Visualization for Data Science
Exploratory Data Analysis

Cholera in 1854 London

Cholera is an infectious disease that affects the small intestine

In 1854, a cholera outbreak swept the Soho district in London. In the first week, there were more than 150 deaths.   It's 1854 (remember: no computers!) and the government has hired your team to find the outbreak's cause

What does your team suspect the cause to be?
What kind of information would you collect to check?
How would you present that information to solve the mystery?

http://www.who.int/wer/2010/wer8513.pdf

Cholera in 1854 London

What does your team suspect the cause to be?
What kind of information would you collect to check?
How would you present that information to solve the mystery?

Learning Outcomes

- Describe what is exploratory data analysis and why it is important
- Explore a dataset by visualizing various one-dimensional and 2D distributions across multiple categories.
- Inspect and describe relationships between variables.
- Detect missing values and suspicious observations in a dataset.

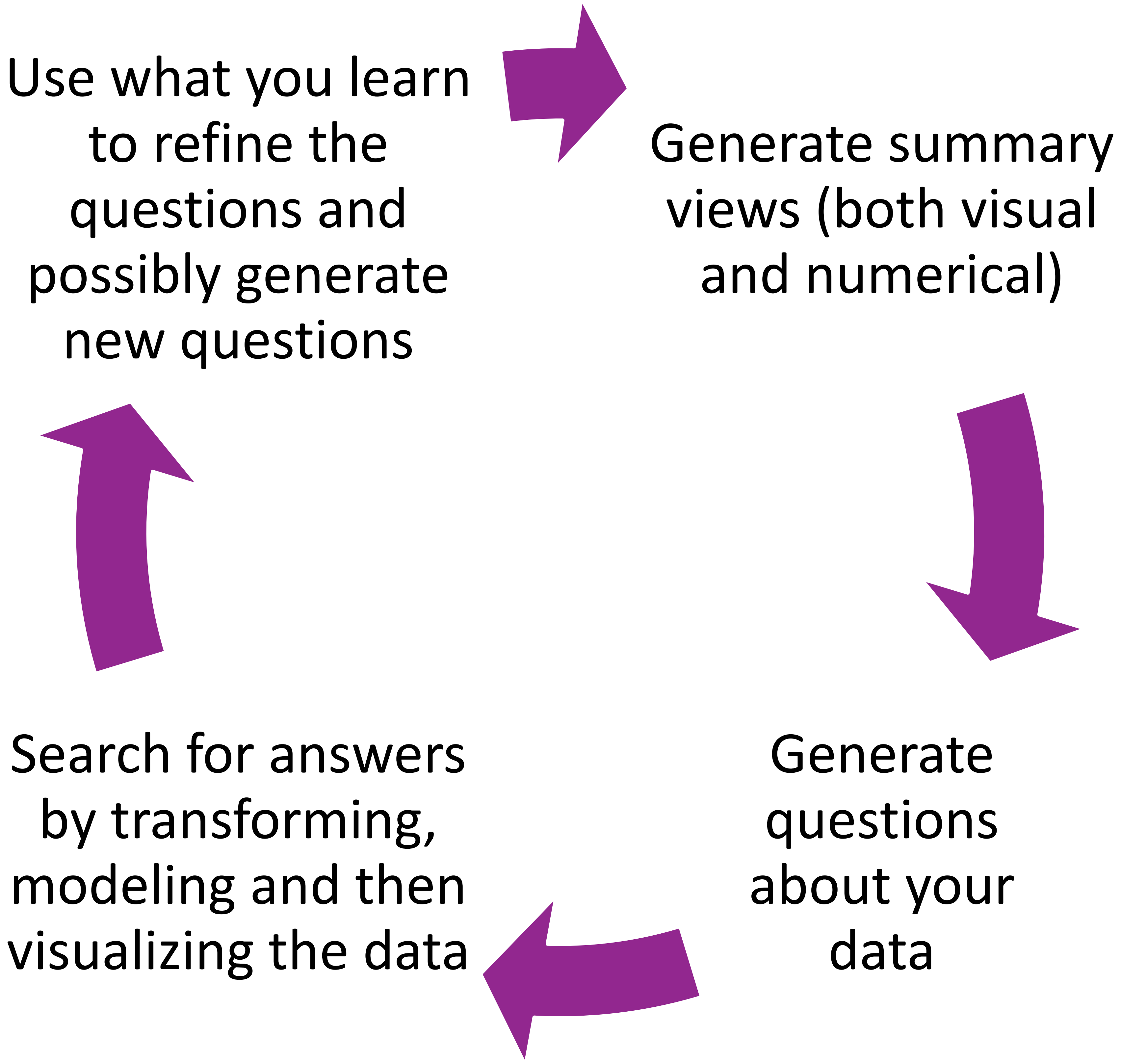
Exploratory Data Analysis (EDA)

Is a process that includes

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Experimental Design and Analysis. Howard J. Seltman

Exploratory Data Analysis Iterative Cycle



Use what you learn to refine the questions and possibly generate new questions

Generate summary views (both visual and numerical)

Generate questions about your data

Search for answers by transforming, modeling and then visualizing the data

https://r4ds.had.co.nz/exploratory-data-analysis.html

Categorizing Exploratory Data Analysis

Medium

- Numerical Summaries
- Visual Data Analysis

Attribute

- Univariate - one column at a time
- Multivariate – two or more variables at a time, looking for relationships.
- Attribute Type – categorical or quantitative

Role

- Outcome
- Explanatory

Categories of EDA

- Univariate Numerical Summaries
- Univariate Visual Idioms
- Multivariate Numerical Summaries
- Multivariate Visual Idioms

Univariate Numerical Summaries

- Categorical Variable
  - Range of values
  - Frequency of each value (proportion)
- Quantitative Variable
  - Distribution: center, spread, modality, shape and outliers
  - Central Tendency: mean, mode, median
  - Spread: variance, standard deviations, interquartile range

Tukey advocates for focusing on max, min, median and quartiles

Experimental Design and Analysis. Howard J. Seltman

Univariate Visual Idioms

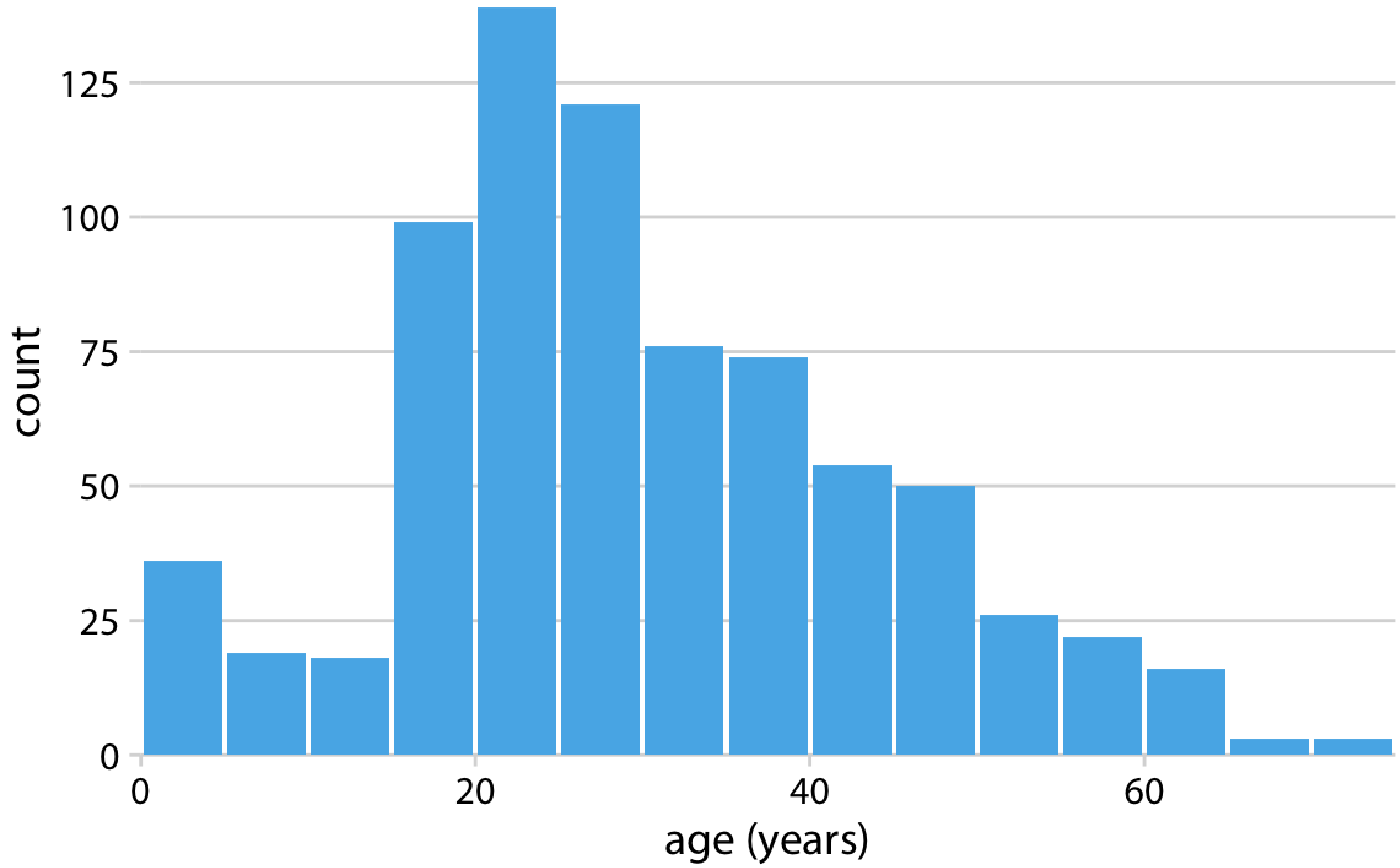
- Distribution of Categorical Variable
  - Bar Charts
- Distribution of Quantitative Variable
  - Histogram
  - Density Plots

Univariate Visual Idiom: Histograms

Histograms are used to visualize the shape, center, range and variation for a continuous variable.

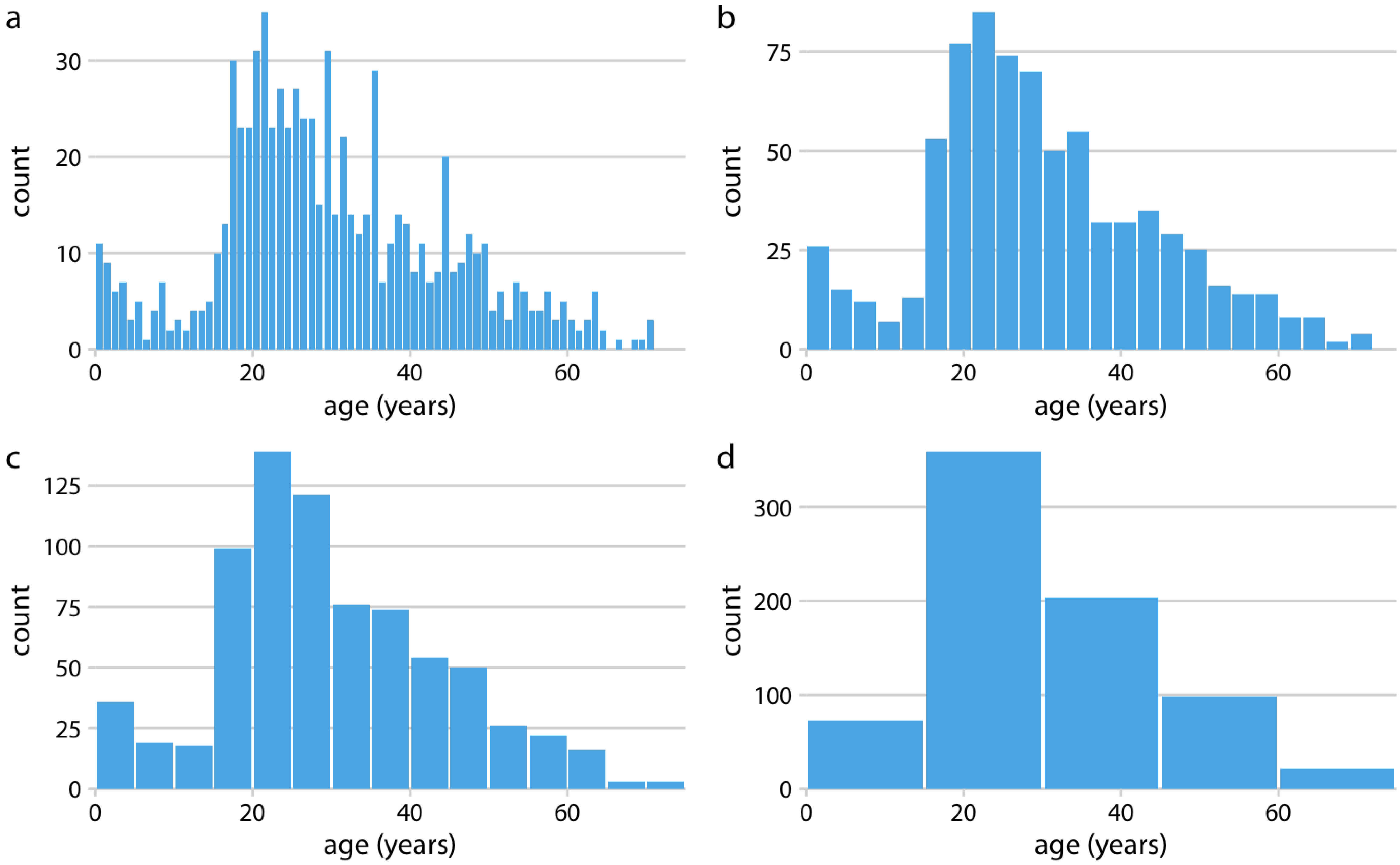
The size of the bin is extremely important.

Design Tip: Vary bin size during EDA



https://clauswilke.com/dataviz/histograms-density-plots.html

Univariate Visual Idiom: Histograms

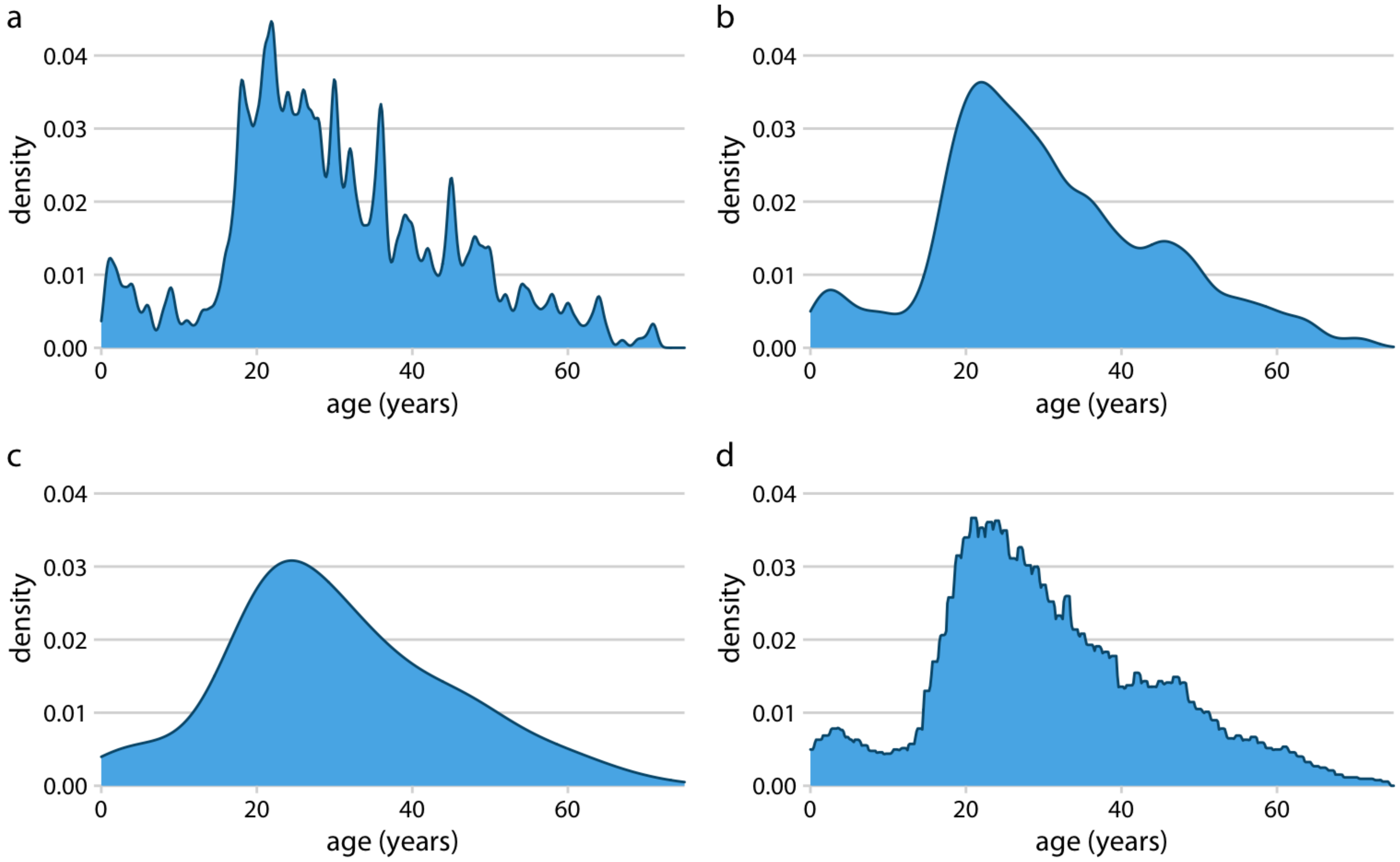


Histograms depend on the chosen bin width. Here, the same age distribution of Titanic passengers is shown with four different bin widths: (a) one year; (b) three years; (c) five years; (d) fifteen years.

https://clauswilke.com/dataviz/histograms-density-plots.html

Univariate Visual Idioms: Density Plot

A density plot is a representation of the distribution of a numeric variable. It uses the kernel density estimate to show the probability density function of a variable. It is basically a smoothed out version of the histogram.



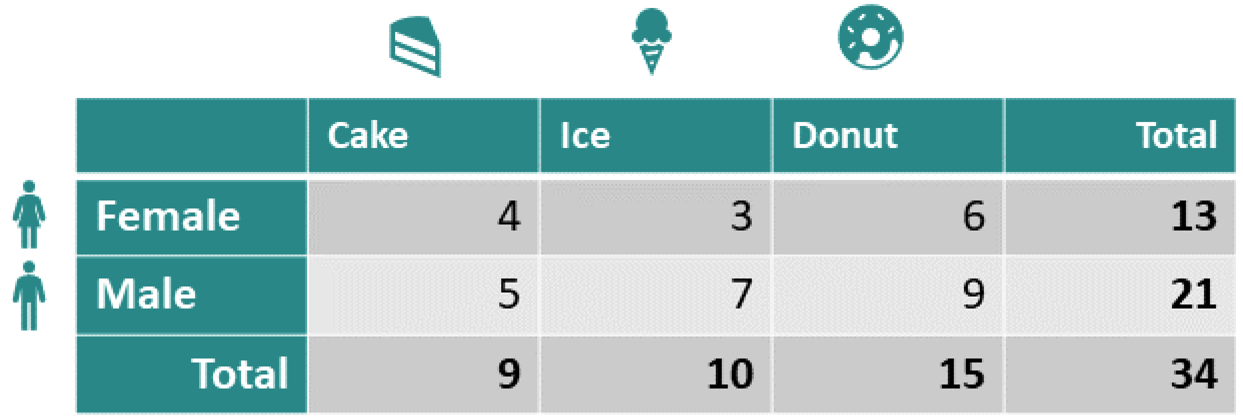
Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: (a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2.

https://clauswilke.com/dataviz/histograms-density-plots.html

Multivariate Numerical Summaries

Categorical Variable

- cross-tabulation
- Univariate statistics by category



	Cake	Ice	Donut	Total
Female	4	3	6	**13**
Male	5	7	9	**21**
Total	**9**	**10**	**15**	**34**

Quantitative Variable

- Correlation & Covariance: a measure of how much (and in what direction) should we expect one variable to change when the other changes.
- Correlation & Covariance Matrix for >2 variables

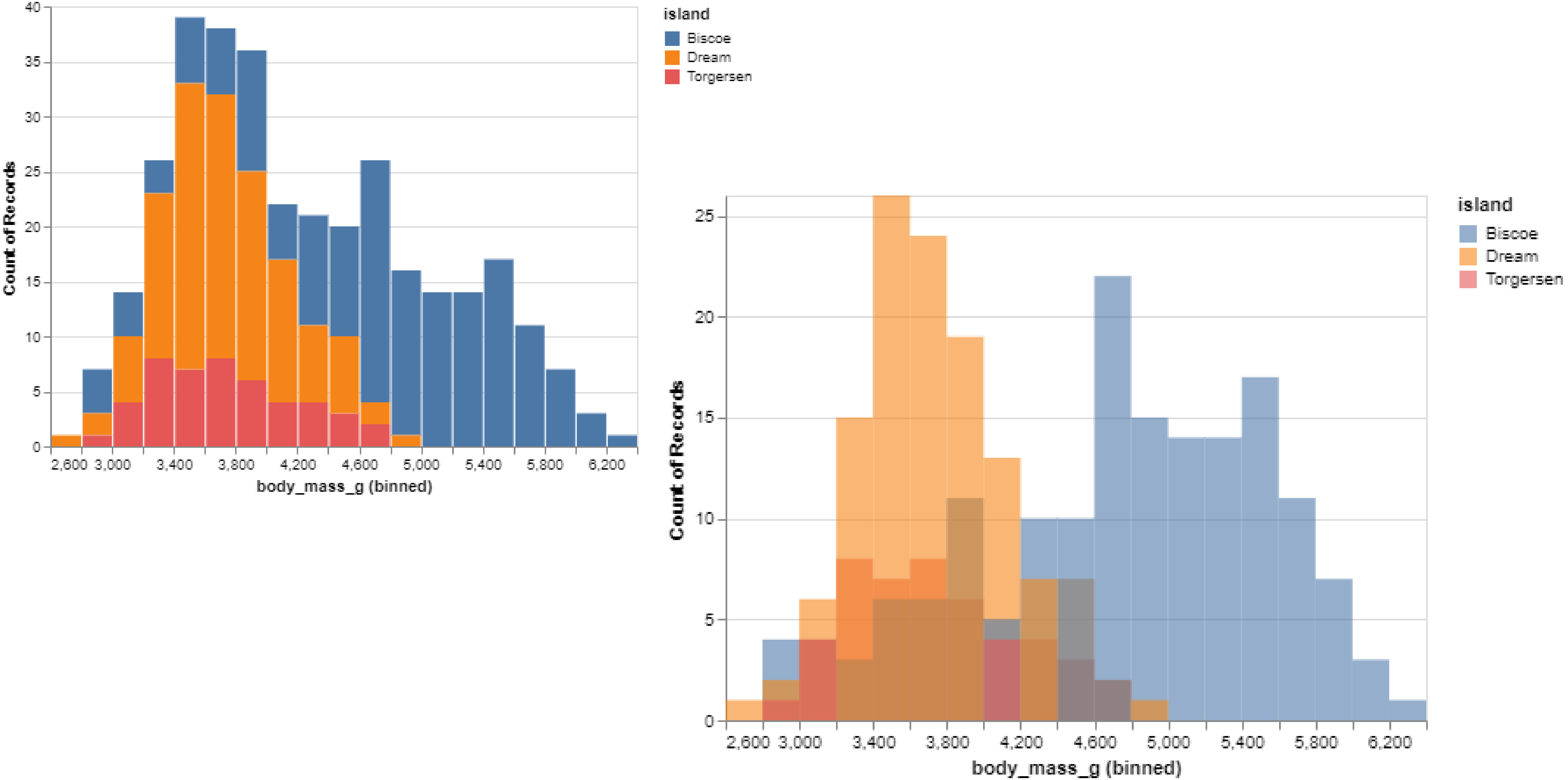
Experimental Design and Analysis. Howard J. Seltman

https://datatab.net/tutorial/cross-table

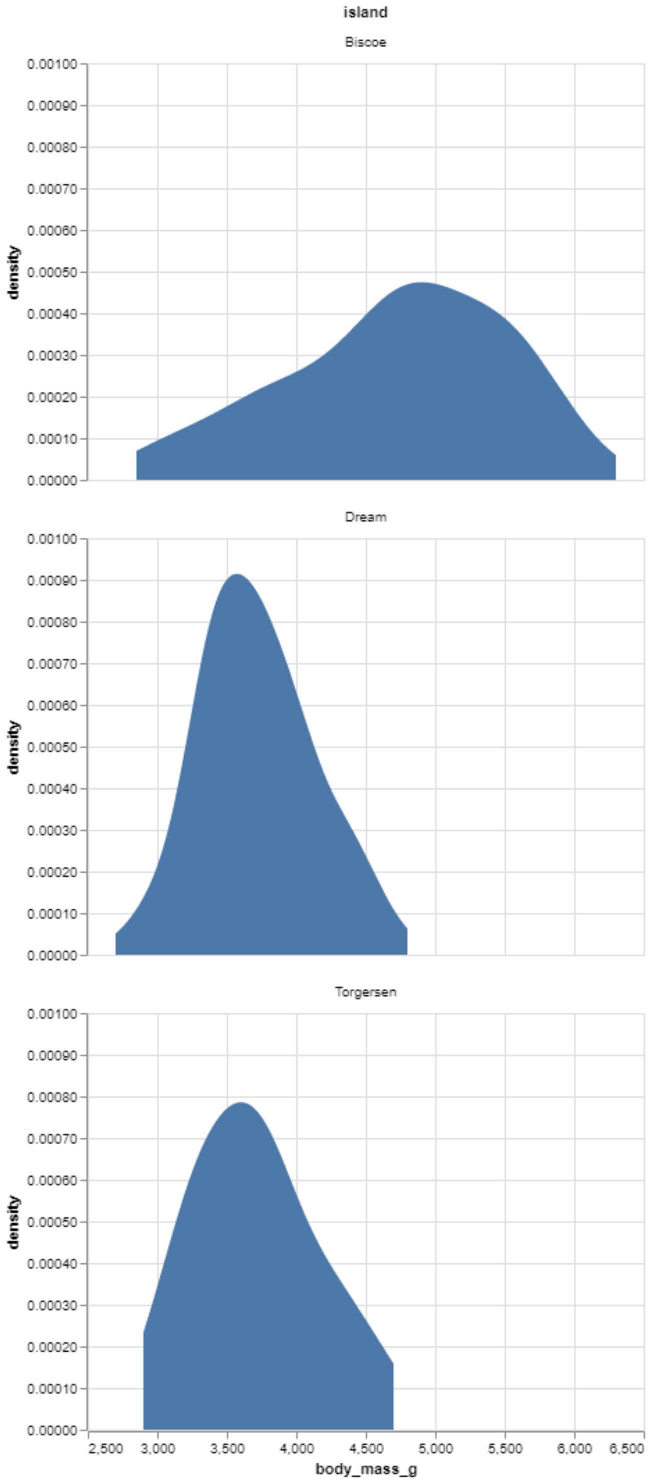
Multivariate Visual Idioms

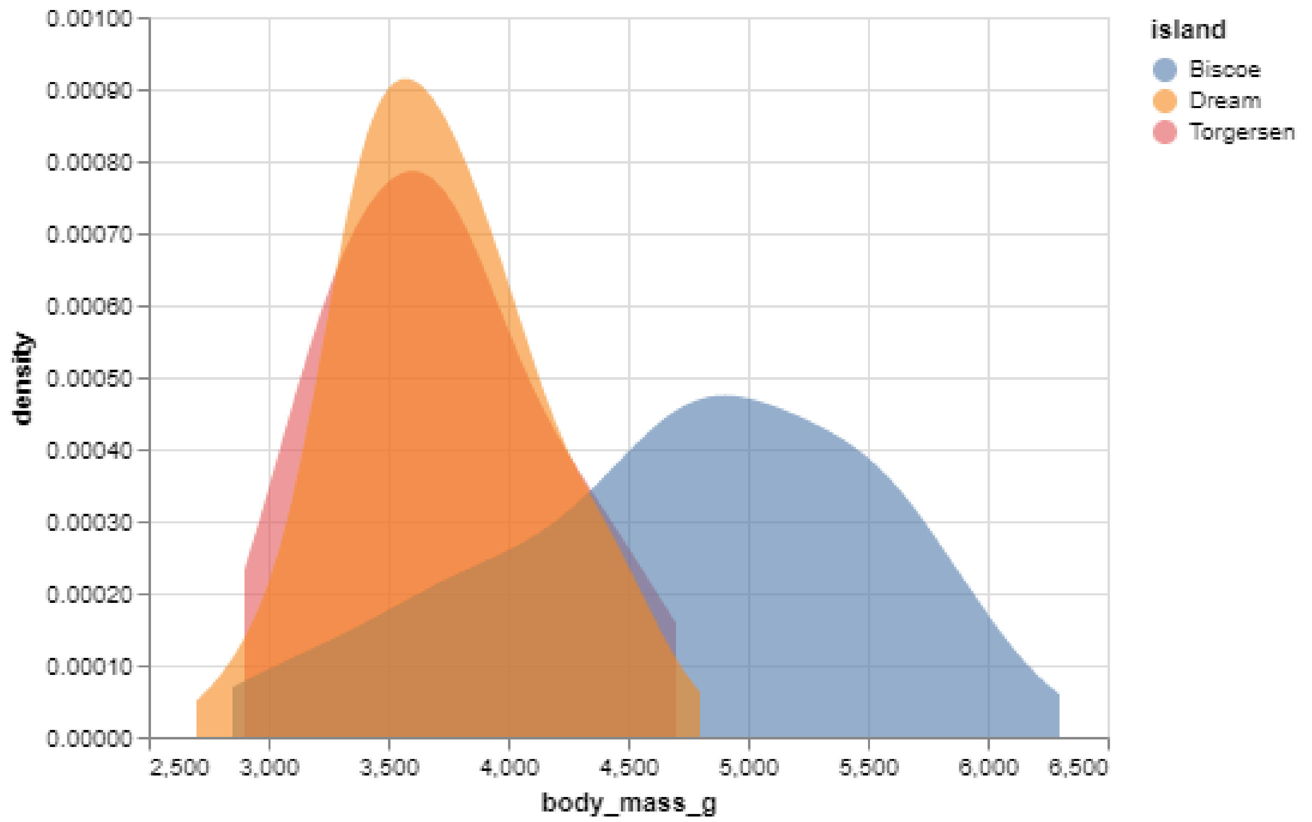
- Categorical Data
  - Stacked Bar Charts
- Quantitative Data
  - Overlapping Density Plots
  - Scatterplots
  - Box-plots & Violin Plots (see T8)
  - Univariate graphs by category – typically used when we have one explanatory variable (categorical) and one outcome (quantitative) variable

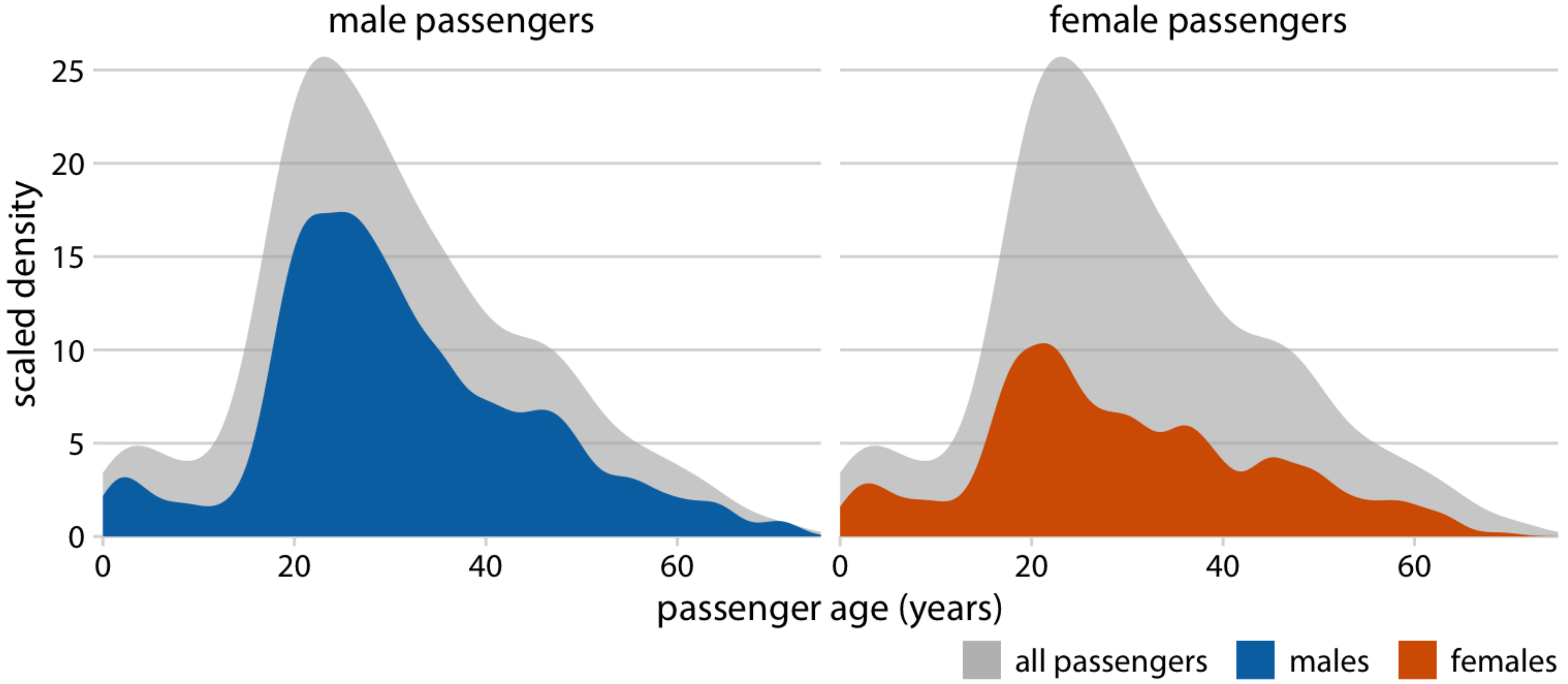
Bivariate Visual Idioms: Stacked Bar Chart



Multivariate Visual Idioms: Overlapping & Faceted Density Plots

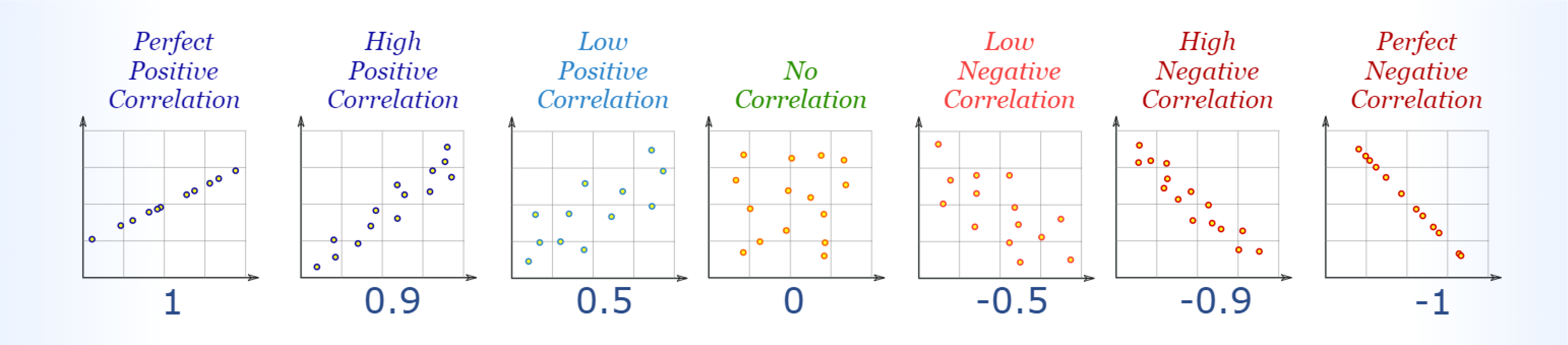






https://clauswilke.com/dataviz/histograms-density-plots.html

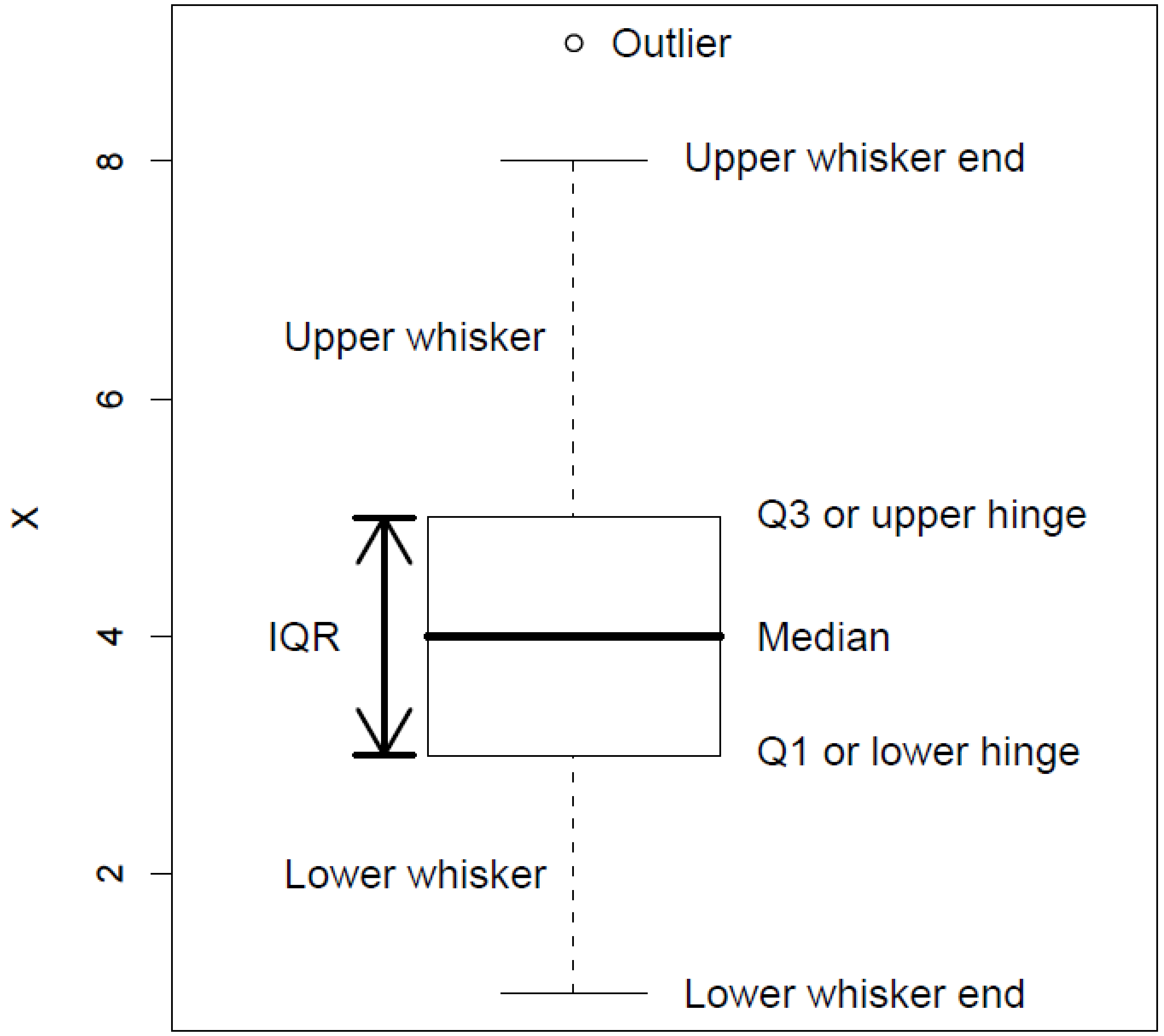
Bivariate Visual Idioms: Scatterplot



https://www.mathsisfun.com/data/scatter-xy-plots.html

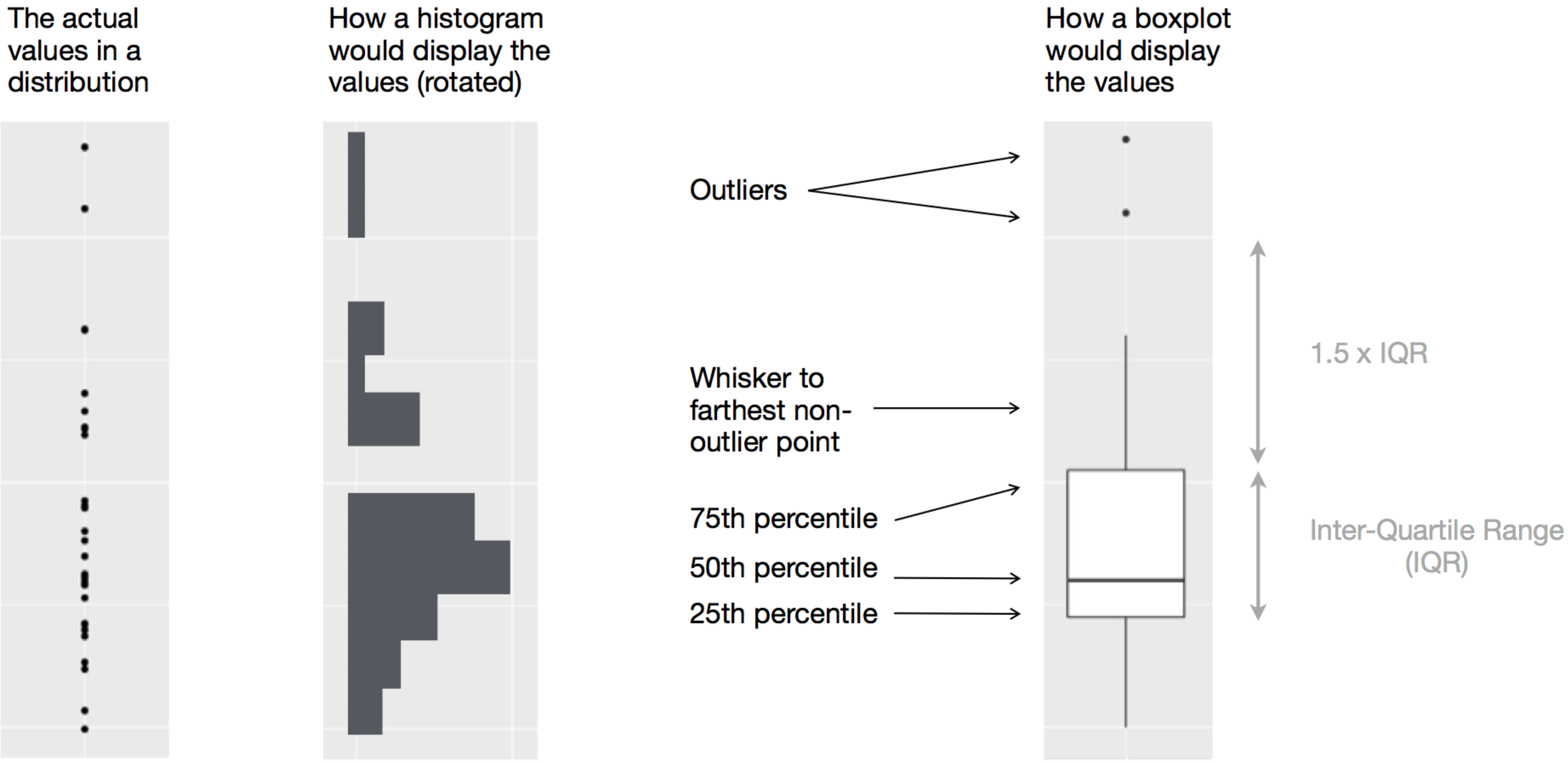
Multivariate Visual Idioms: Boxplots

Very good at representing data related to the central tendency, symmetry and skew.



Experimental Design and Analysis. Howard J. Seltman

Multivariate Visual Idioms: Boxplots


The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

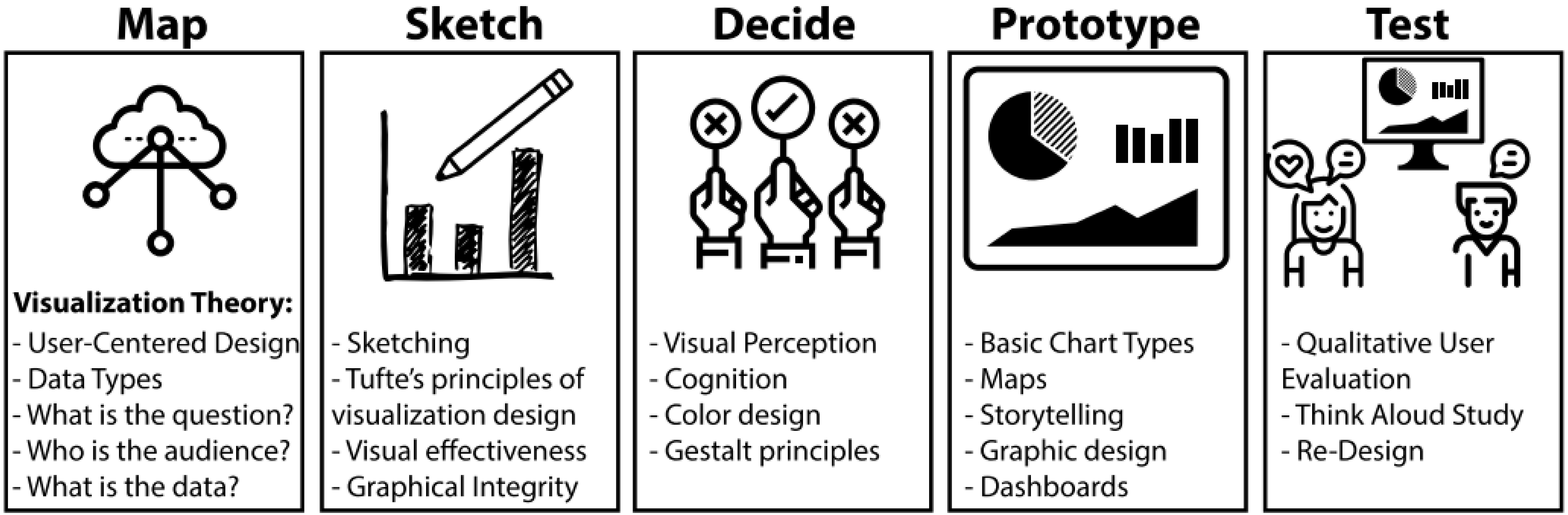
https://r4ds.had.co.nz/exploratory-data-analysis.html

“In a nutshell: You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science ***{ it is a very important art!}****”*

- Howard J. Seltman

Feedback Time

https://ubc.ca1.qualtrics.com/jfe/form/SV_20kDF8H6woQf1qu



Map

Visualization Theory:

- User-Centered Design
- Data Types
- What is the question?
- Who is the audience?
- What is the data?

Sketch

- Sketching
- Tufte's principles of visualization design
- Visual effectiveness
- Graphical Integrity

Decide

- Visual Perception
- Cognition
- Color design
- Gestalt principles

Prototype

- Basic Chart Types
- Maps
- Storytelling
- Graphic design
- Dashboards

Test

- Qualitative User Evaluation
- Think Aloud Study
- Re-Design

https://ieeexplore.ieee.org/document/9547834

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.
- Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Graphical excellence is nearly always multivariate.
- And graphical excellence requires telling the truth.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*.

Sketching Basics

- Not a prototype
- No artistic skill required
- Helpful in making decisions

Sketch Individual Exercise – 10 minutes

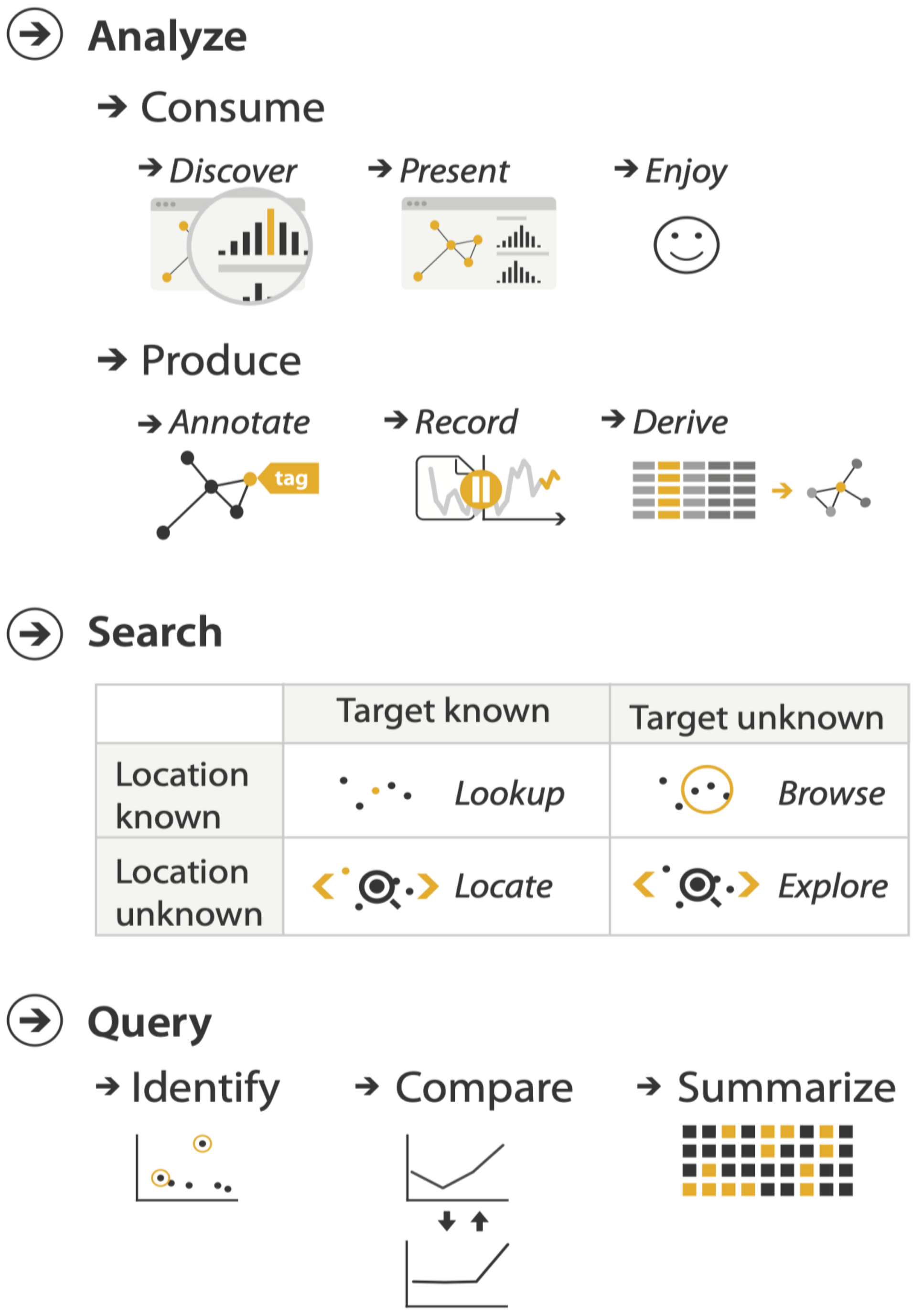
- Dataset - https://www.thesquirrelcensus.com/data

You have a list of questions that was generated by your peers, your task for today is to first create sketches for each question. In addition, you need to create an aesthetic pleasing novel visualization for at least one of the questions you have.

NOTE: This is purely brainstorming visualizations, do not get stuck in a technique, create low fidelity representations. Do not critique the visualizations of others or even yours, just do it.

Task Abstraction

Different ways to classify your tasks (high, mid or low), different levels of abstraction



Analyze

- Consume
  - *Discover*
  - *Present*
  - *Enjoy*
- Produce
  - *Annotate*
  - *Record*
  - *Derive*

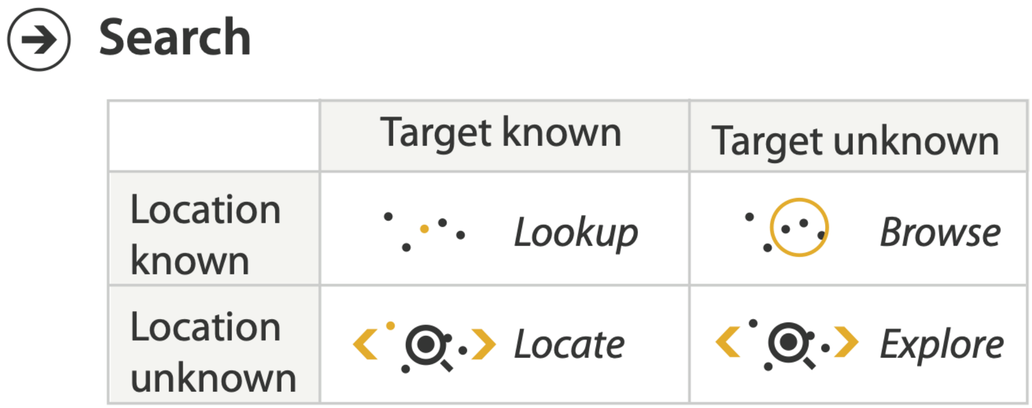
Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

- Identify
- Compare
- Summarize

- Lookup
  - What is the address of IKB Learning Center
- Locate
  - Where is IKB Learning Center
- Browse
  - What buildings are near IKB Learning Center
- Explore
  - What is south of Agronomy



Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Low-level classification of tasks

- Retrieve Value - How long is the movie Gone with the Wind?
- Filter - What comedies have won awards?
- Compute Derived Value - How many awards have MGM studio won in total?
- Find Extremum - What director/film has won the most awards?
- Sort - Rank movies by most number of awards won
- Determine Range - What is the range of film lengths?
- Characterize Distribution - What is the age distribution of actors?
- Find Anomalies - Are there exceptions to the relationship between number of awards won and total movies made by an actor?
- Cluster - Is there a cluster of typical film lengths?
- Correlate - Is there a trend of increasing film length over the years?

https://faculty.cc.gatech.edu/~stasko/papers/infovis05.pdf

Decide Group Exercise – 20 minutes

- Group size 3 – 4
- Dataset - https://www.thesquirrelcensus.com/data

You have a list of questions that was generated by your peers, now that you have your sketches, you need to fine tune the questions and then decide which sketch is most appropriate for the question/task. Now you are working in your group so first come up with a way to offer feedback and make decisions.