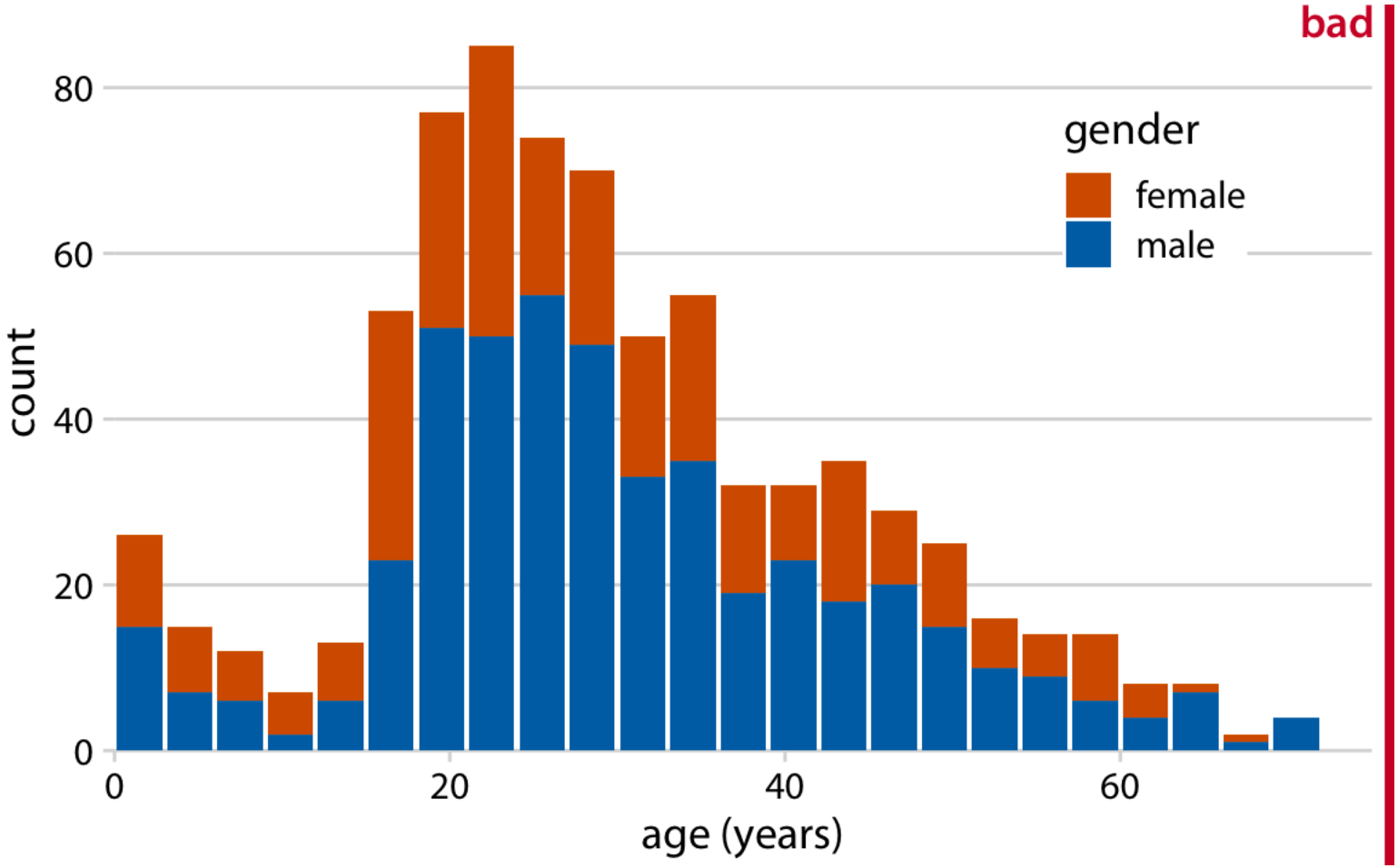
Sina Plot

- When data is sparse, plot raw data points
- Ex. Violin + Sina plot



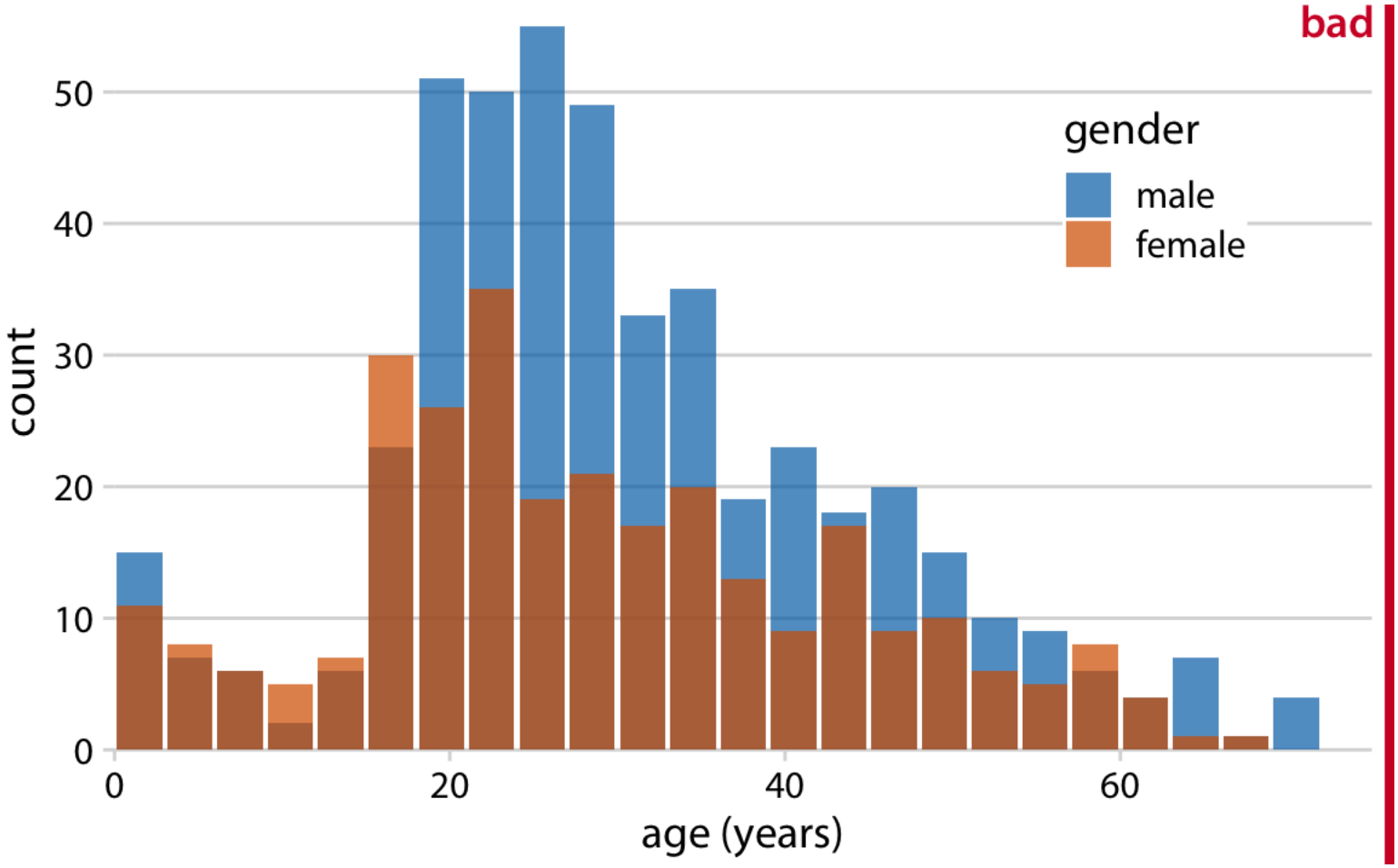
Stacked Histograms

Stacked Histogram of the ages of Titanic passengers stratified by gender



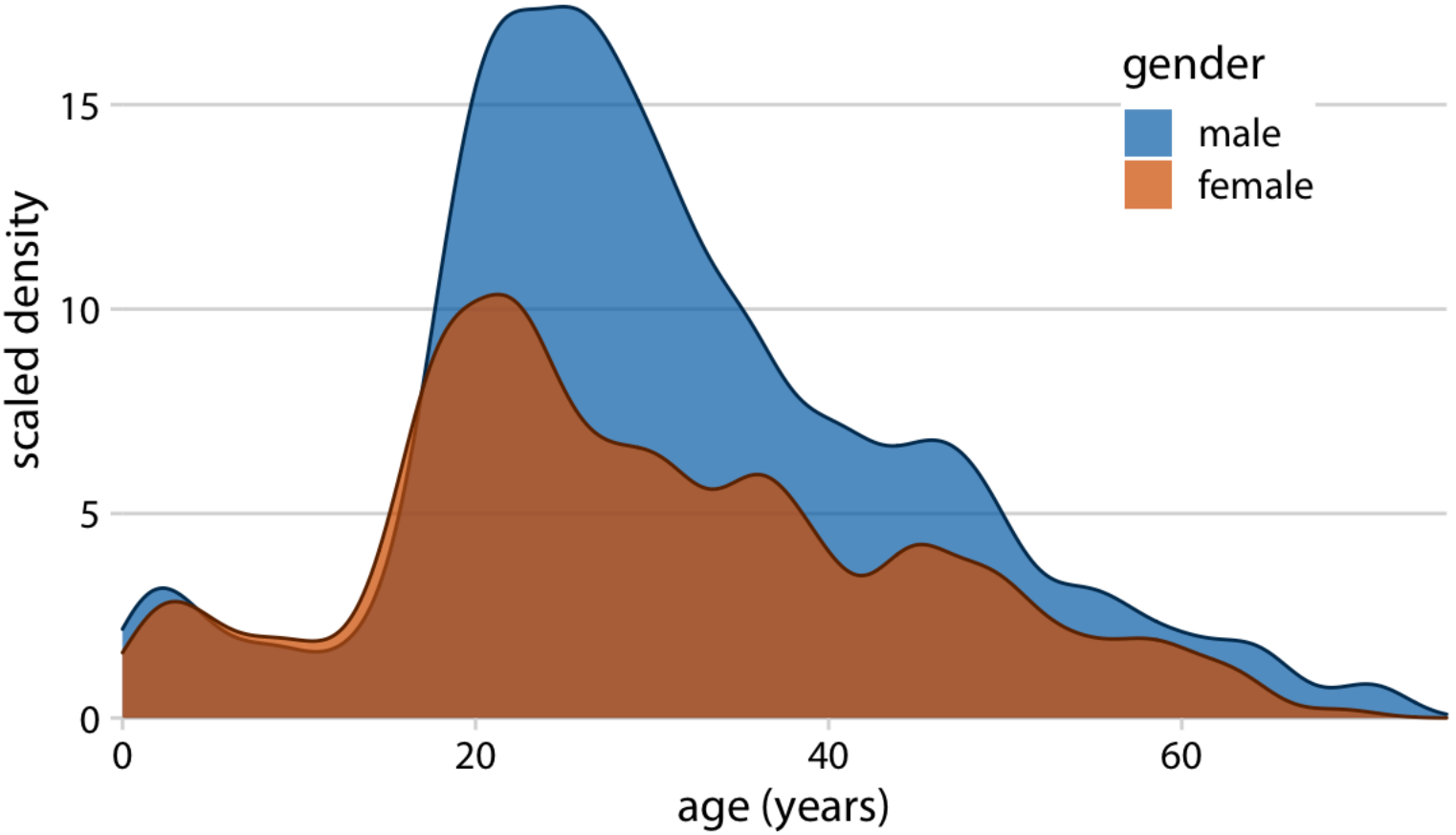
Overlapping Histogram

Overlapping histogram of age distributions of male and female Titanic passengers

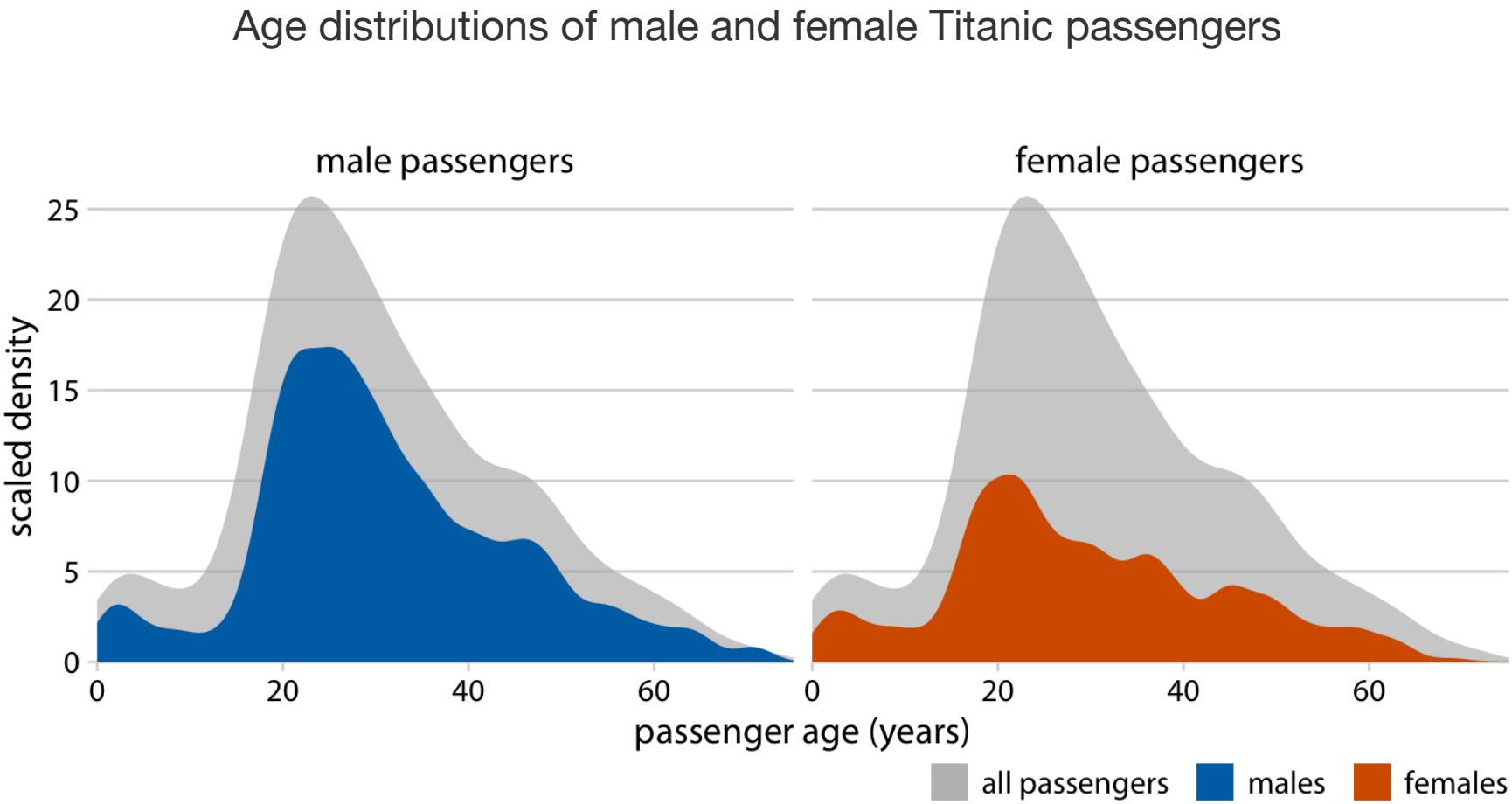


Overlapping Density Plot

Density estimates of the ages of male and female Titanic passengers



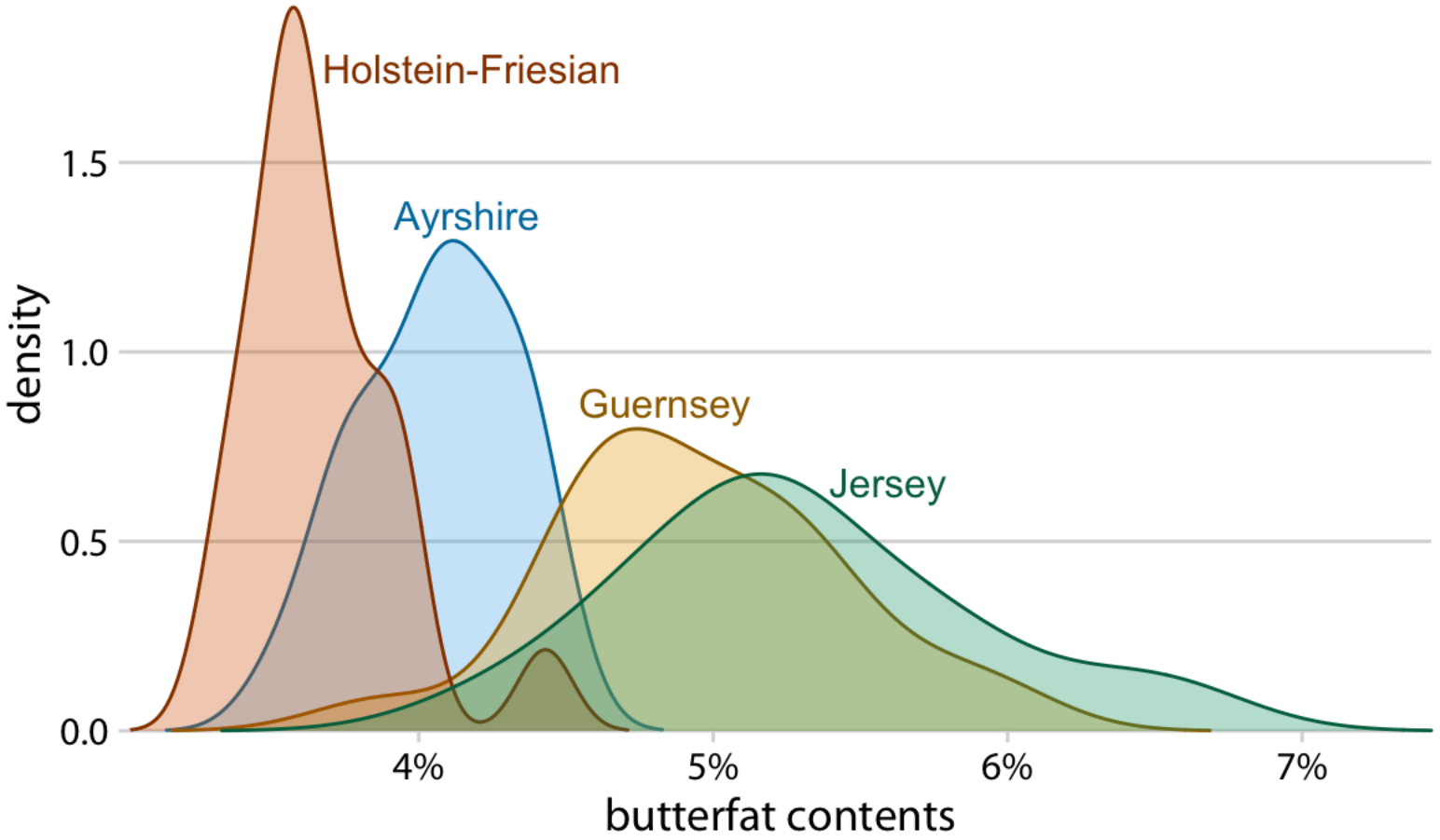
Faceted Density Plot



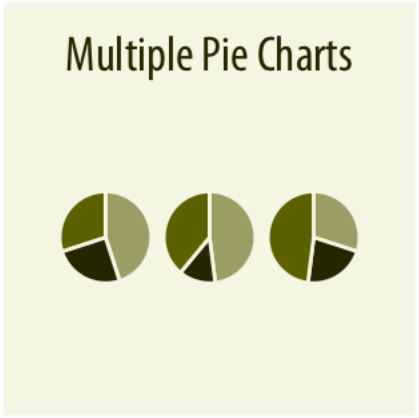
Overlapping Density Plot

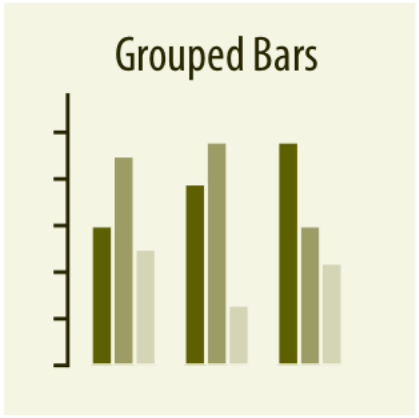
To visualize several distributions at once, kernel density plots will generally work better than histograms.

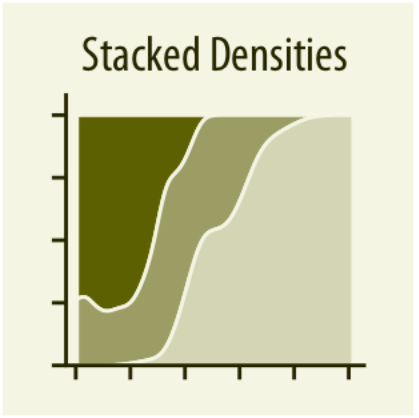
Density estimates of the butterfat percentage in the milk of four cattle breeds

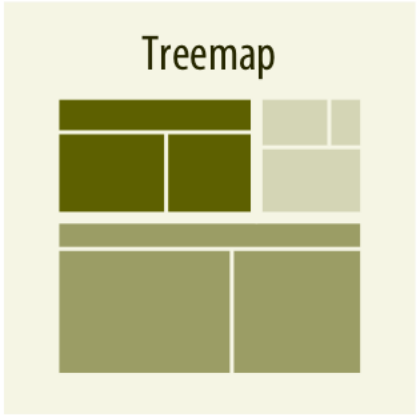


Proportions: Multiple





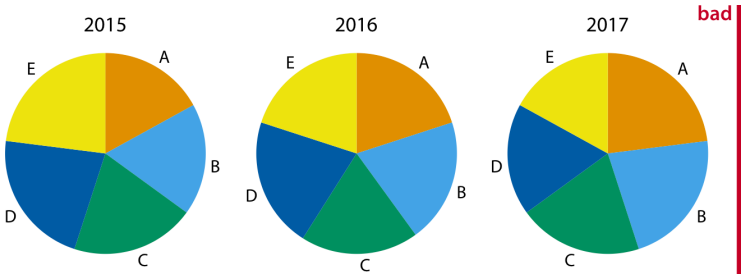




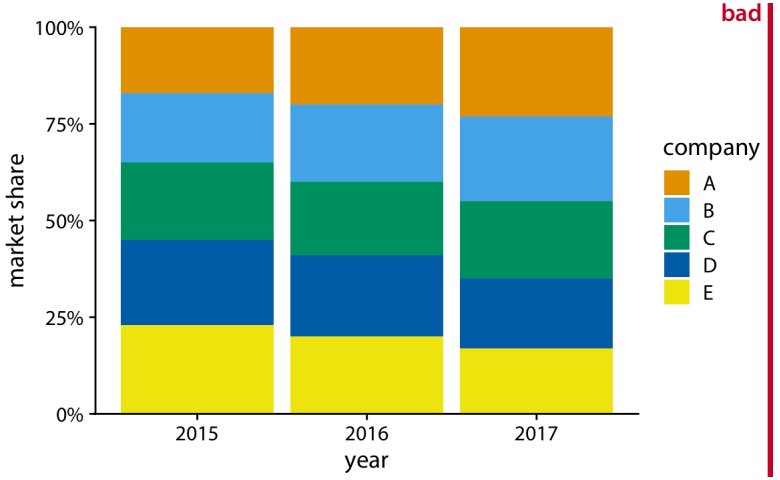
Wilke 2018

Side-by-Side Proportion Plots

Market share of five hypothetical companies for the years 2015–2017



Side-by-side Pie Chart

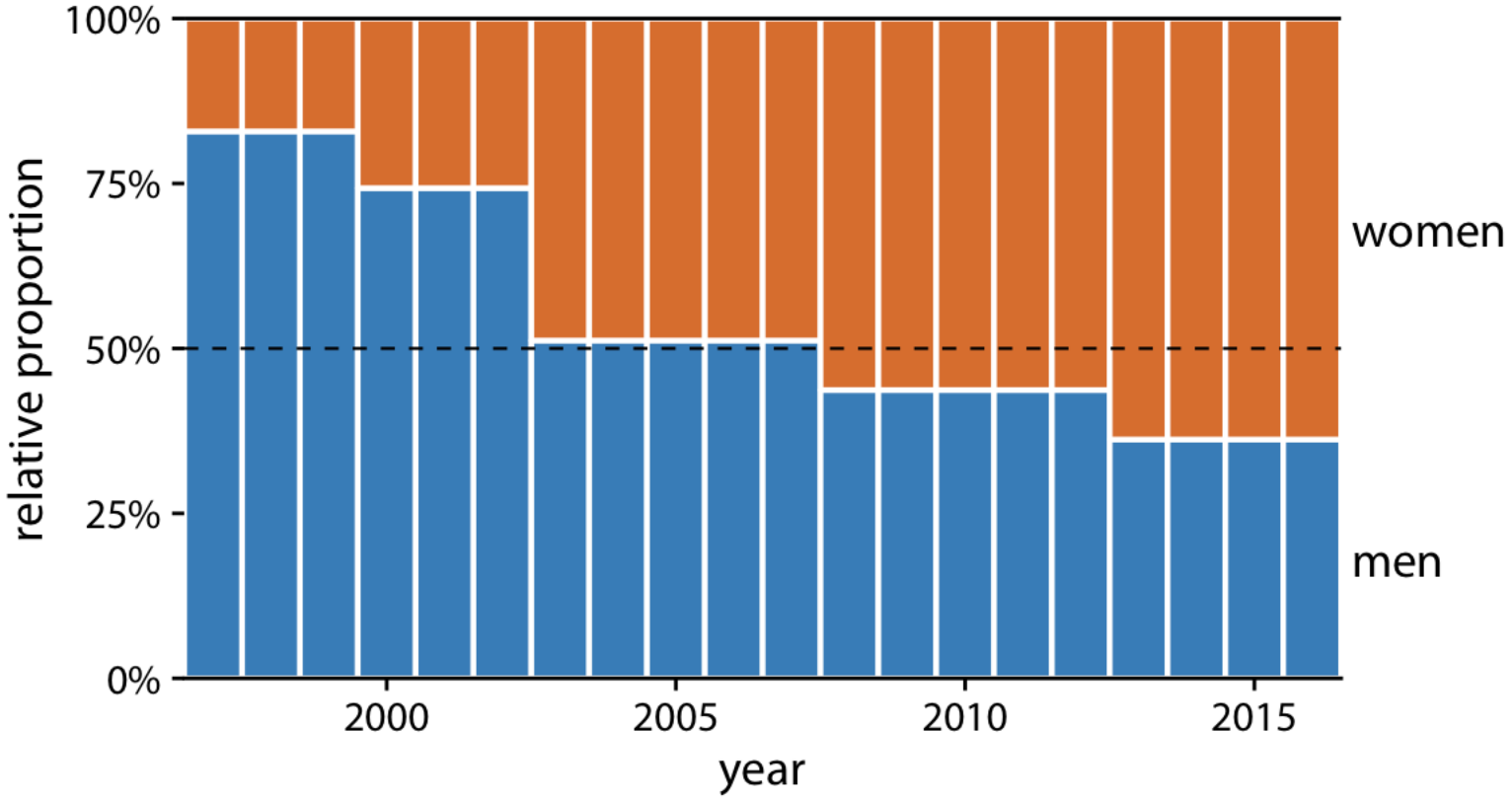


Stacked Bar Chart

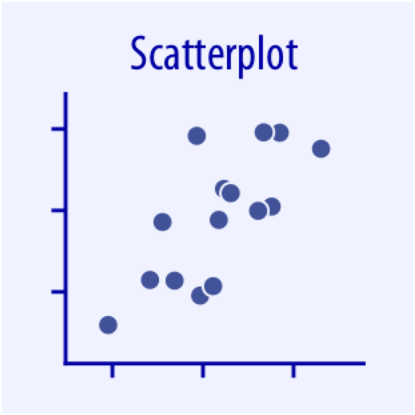
Side-by-side Bar Chart

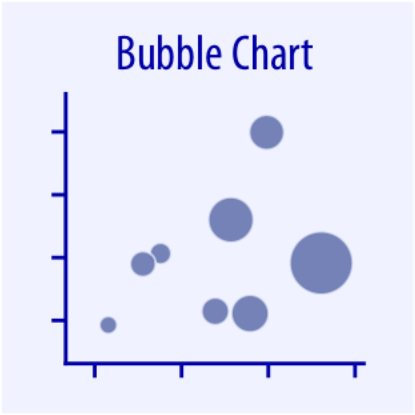
Stacked Bar Plots

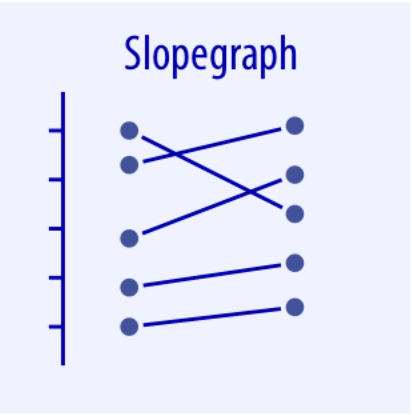
Change in the gender composition of the Rwandan parliament over time, 1997 to 2016



x-y Relationships

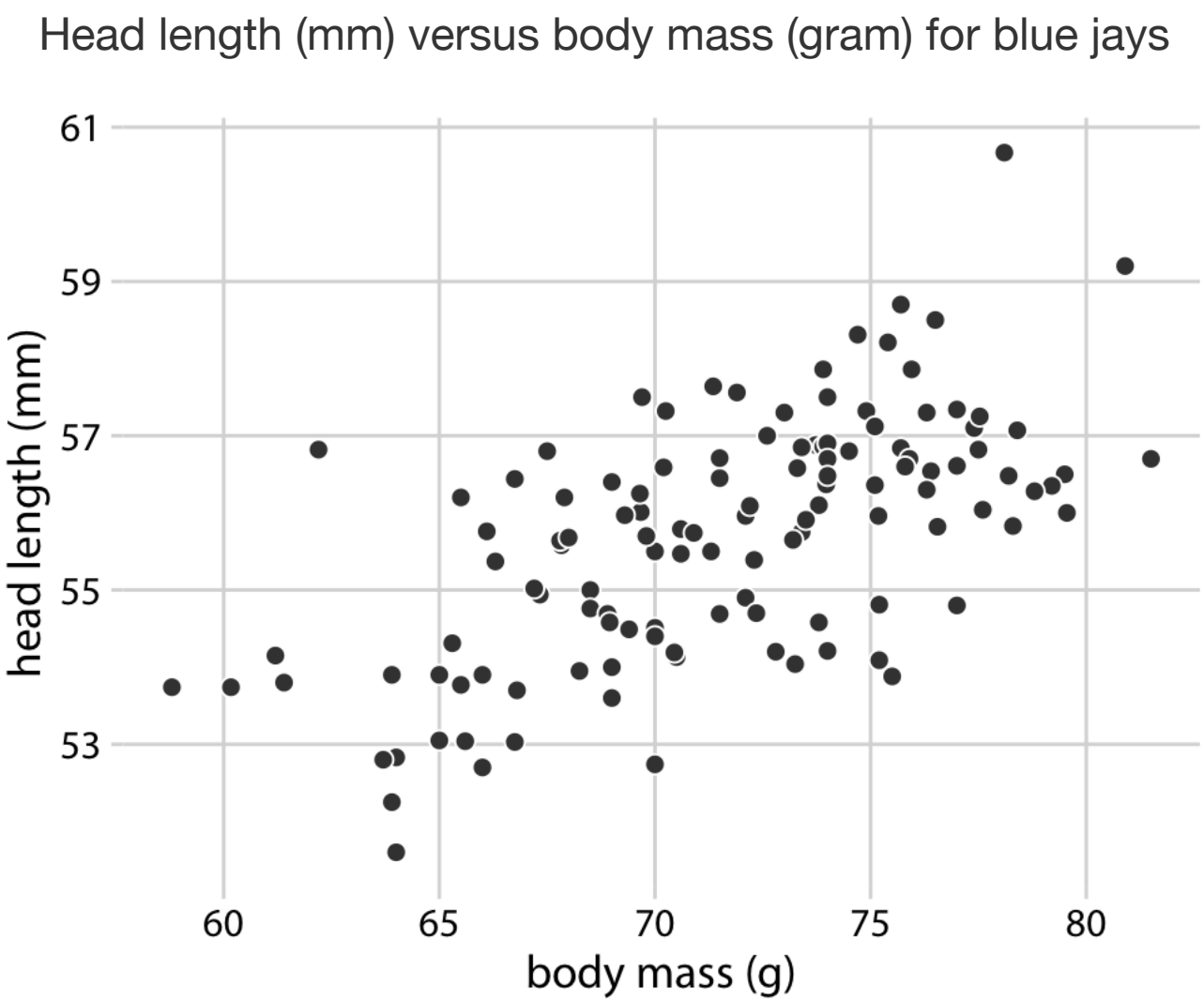






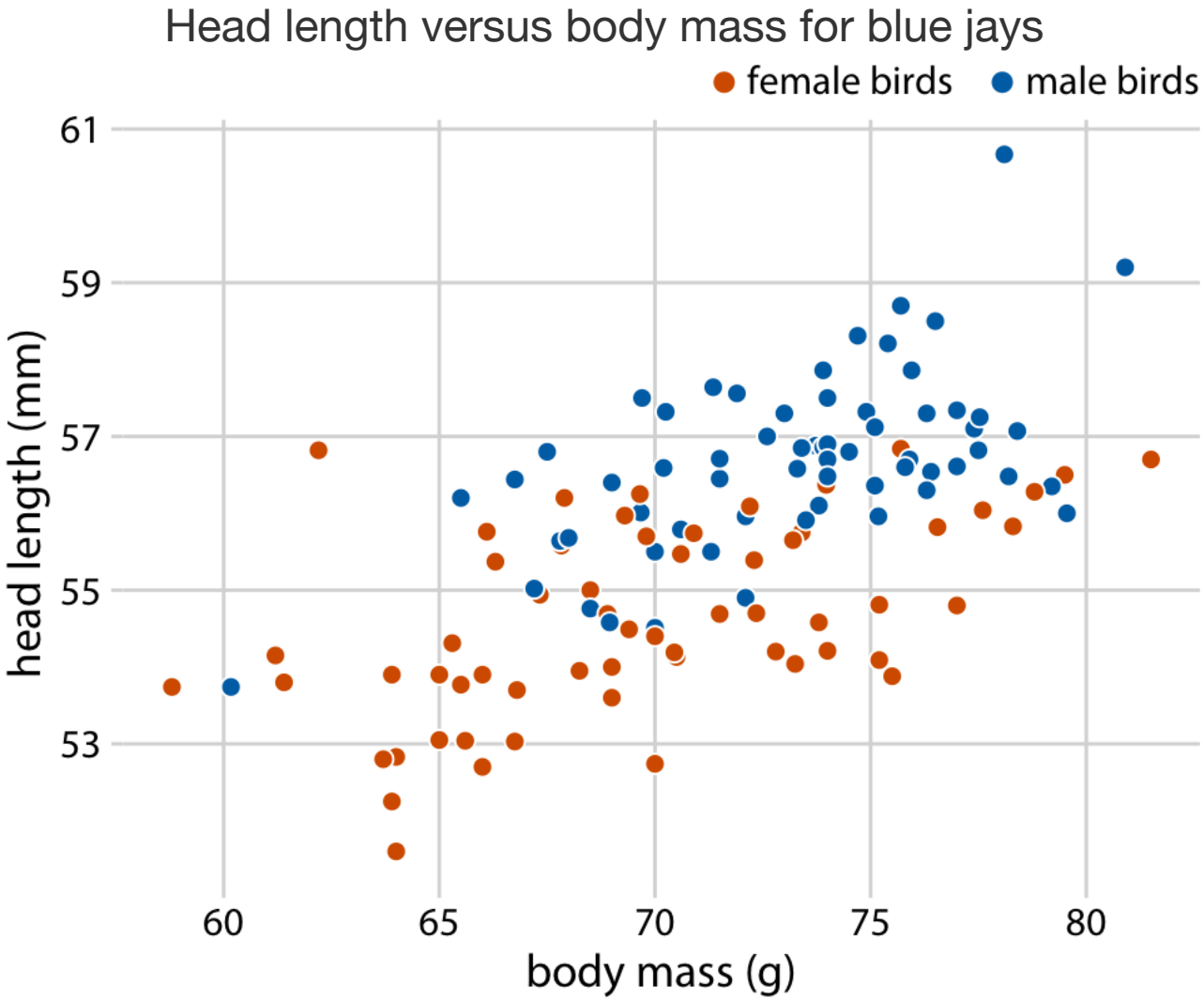
Wilke 2018

Scatter Plots



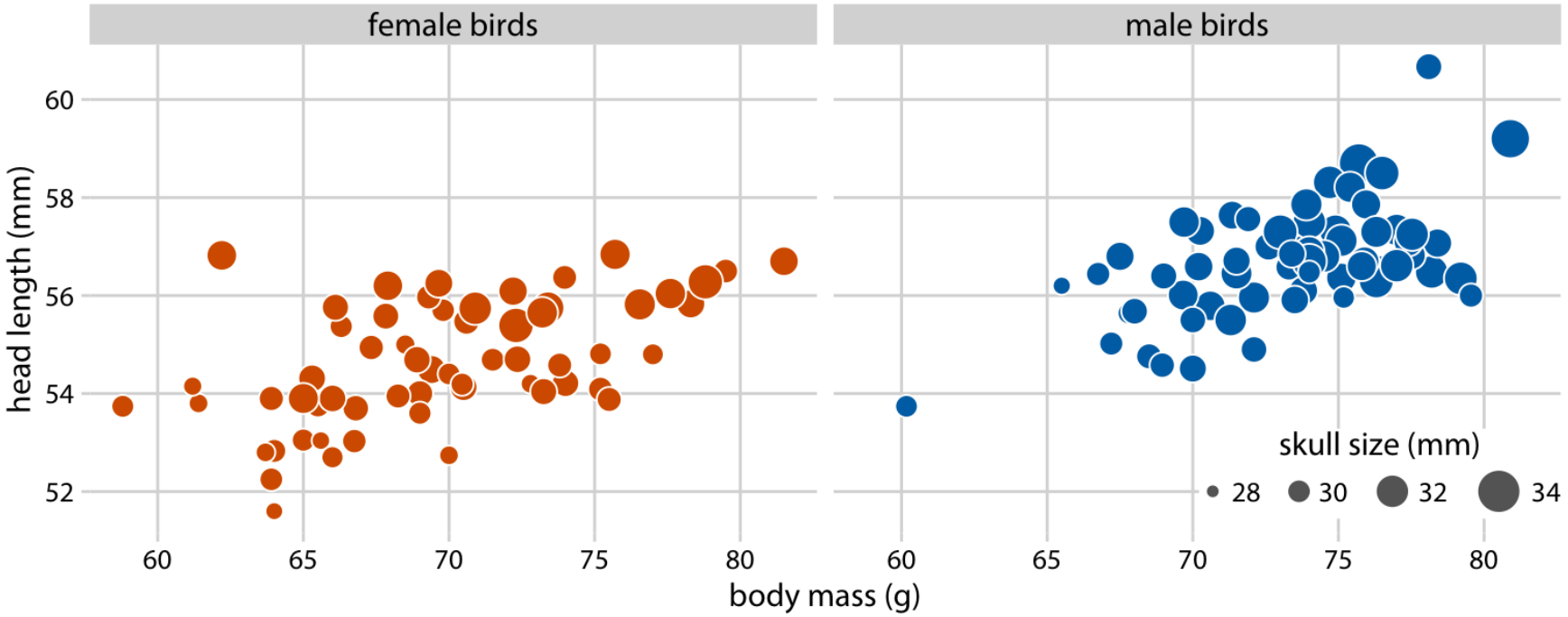
Scatter Plots – Additional Variables

Add color to differentiate categorical variable

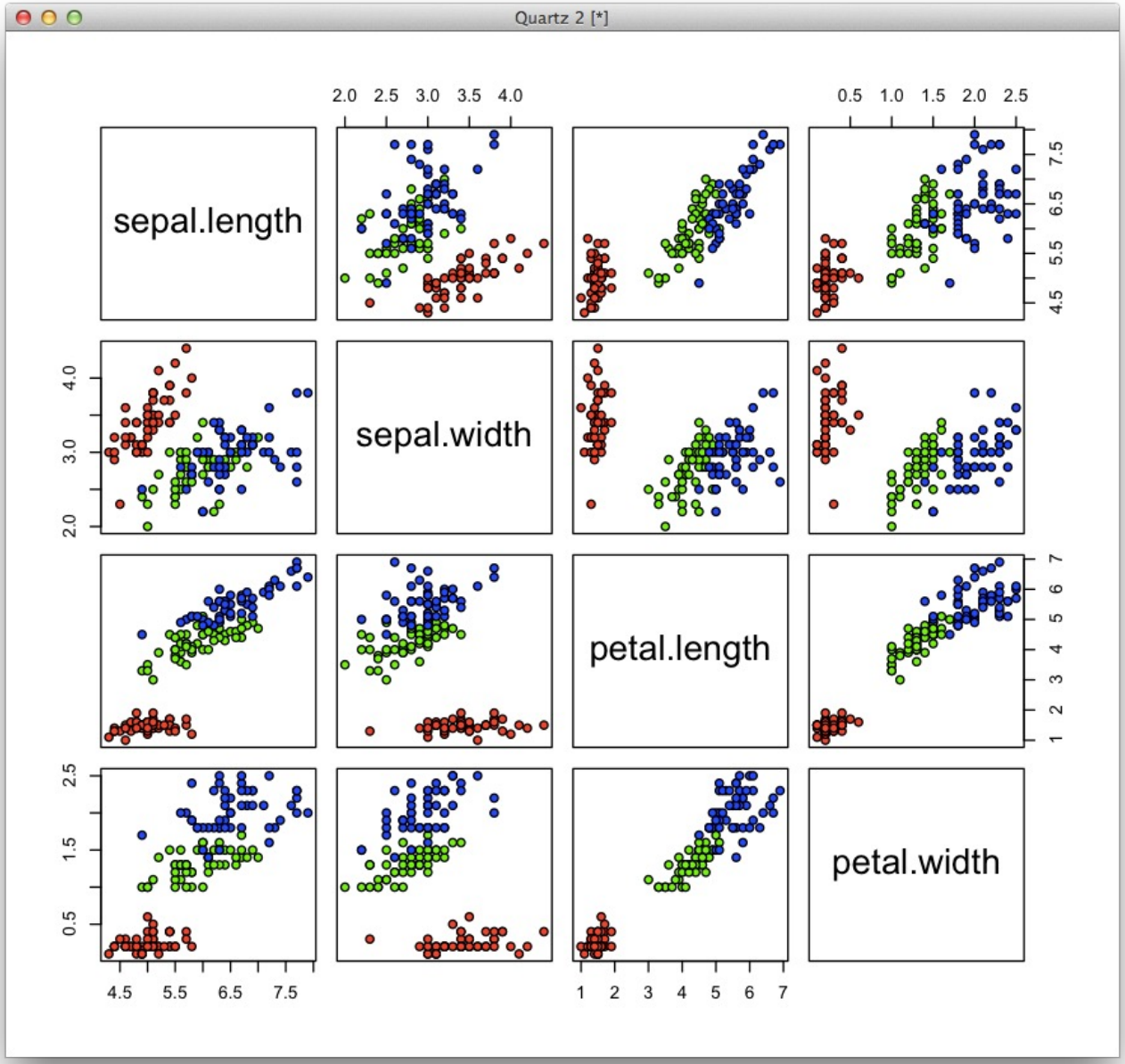


Scatter Plot – Additional Variables

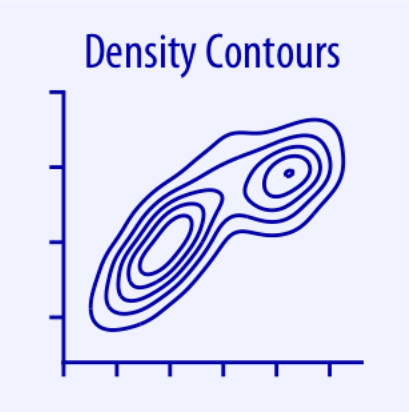
Head length versus body mass for blue jays.

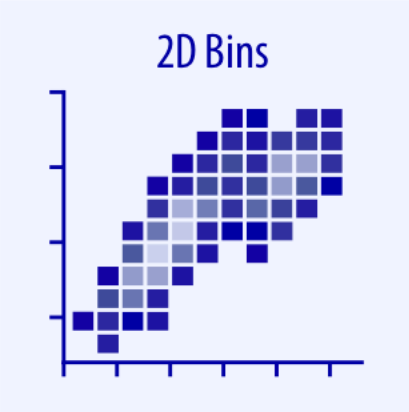


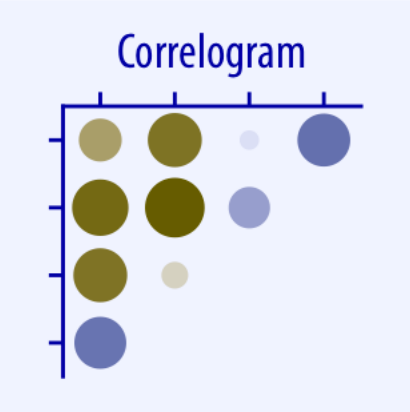
Scatter Plot Matrix



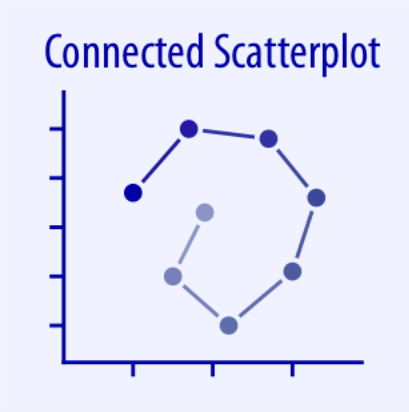
x-y Relationships with lots of data

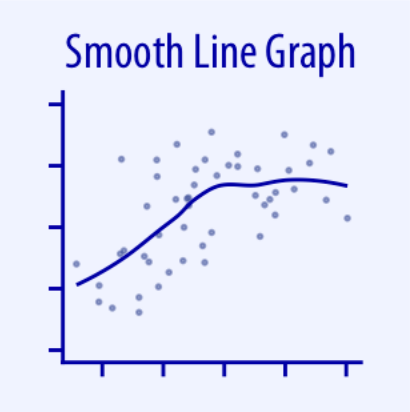






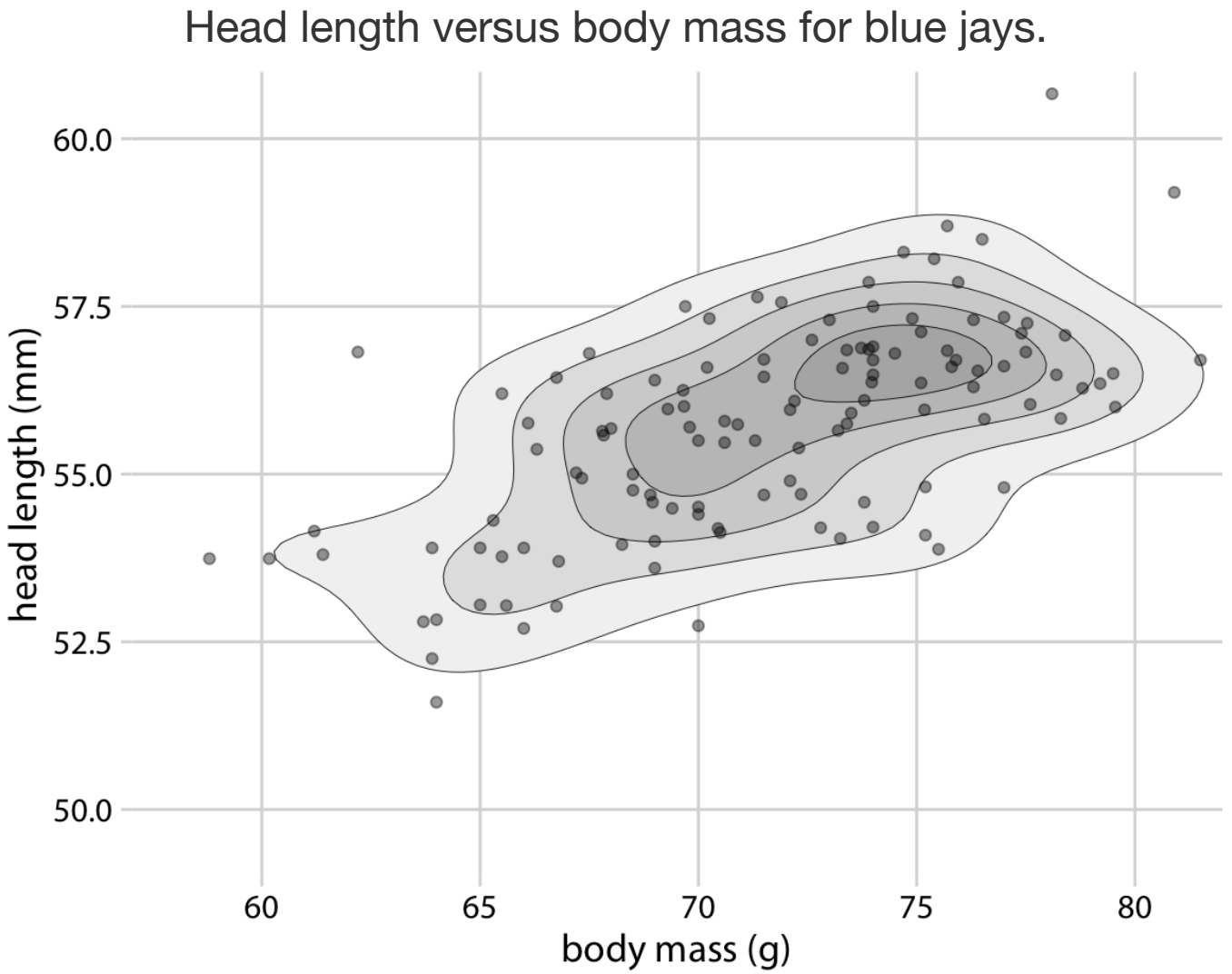
Adding Time



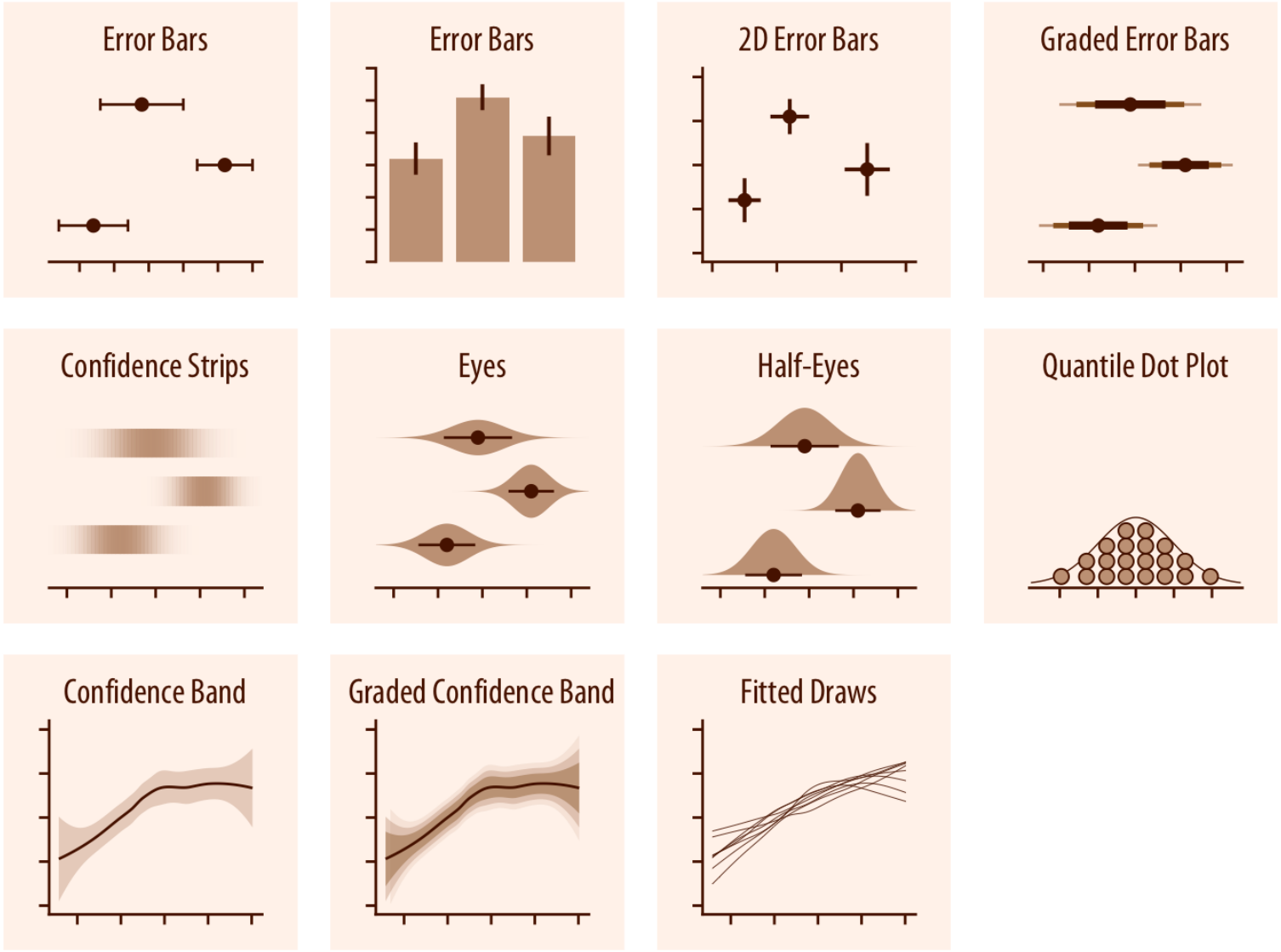


Wilke 2018

Contour Plots



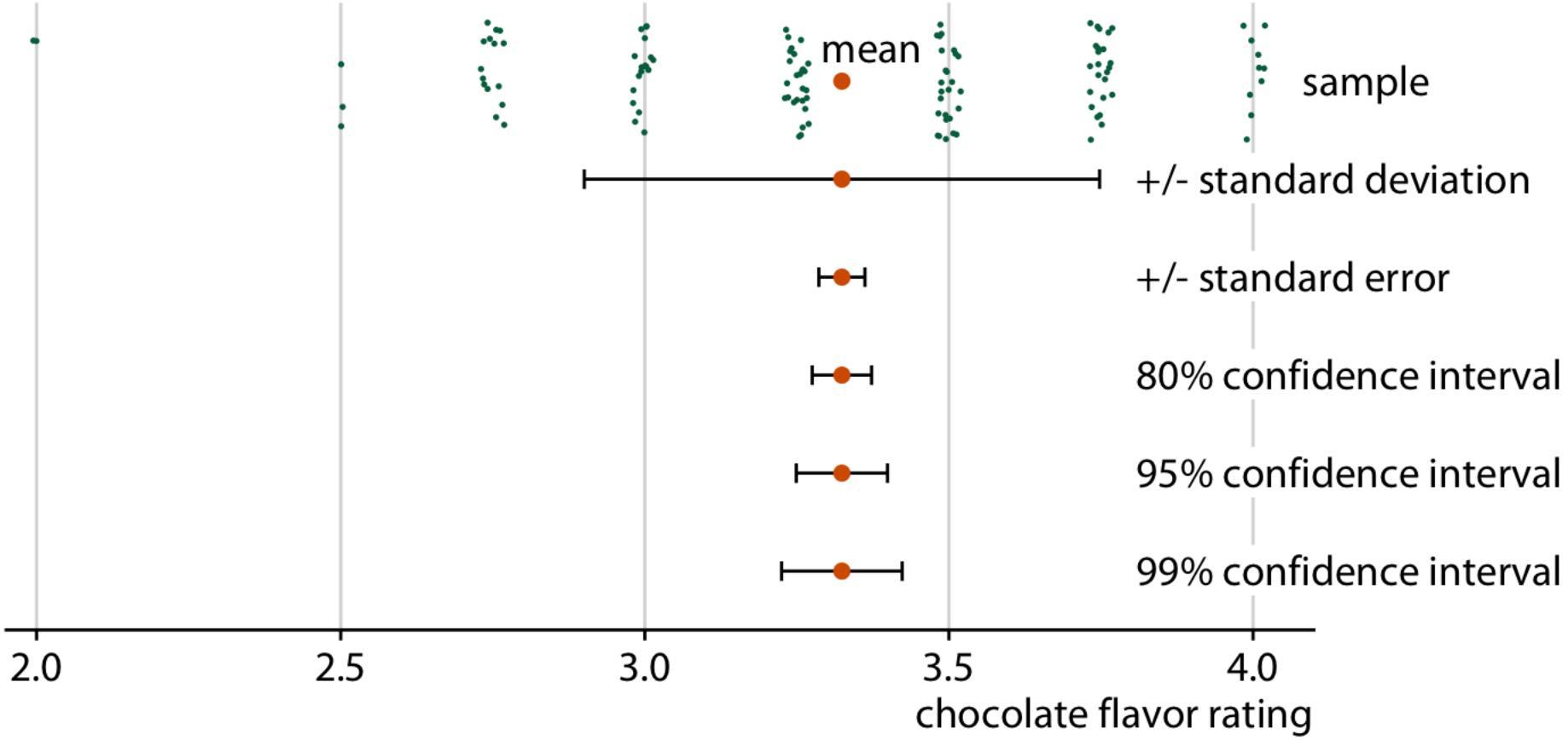
Uncertainty



Wilke 2018

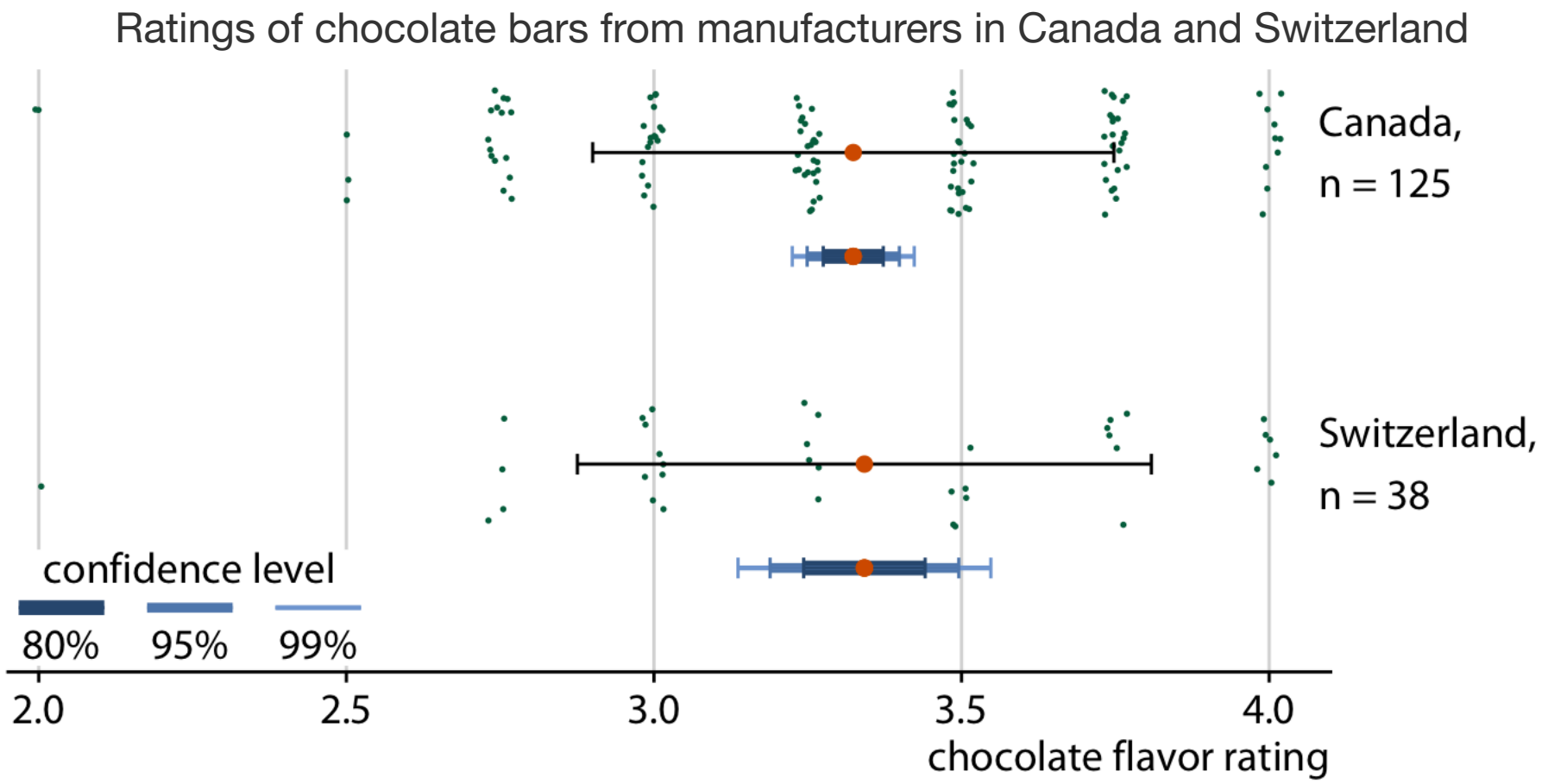
Uncertainty – Error Bars

Relationship between sample, sample mean, standard deviation, standard error, and confidence intervals

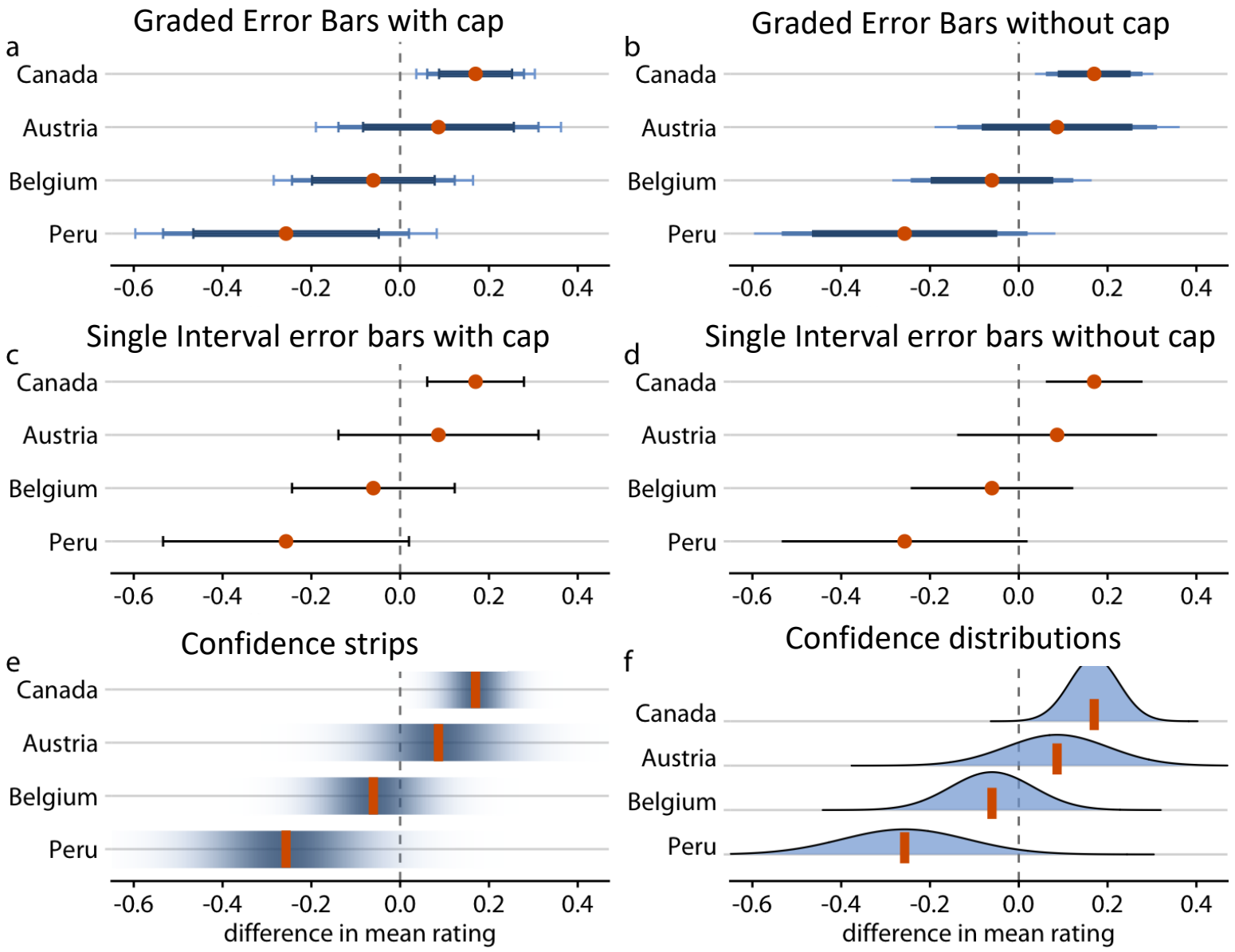


Wilke 2018

Graded Error Bars



Uncertainty Information



Tufte 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
  - requires telling the truth about the data.

Colors Can Distinguish Groups

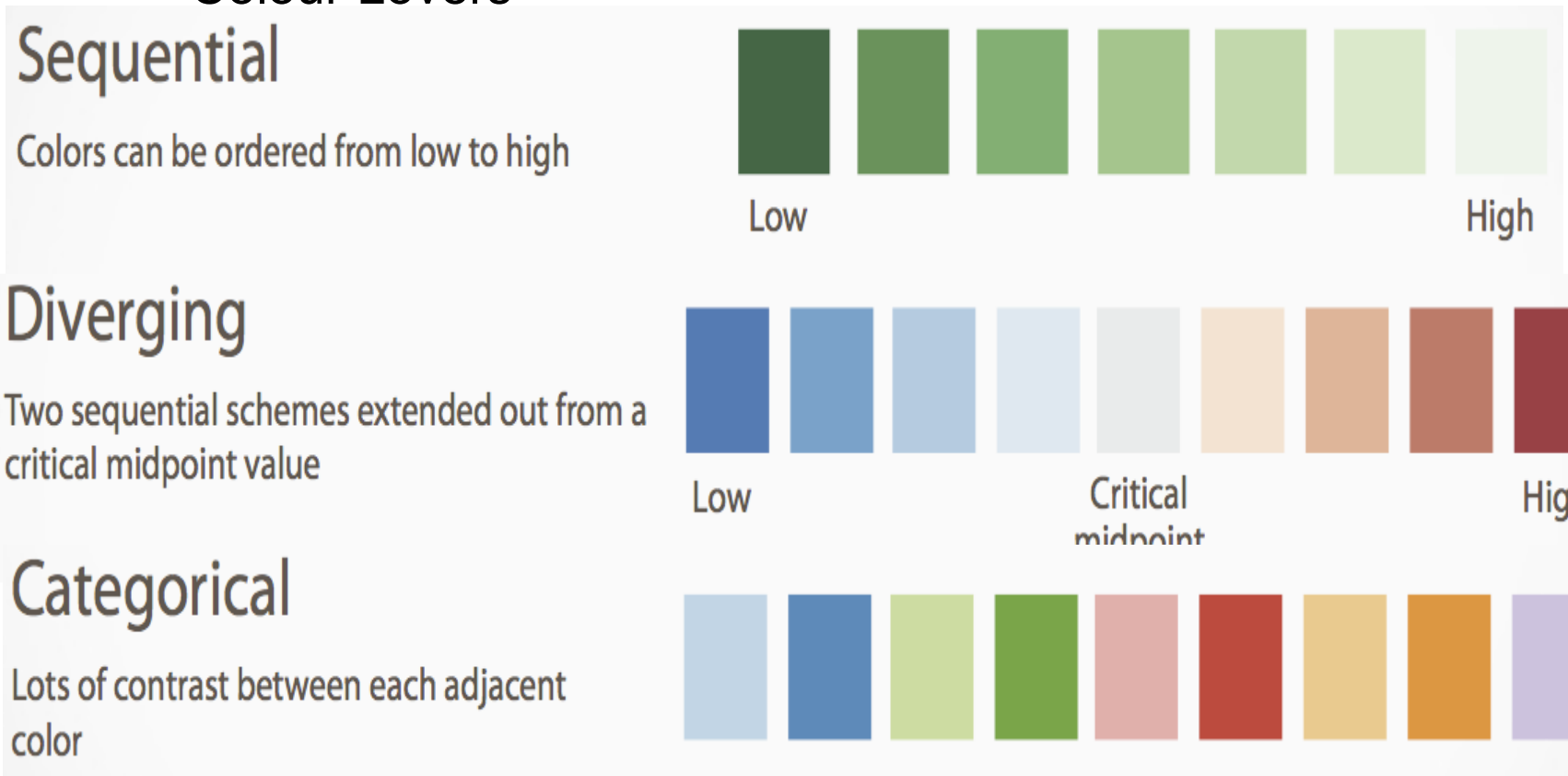
Wilke 2018

Colors Can Show Data

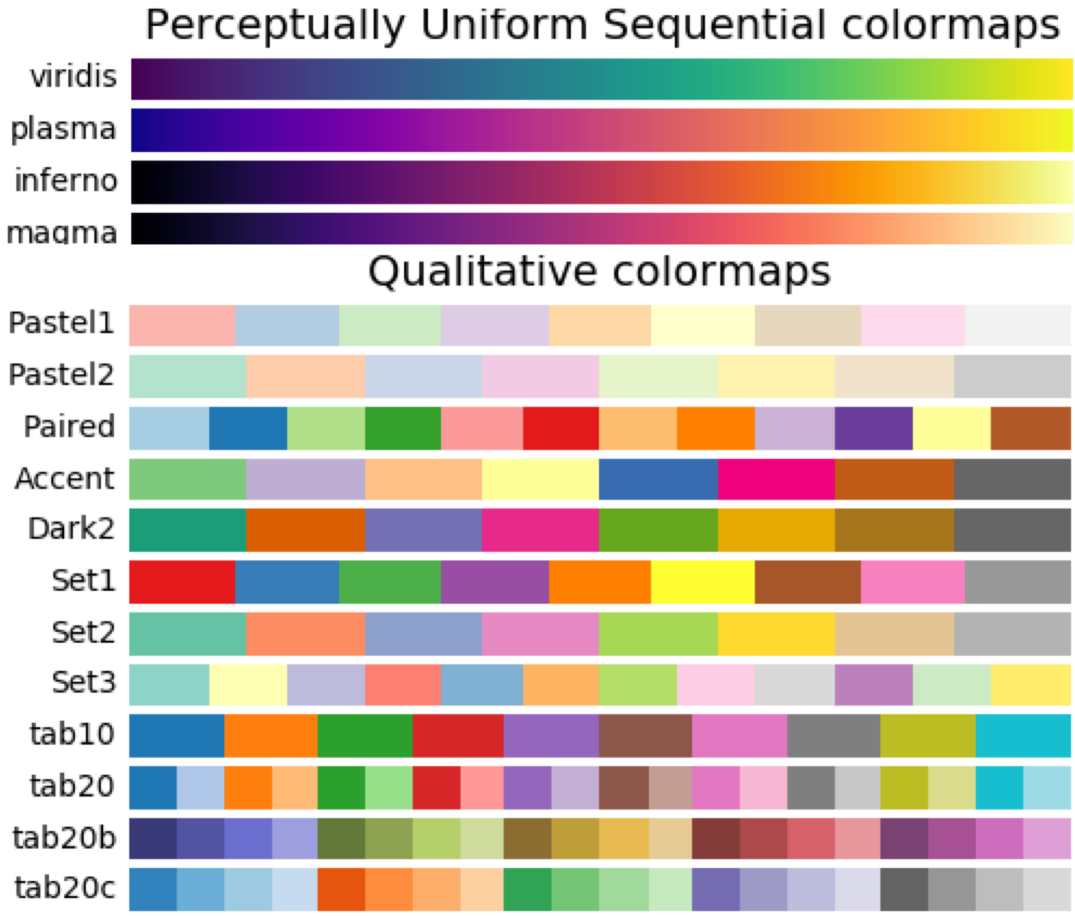
Wilke 2018

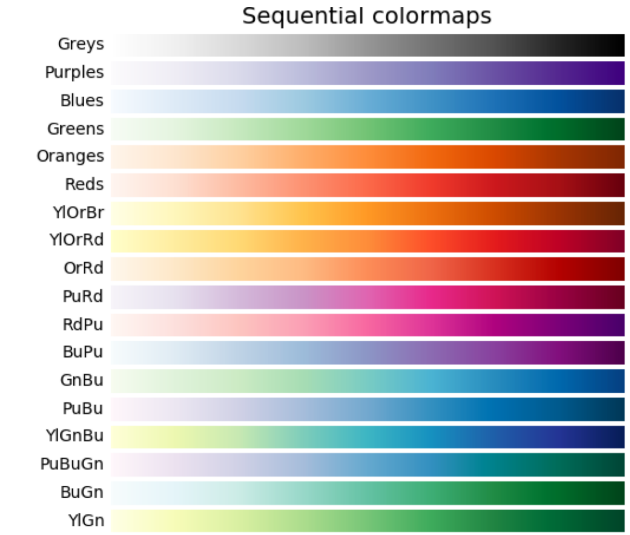
Color Schemes

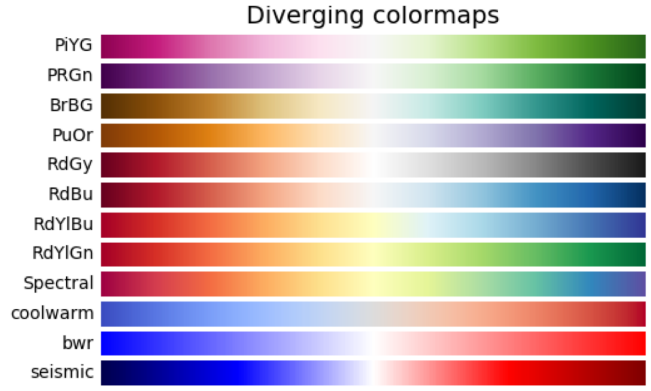
- Use online resources to discover and record your color schemes
  - Color Brewer
  - Kuler
  - Colour Lovers



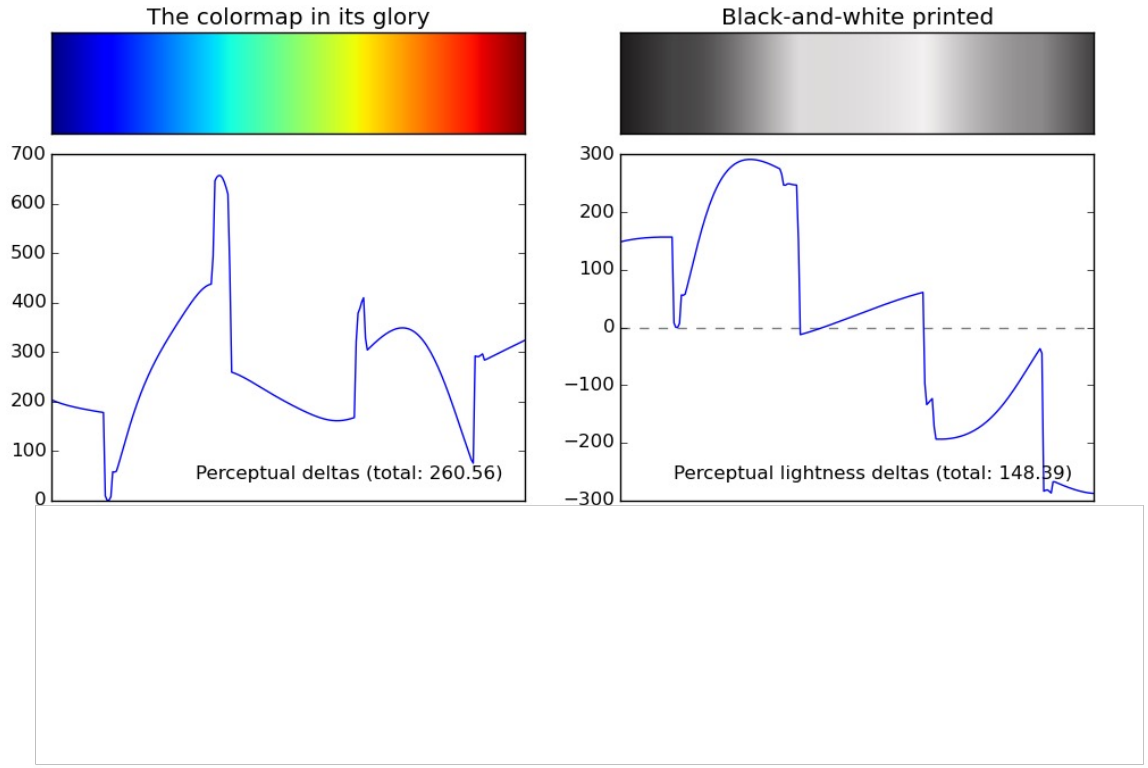
Color Schemes

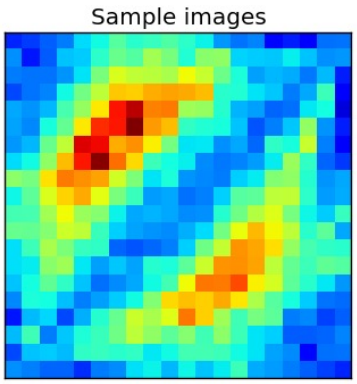


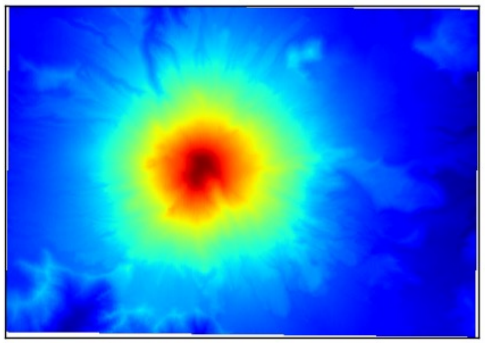


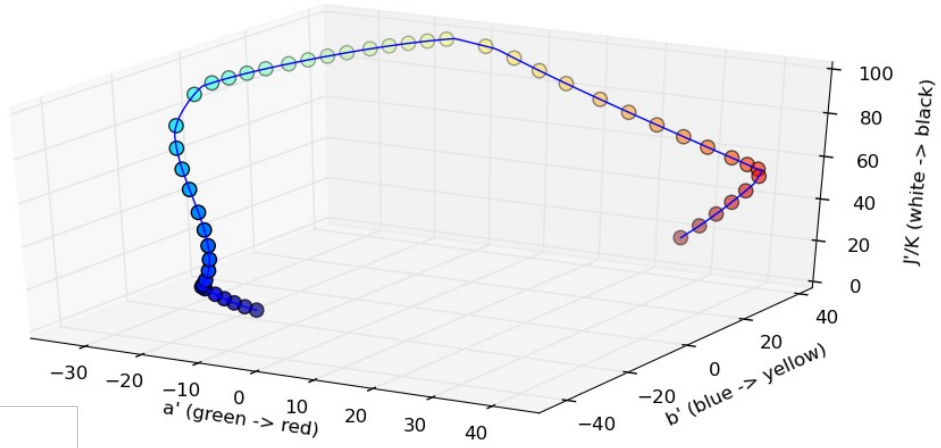


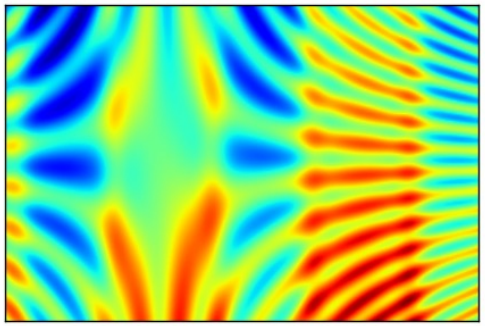
Colormap evaluation: jet





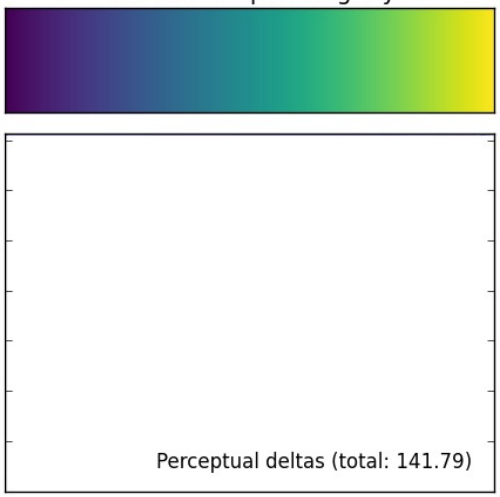




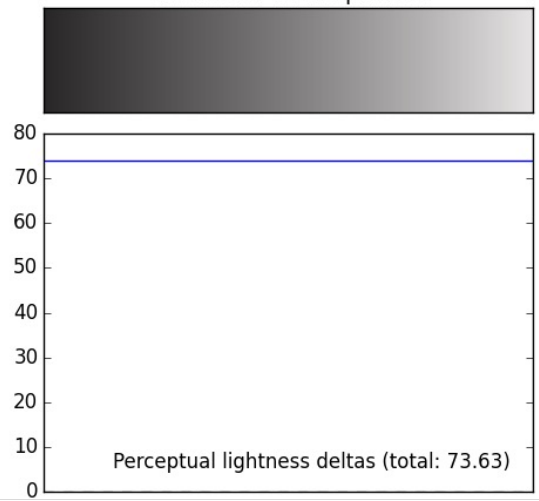


Colormap evaluation: option_d.py

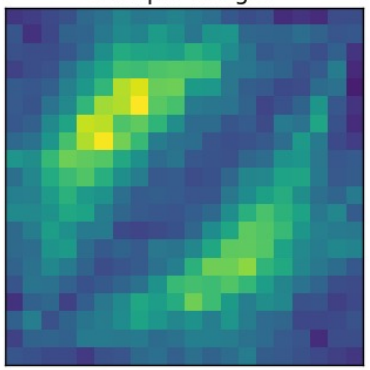
The colormap in its glory

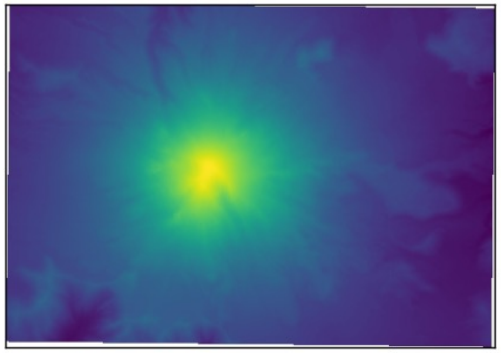


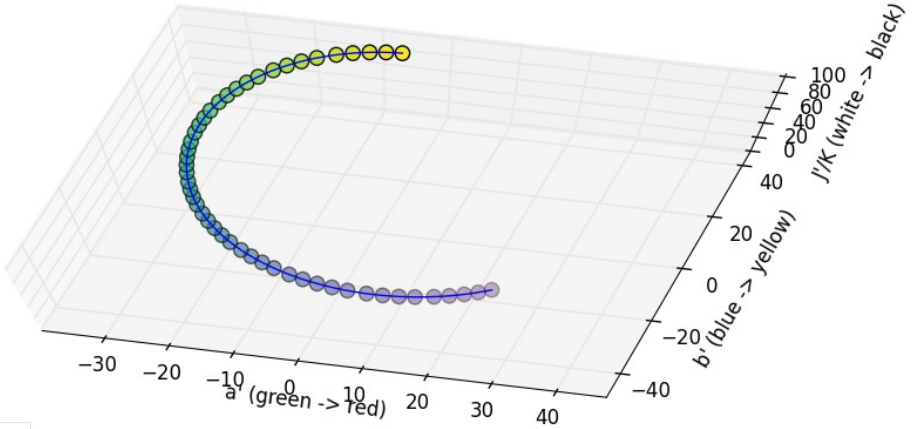
Black-and-white printed

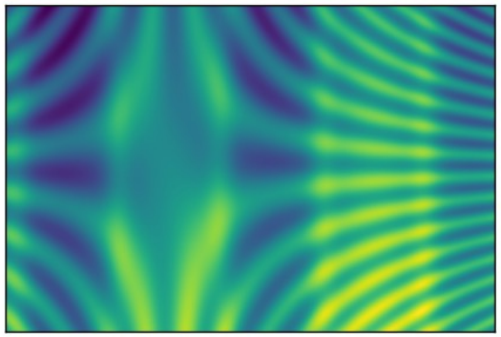


Sample images

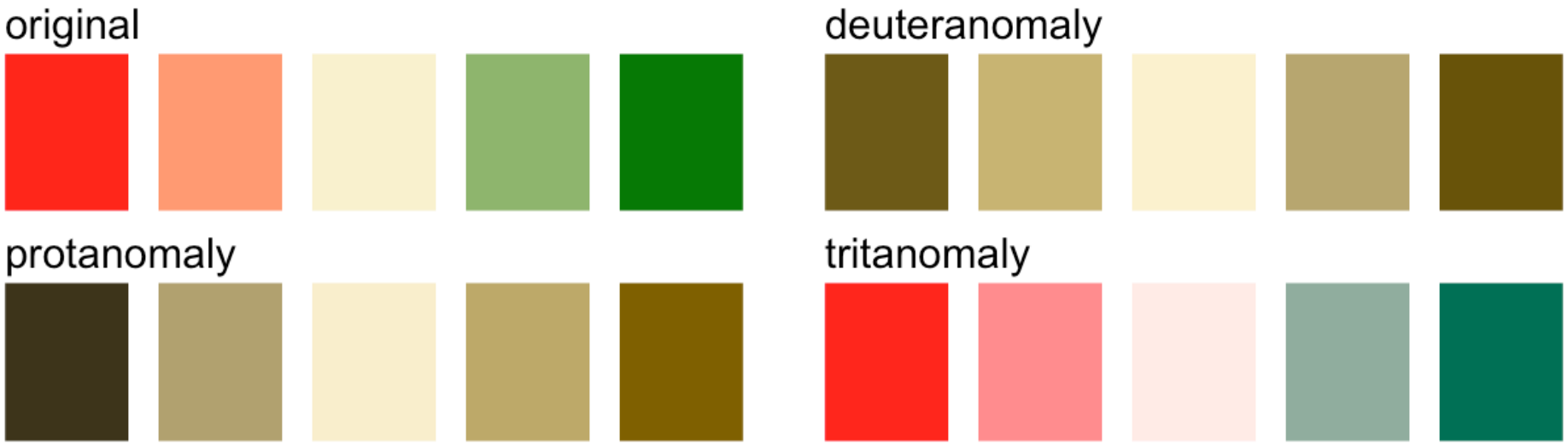








Beware of Not Thinking about Color Blindness

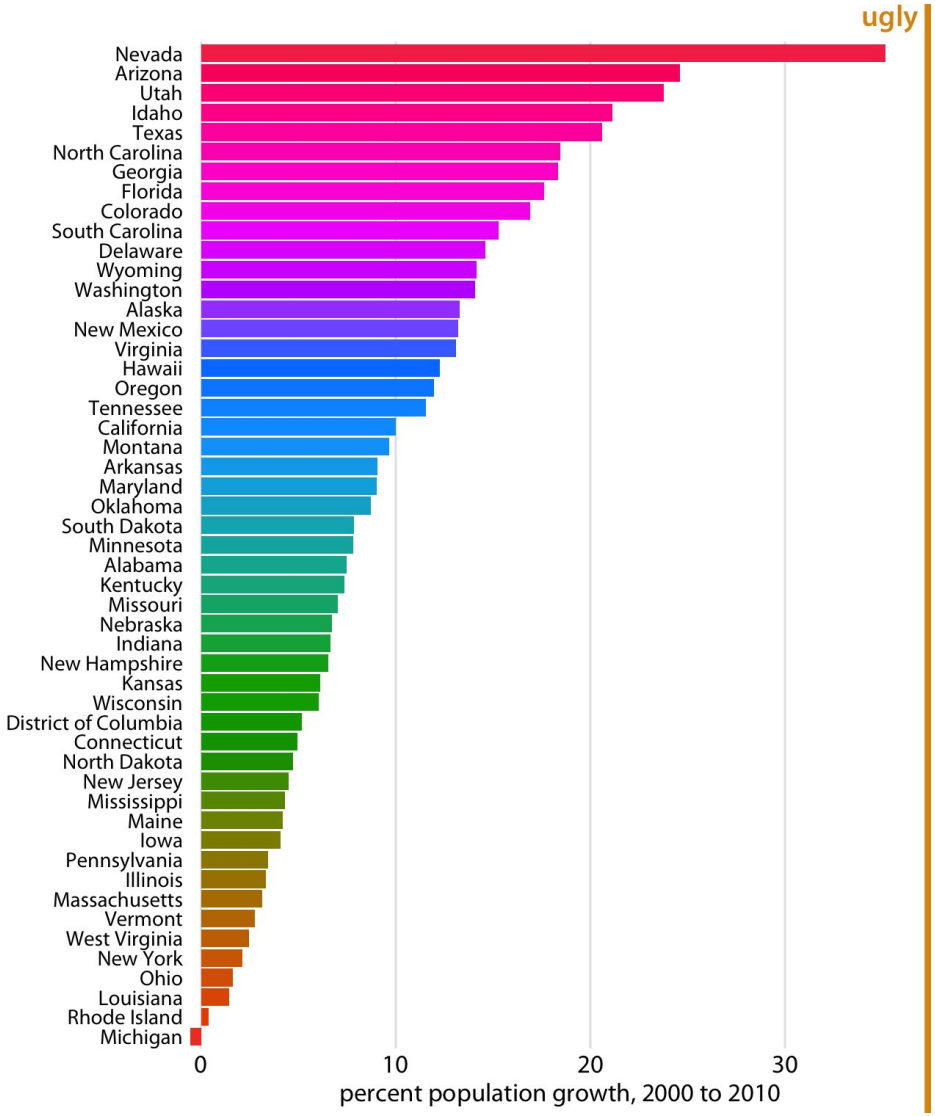


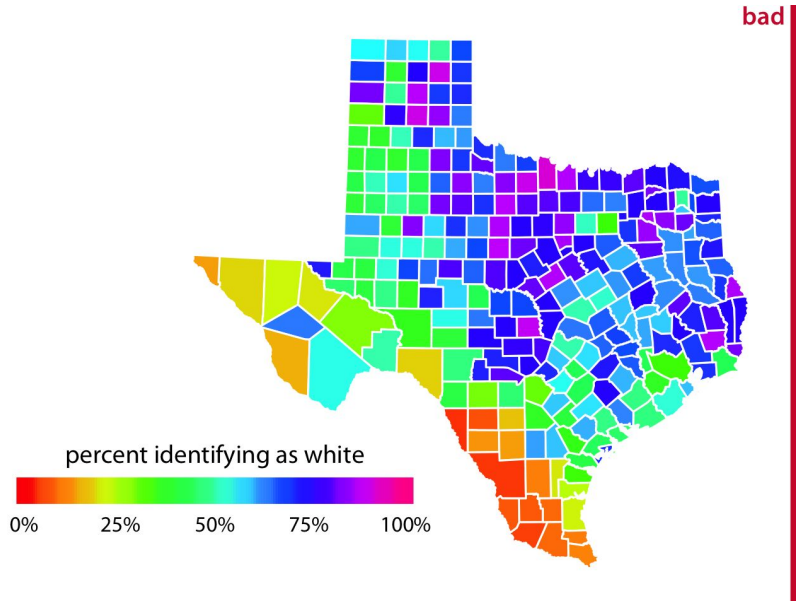
Never use Red-Green!

Use redundant coding: shapes, sizes, etc.

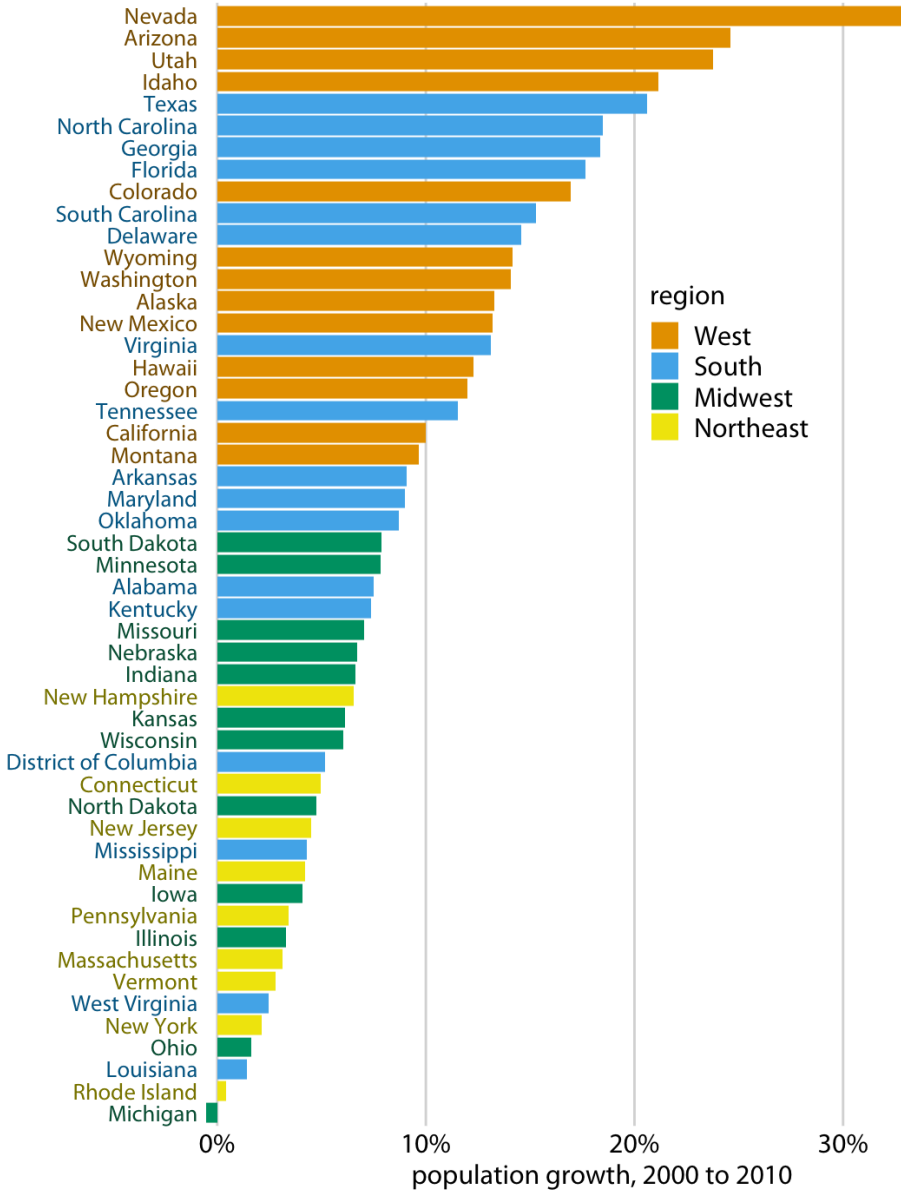
Wilke 2018

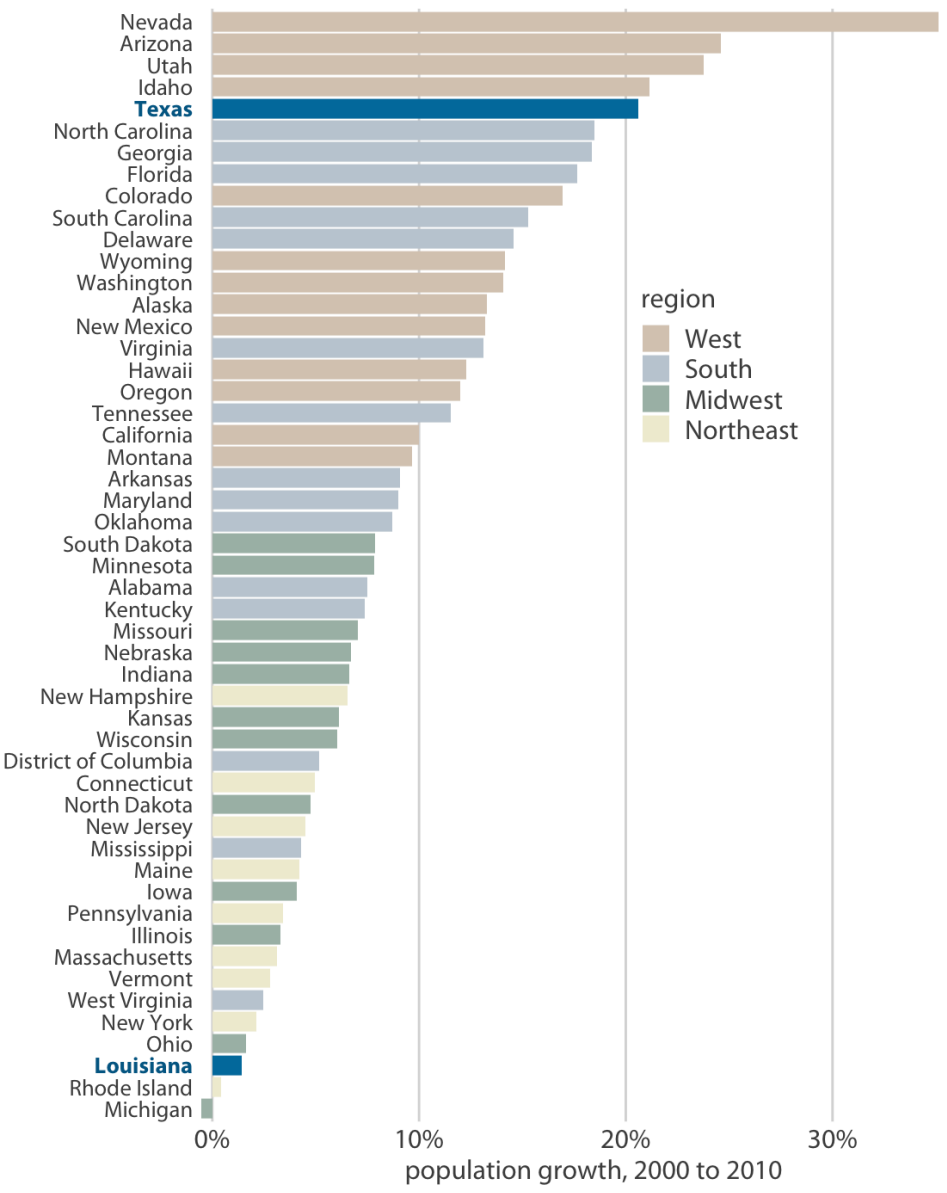
What happens with the wrong palette?



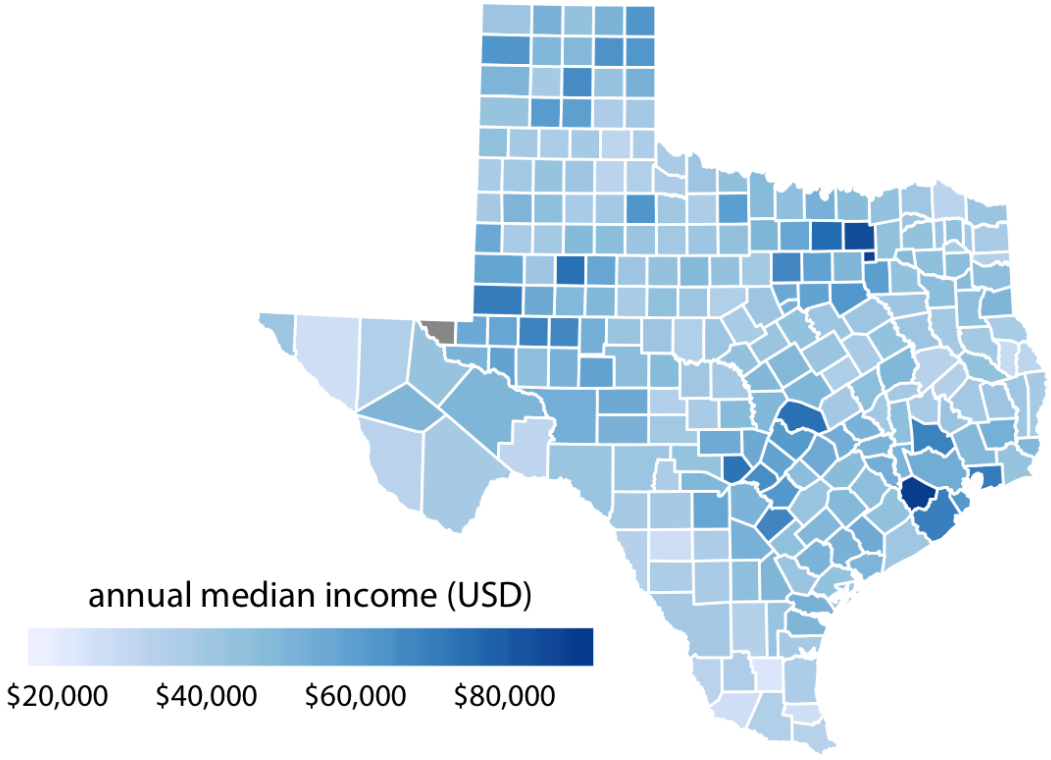


What happens with the wrong palette?

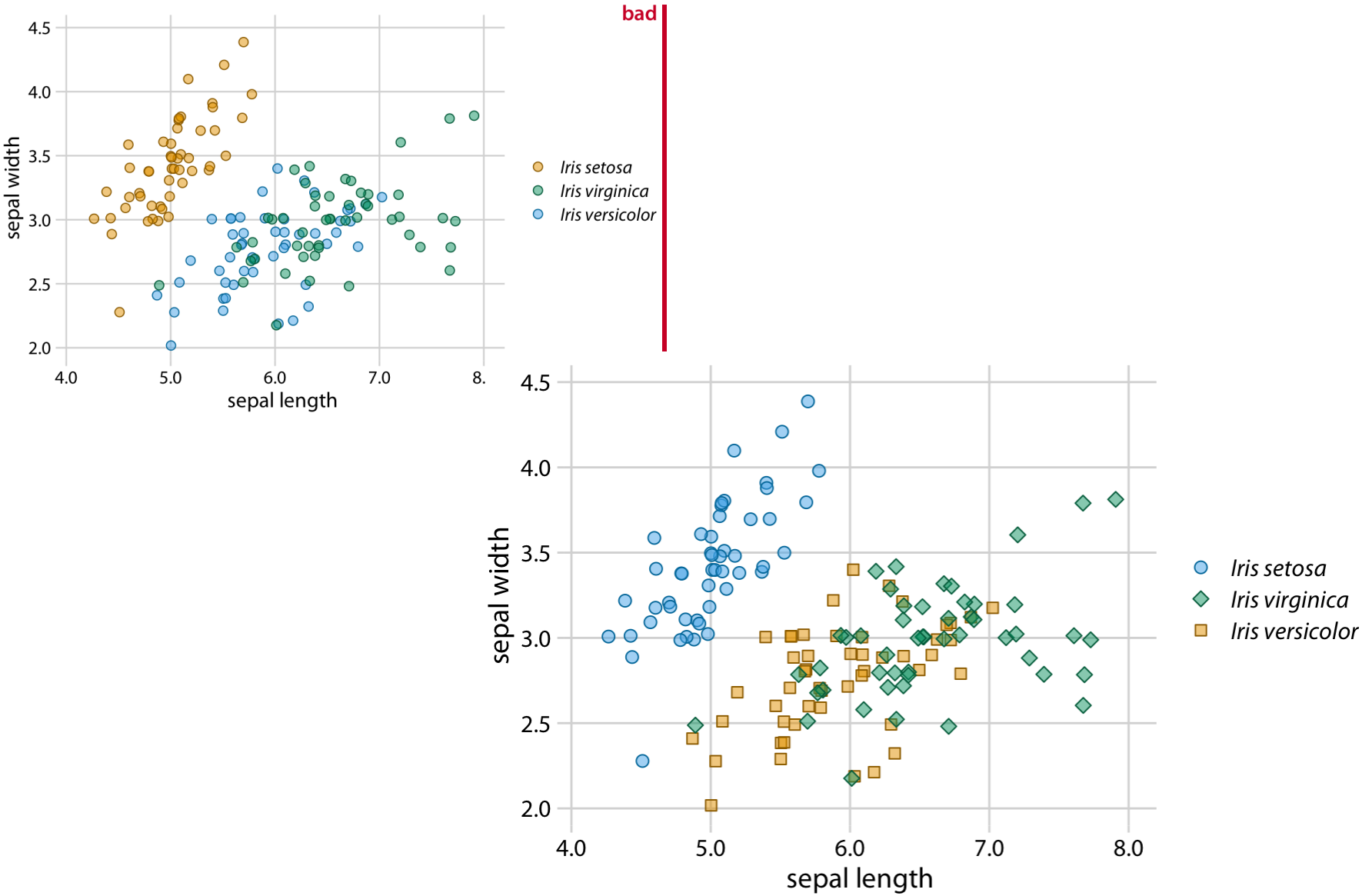




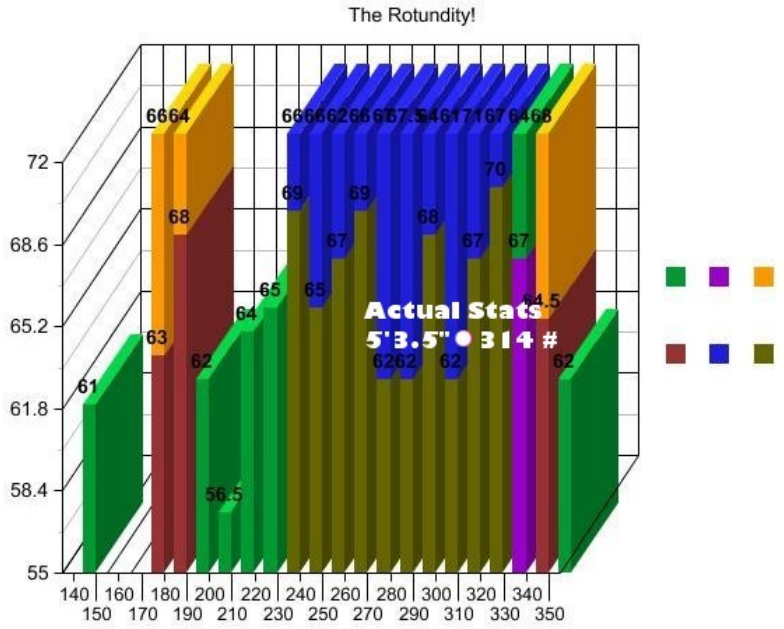
What happens with the wrong palette?

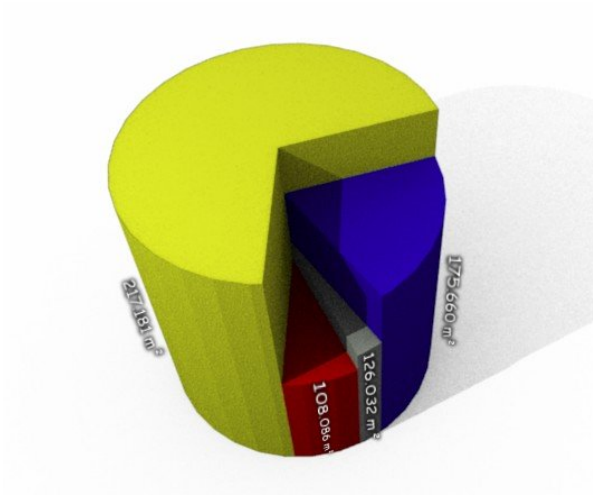


Redundant coding



Bad Visualizations





Summary

- Worthwhile to do visualization of data
  - Humans are great at pattern recognition
  - Helps to understand data distributions, outliers, errors, and other properties
- Many different visualization methods available
- Limitations
  - Generally only useful for up to 3-4 dimensions

Do's and Don't Summarized

- Do
  - Take your time
    Good visualizations take at minimum 10-20 minutes to tweak and refine
    - For an important presentation or paper this can be days of iteration
  - Label everything
  - Think about questions and story
- Do NOT
  - Use 3D unless really necessary
  - Use piecharts (usually)

References

- Kosslyn: Types of Visual Representations
- Lohse et al: How do people perceive common graphic displays
- Bertin, MacKinlay: Perceptual properties and visual features
- Tufte/Wainer: How to mislead with graphs
- Wilke: Fundamentals of Data Visualization https://clauswilke.com/dataviz/index.html