Announcements

HW1

- Due Tue 2/8

Plan

Wrap up Relational Data

- SQLite examples
- DB joins

Visualization and Data Exploration

15-388/688 - Practical Data Science: Visualization and Data Exploration

Pat Virtue
Carnegie Mellon University
Spring 2022

Slide credits: CMU AI, Zico Kolter

Outline

Basics of visualization

Data types and visualization types

→ Software plotting libraries

Outline

Basics of visualization

Data types and visualization types

Software plotting libraries

Two types of visualization

Data exploration visualization: figuring out what is true

Data presentation visualization: convincing other people it is true

This lecture will mostly be focused on the first, some later lectures will touch on the second

“Data exploration” is much broader than just visualization (most of the analysis techniques we will cover fit into it)

Importance of visualization

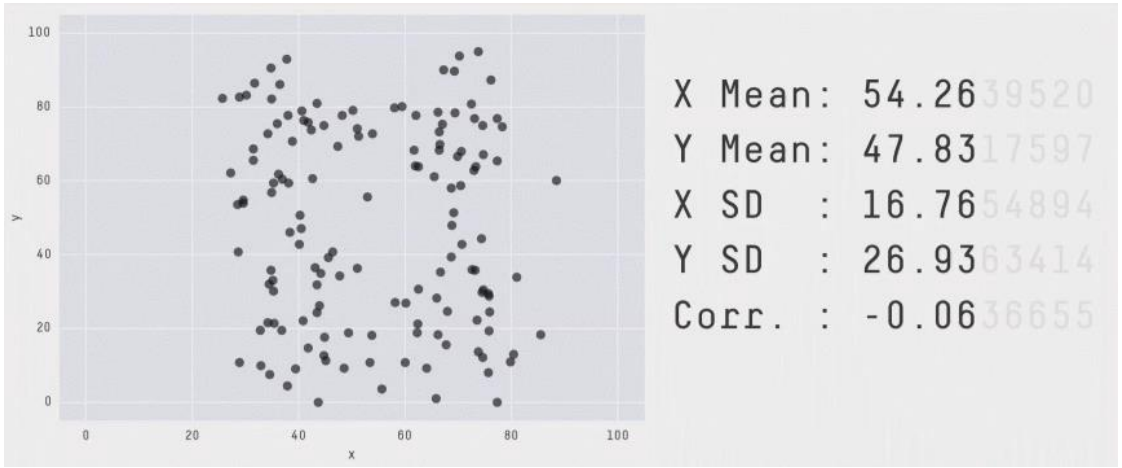
Before you run any analysis, build any machine learning system, etc, always visualize your data

If you can't identify a trend or make a prediction for your dataset, it's unlikely that an automated algorithm will

This is especially important to keep in mind as you hear stories of "superhuman" performance of AI methods (it is possible, but takes a long time, and is not the norm)

Visualization vs. statistics

Visualization almost always presents a more informative (though less quantitative) view of your data than statistics (the noun, not the field)



[Source: https://twitter.com/JustinMatejka/status/770682771656368128 Credit: @JustinMatejka, @albertocairo]

This is a mathematical property: n data points and m equations to satisfy, with $n > m$

Outline

Basics of visualization

Data types and visualization types

Software plotting libraries

Data types

Nominal: categorical data, no ordering

Example – Pet: {dog, cat, rabbit, …}

Operations: $=, \neq$

Ordinal: categorical data, with ordering

Example – Rating: {1,2,3,4,5}

Operations: $=, \neq, \geq, \leq, >, <$

Interval: numerical data, zero doesn't mean zero "quantity"

Example – Temperature Fahrenheit

Operations: $=, \neq, \geq, \leq, >, <, +, -$

Ratio: numerical data, zero has meaning related to zero "quantity"

Example – Temperature Kelvin

Operations: $=, \neq, \geq, \leq, >, <, +, -, \div$

Poll 1

What data type is temperature data in Celsius

A. Nominal

B. Ordinal

✓ C. Interval

D. Ratio

Poll 2

What data type is data for number of children?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

Visualization Types

Most discussion of visualization types emphasizes what elements the chart is trying to convey

Instead, we are going to focus on the type and dimensionality of the underlying data

Visualization types (not an exhaustive list):

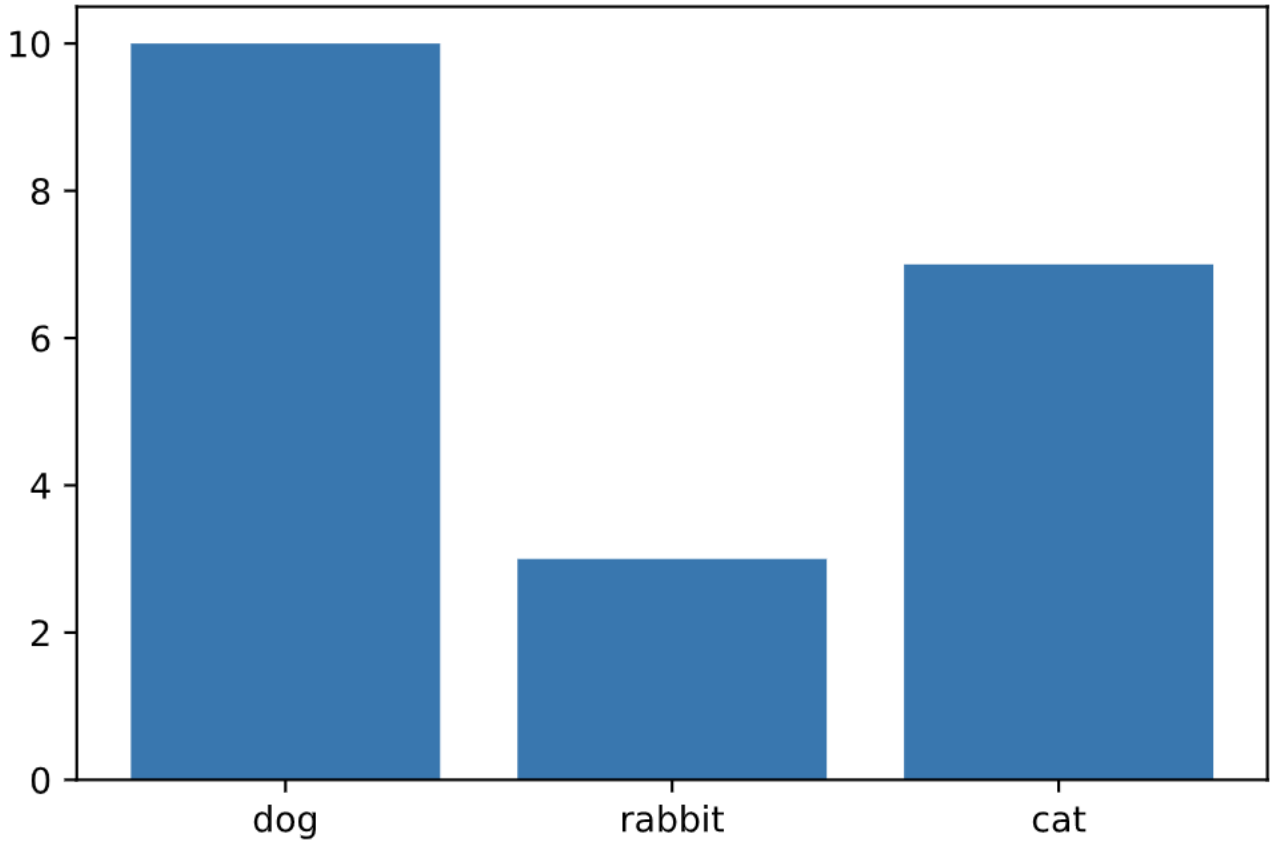
- 1D: bar chart, pie chart, histogram
- 2D: scatter plot, line plot, box and whisker plot, heatmap
- 3D+: scatter matrix, bubble chart

1D DATA

Bar chart

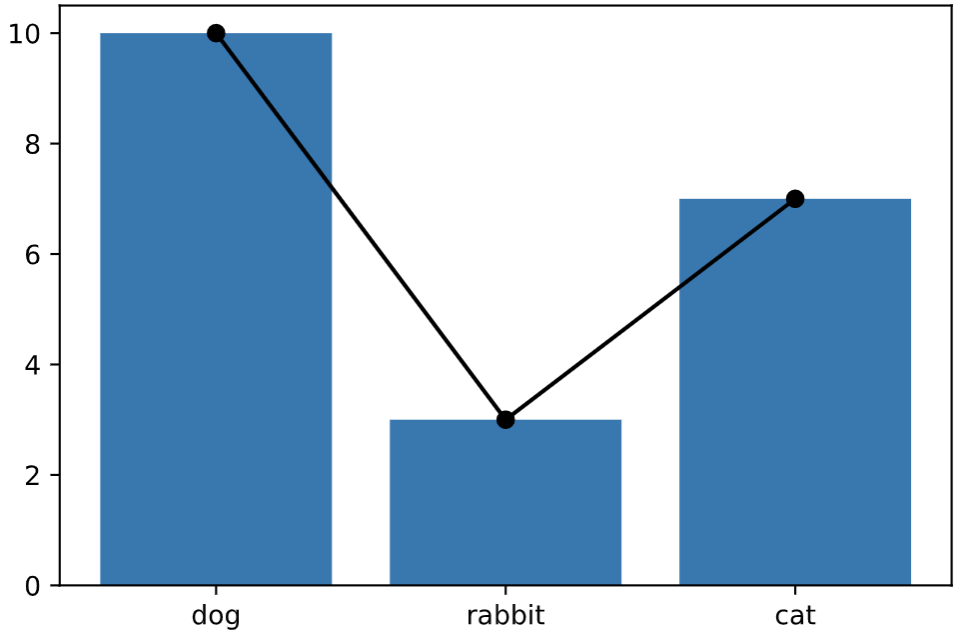
	Data
Nominal	✔
Ordinal	✔
Interval	✗
Ratio	✗

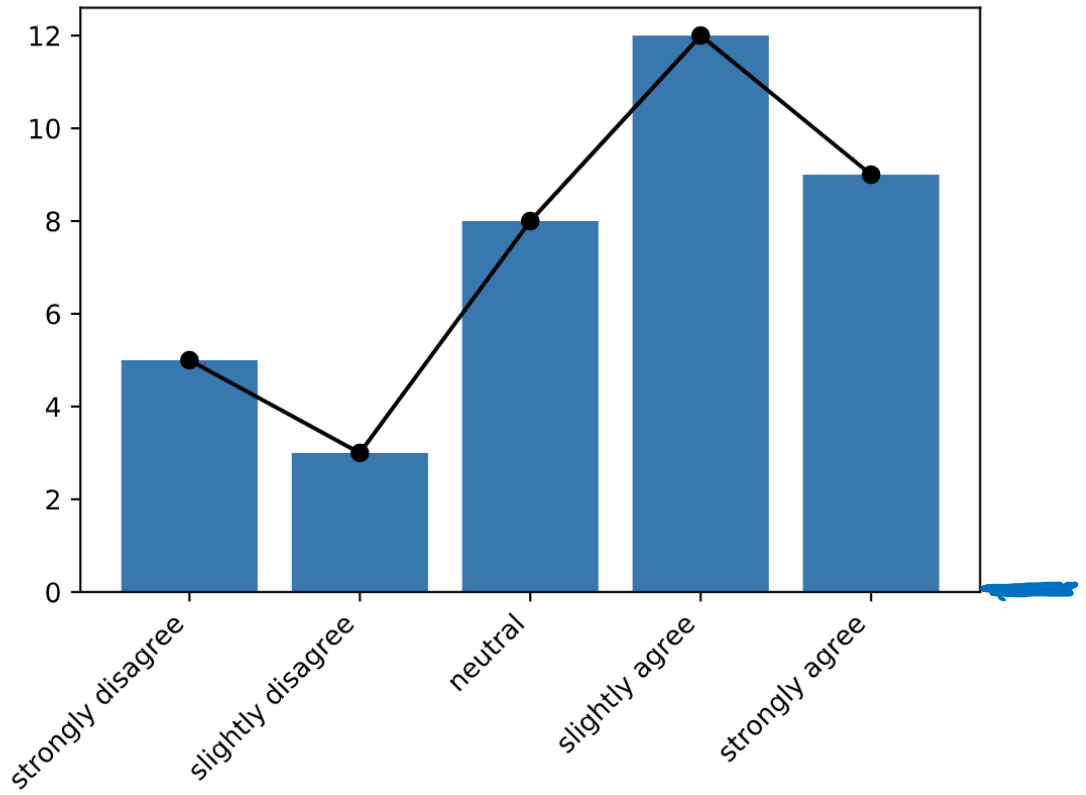
Suggestions, not rules



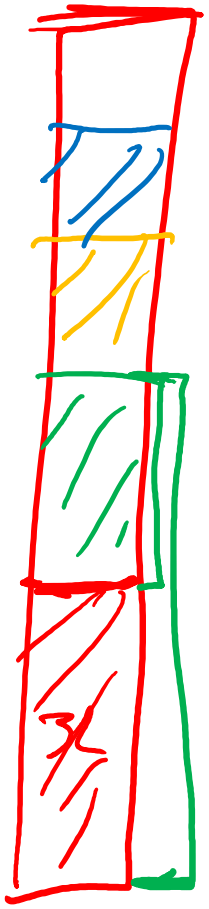
Bar chart (bad)

Don't use lines within a bar chart for categorial or ordinal features!

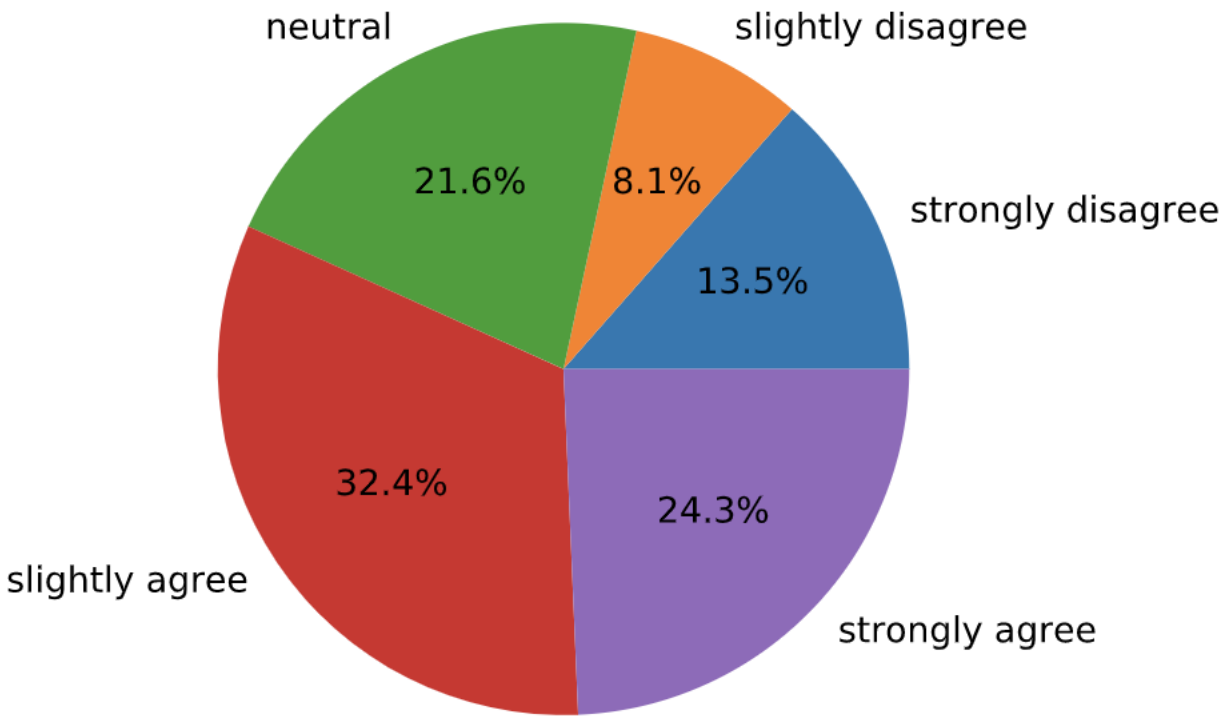




Pie chart

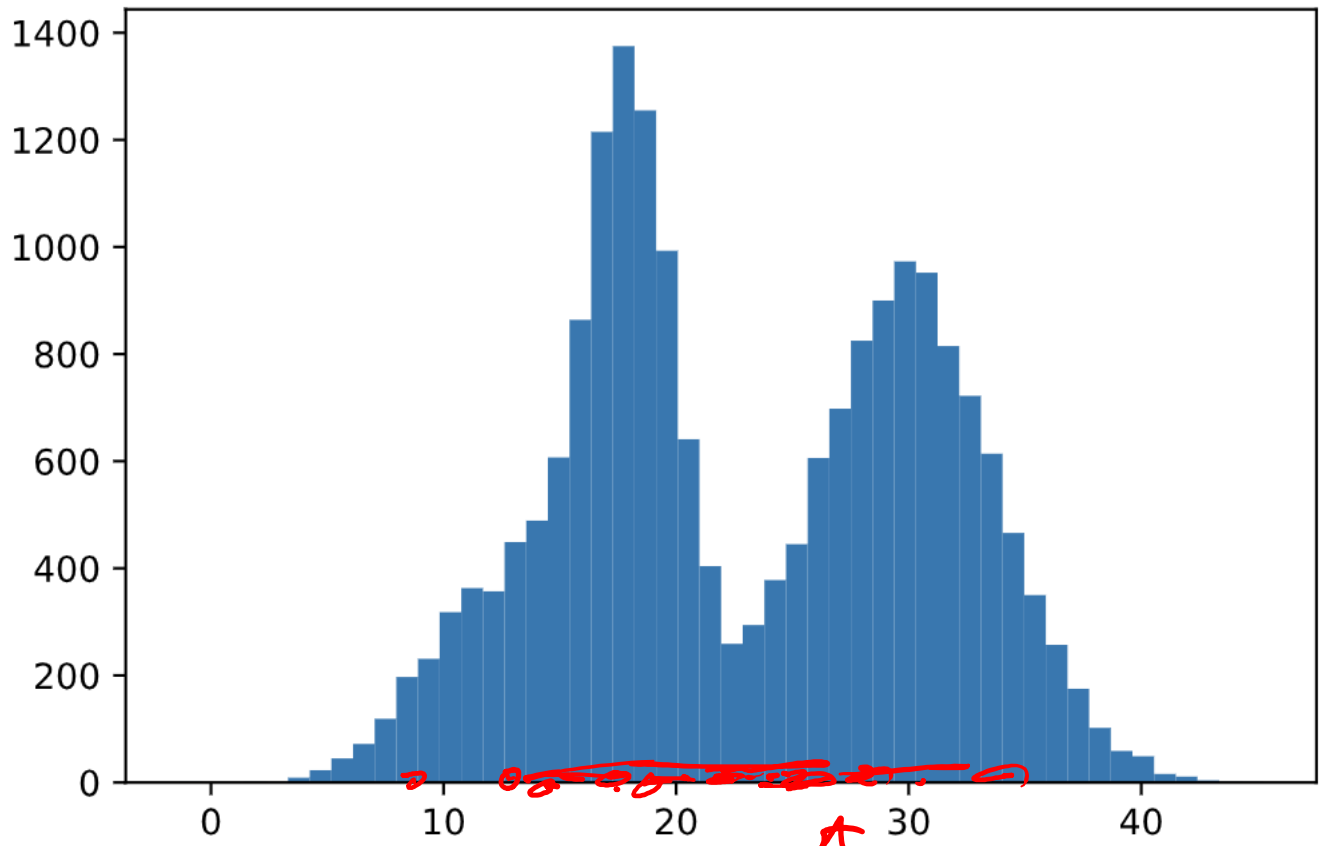


	Data
Nominal	✗
Ordinal	✗
Interval	✗
Ratio	✗



Histogram

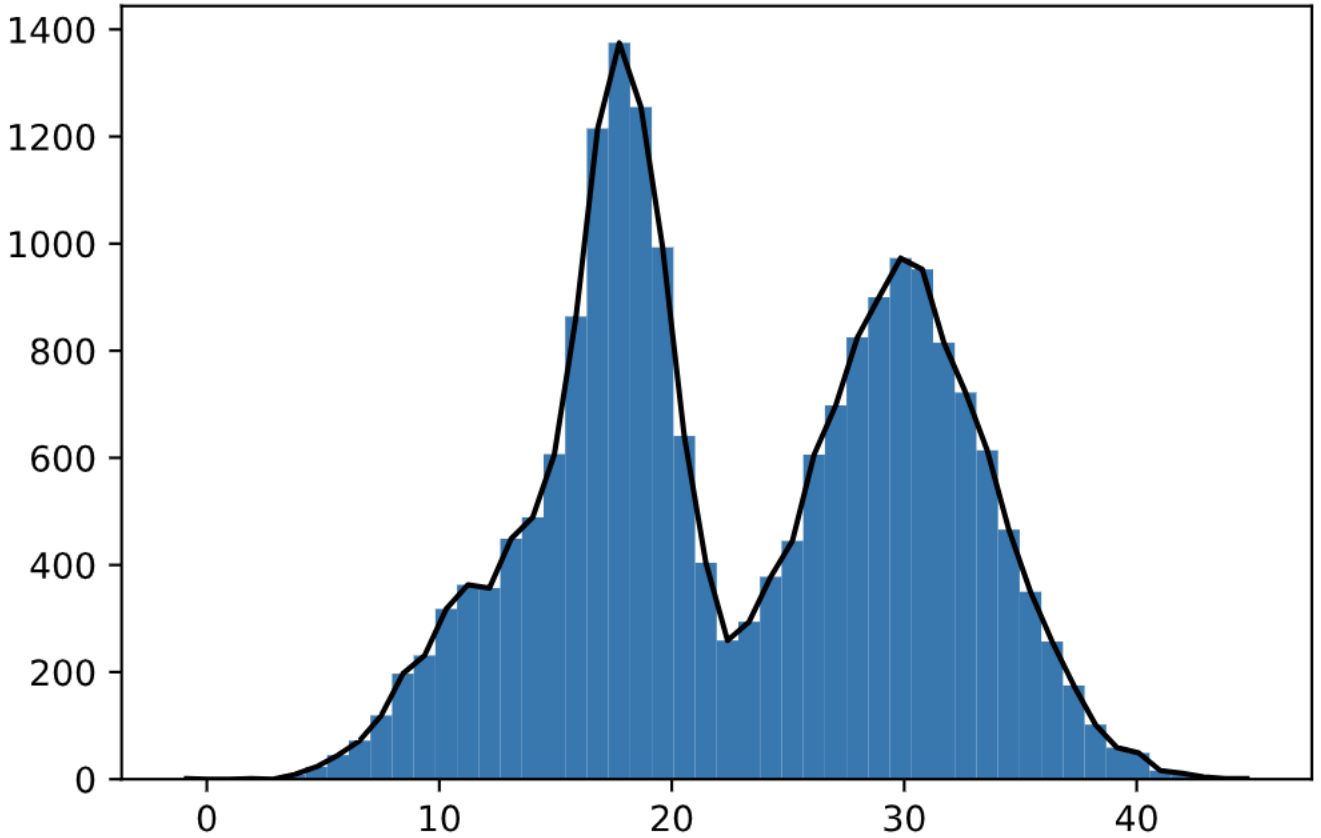
	Data
Nominal	✗
Ordinal	✗
Interval	✔
Ratio	✔



Data → Count per bin (range of values) → Plot with bar chart

Histogram

OK to use lines within a histogram (but not very informative)

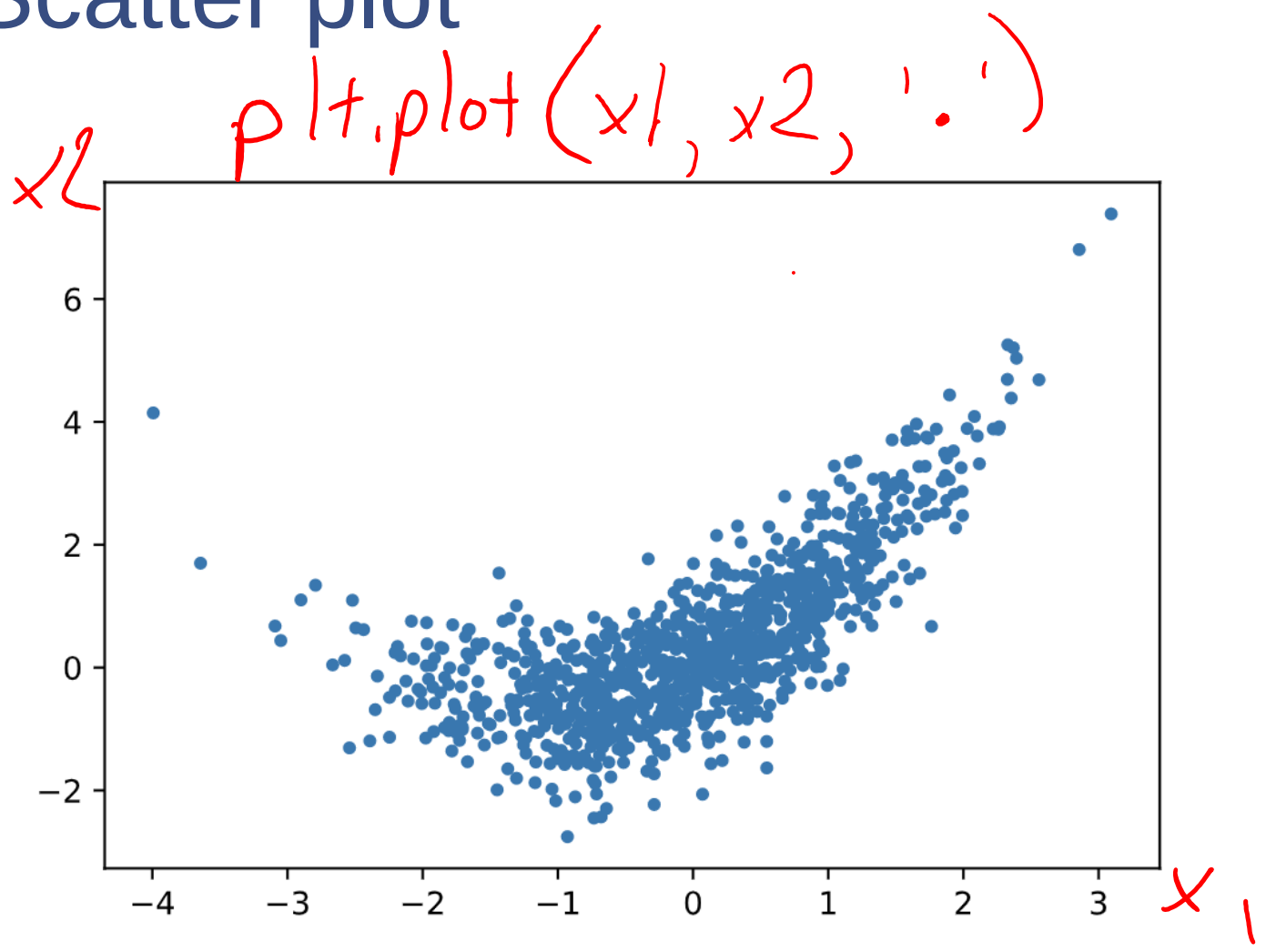


2D DATA

Scatter plot

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓

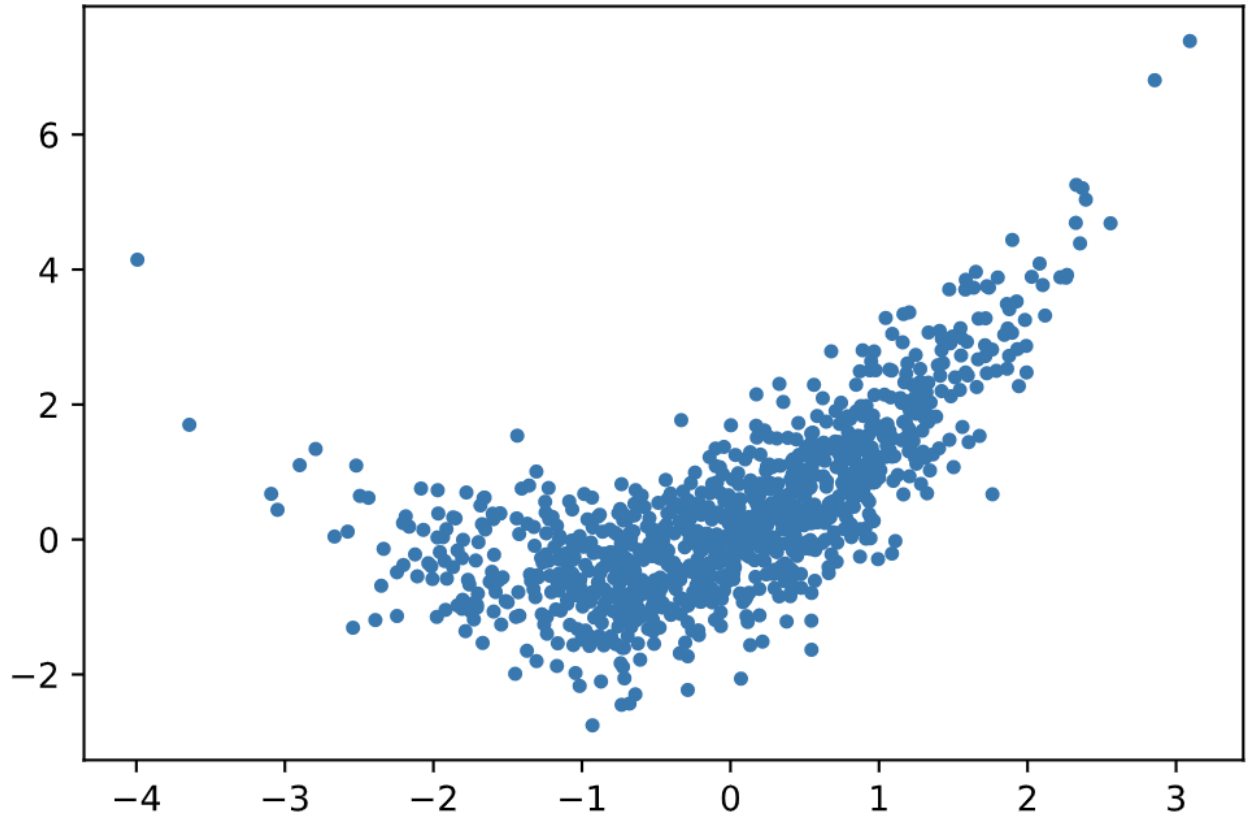
Why not ordinal data in first dimension?



Scatter plot

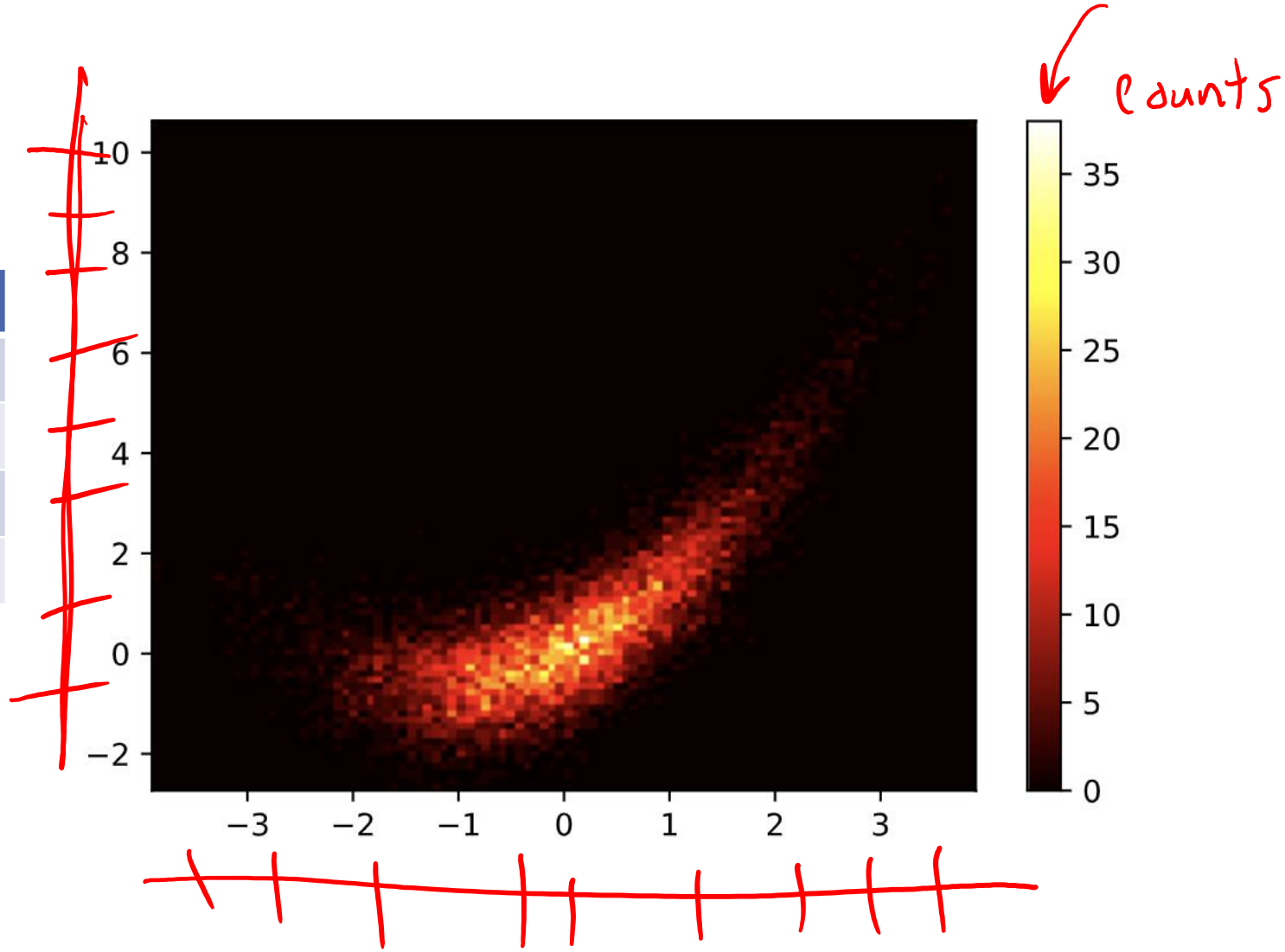
	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓

Why not ordinal data in first dimension?



Heatmap (density, or 2D histogram)

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓



Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

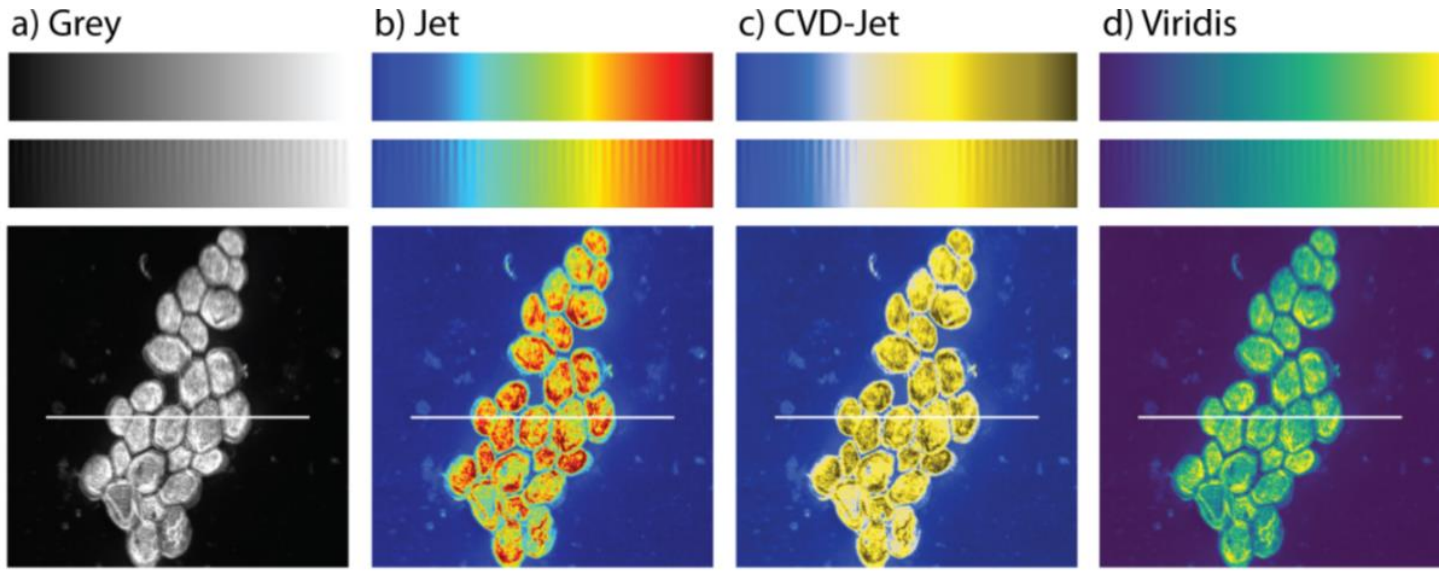


Image: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199239

Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

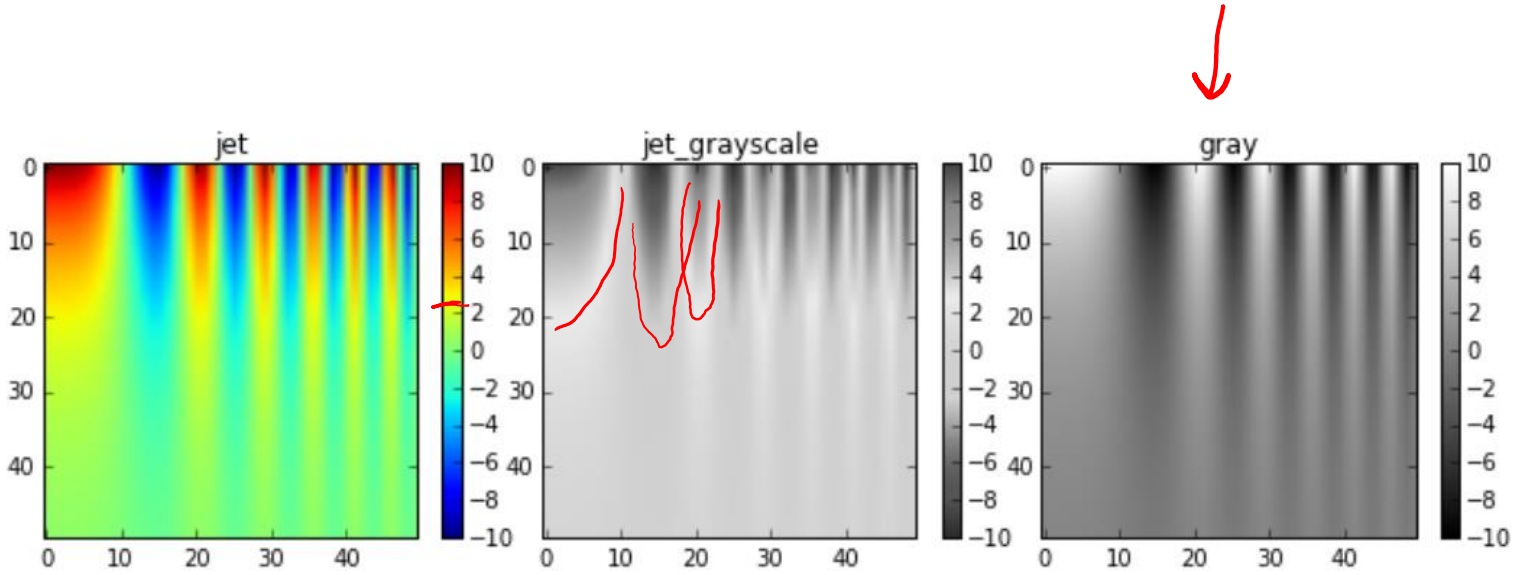


Image: https://jakevdp.github.io/blog/2014/10/16/how-bad-is-your-colormap/

Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

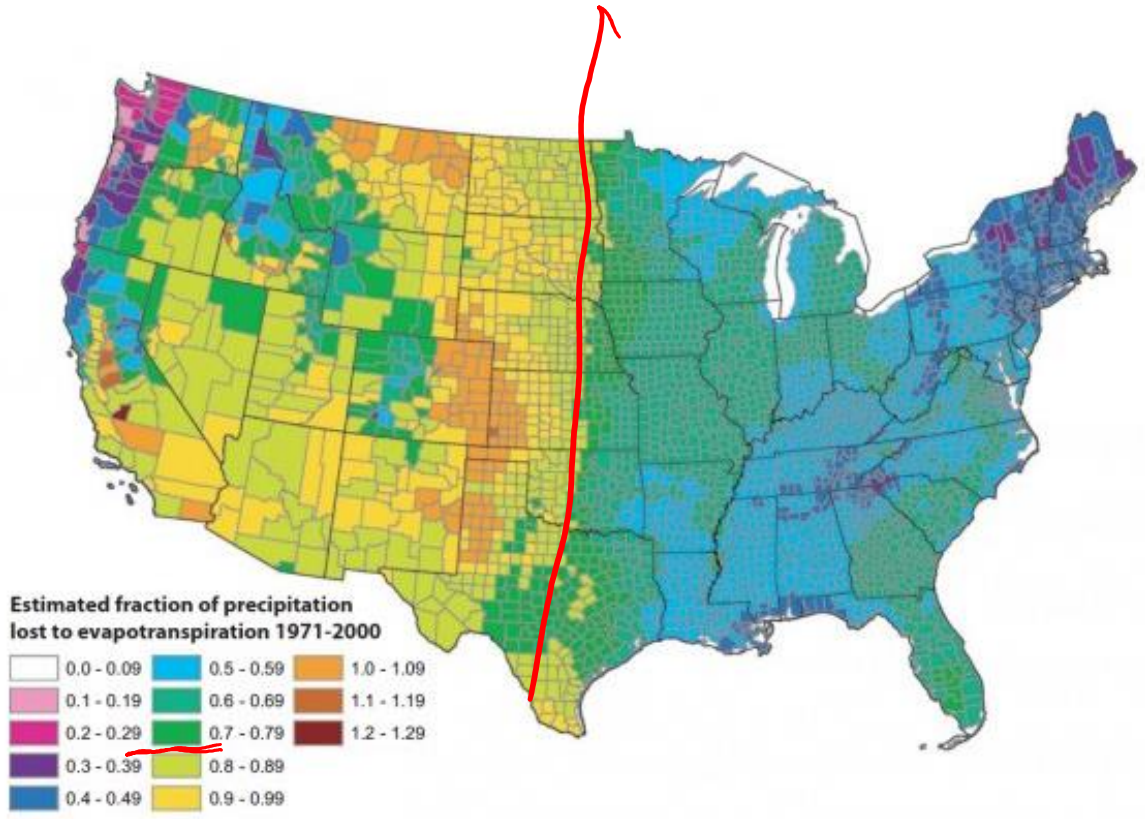


Image: https://eagereyes.org/basics/rainbow-color-map

Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

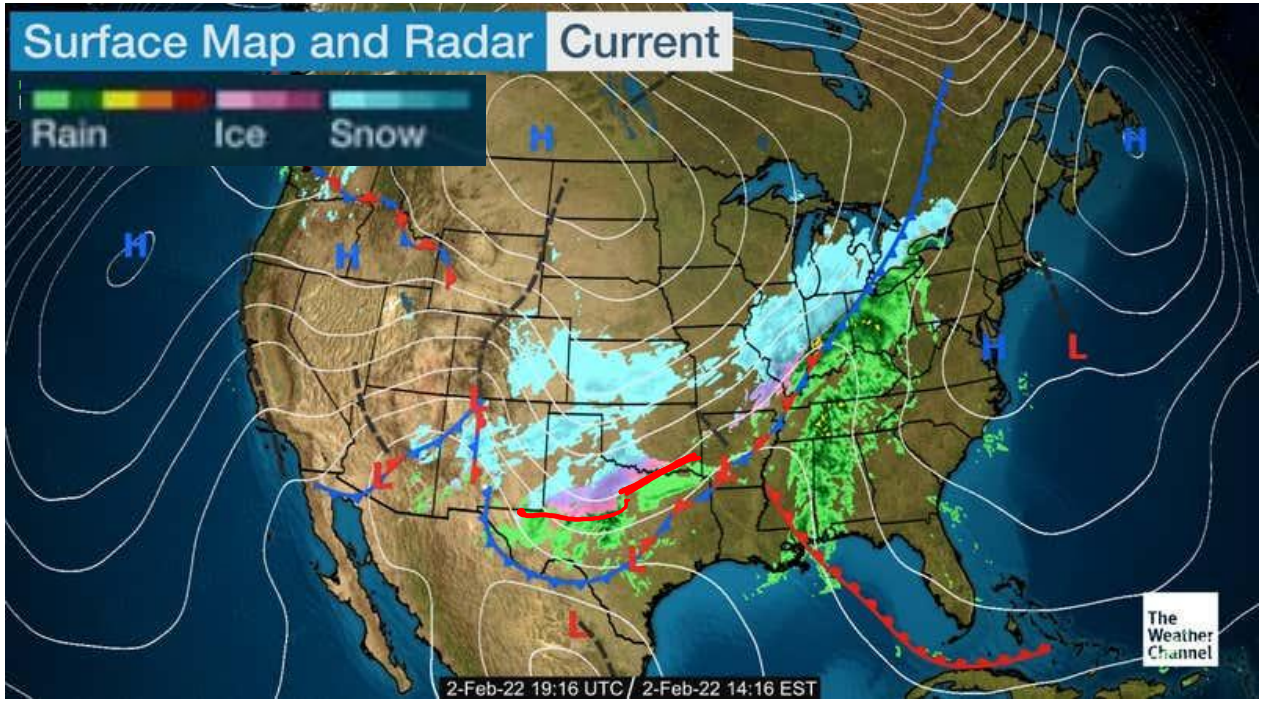


Image: https://weather.com/maps/currentusweather

Choose Colormaps Carefully

Several factors to consider. For example:

- Accessibility
- Printing in grayscale
- Unintentional boundaries
- Intentional boundaries
- Color semantics

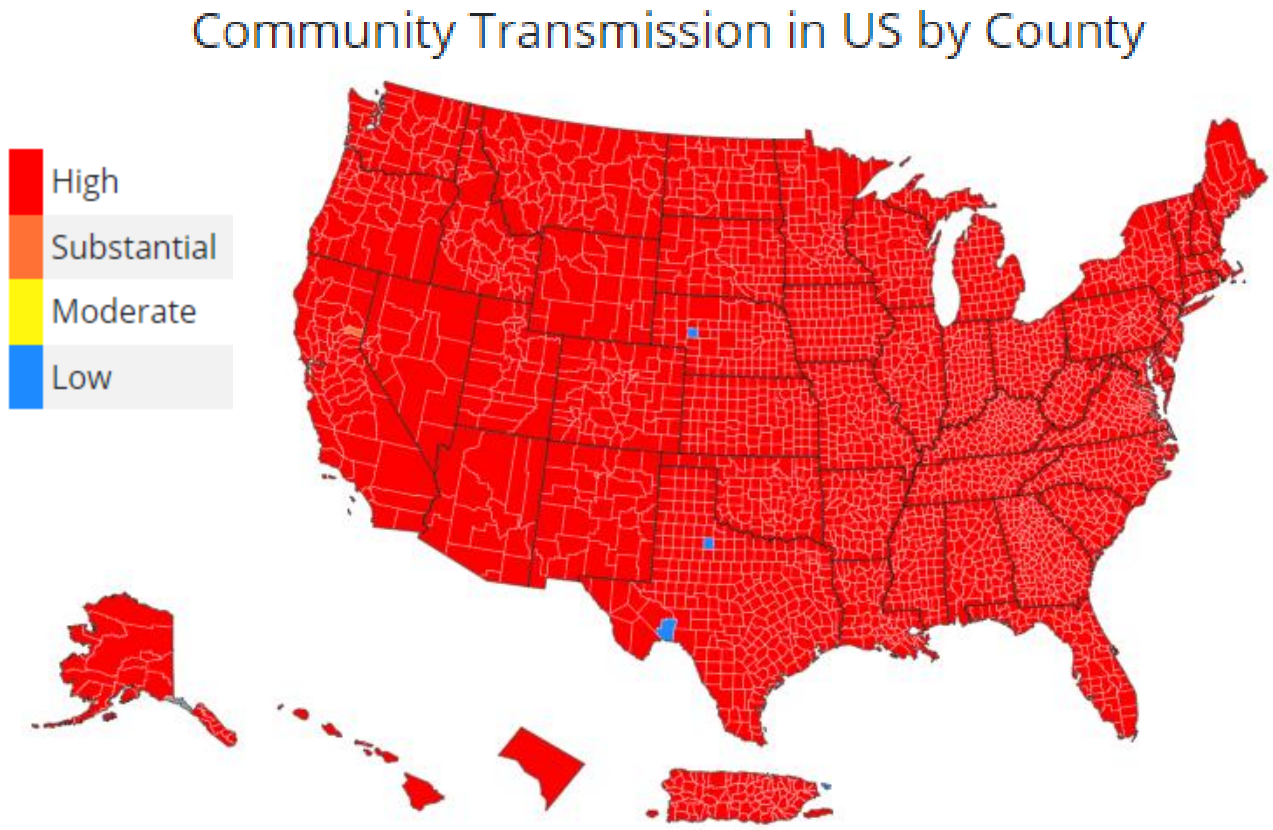
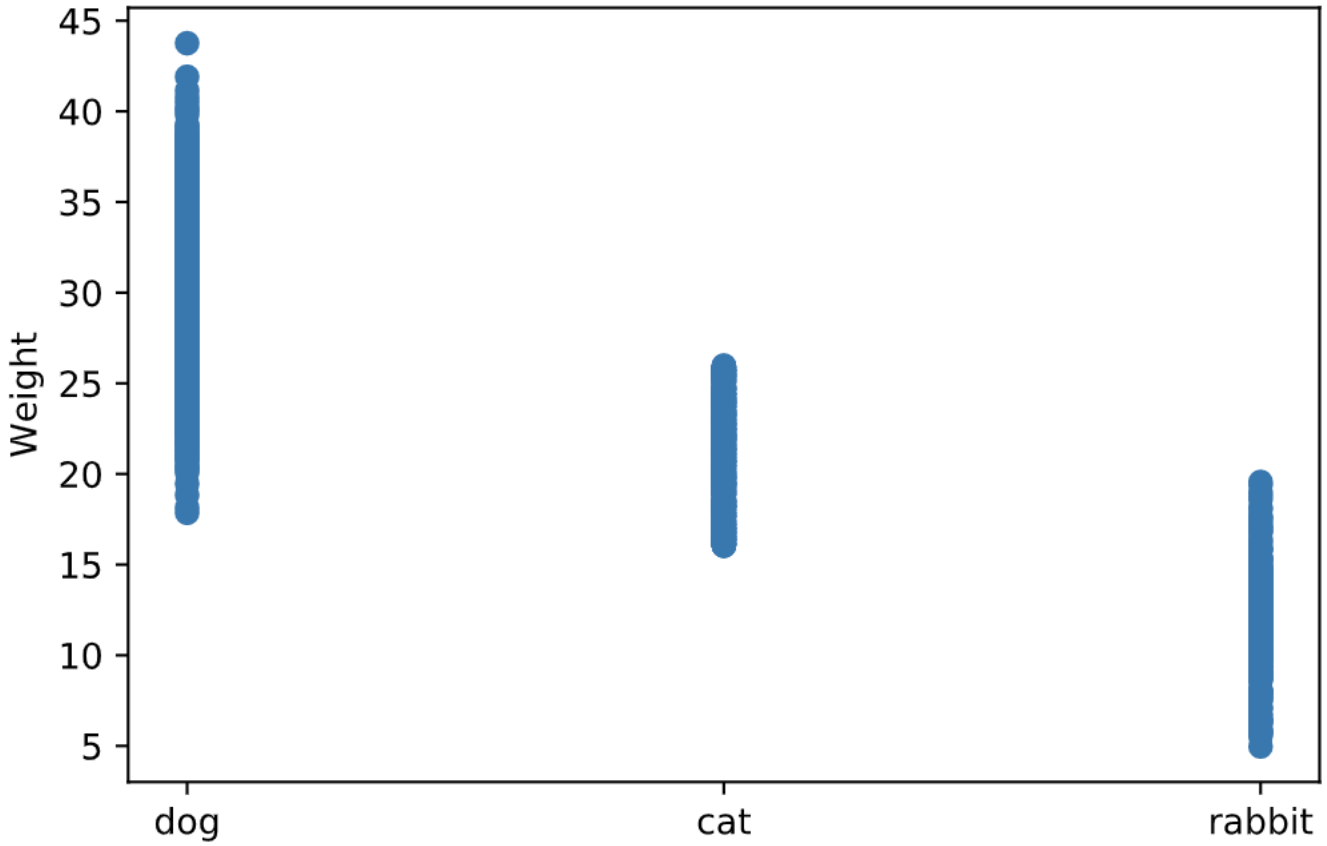


Image: https://covid.cdc.gov/covid-data-tracker/#county-view

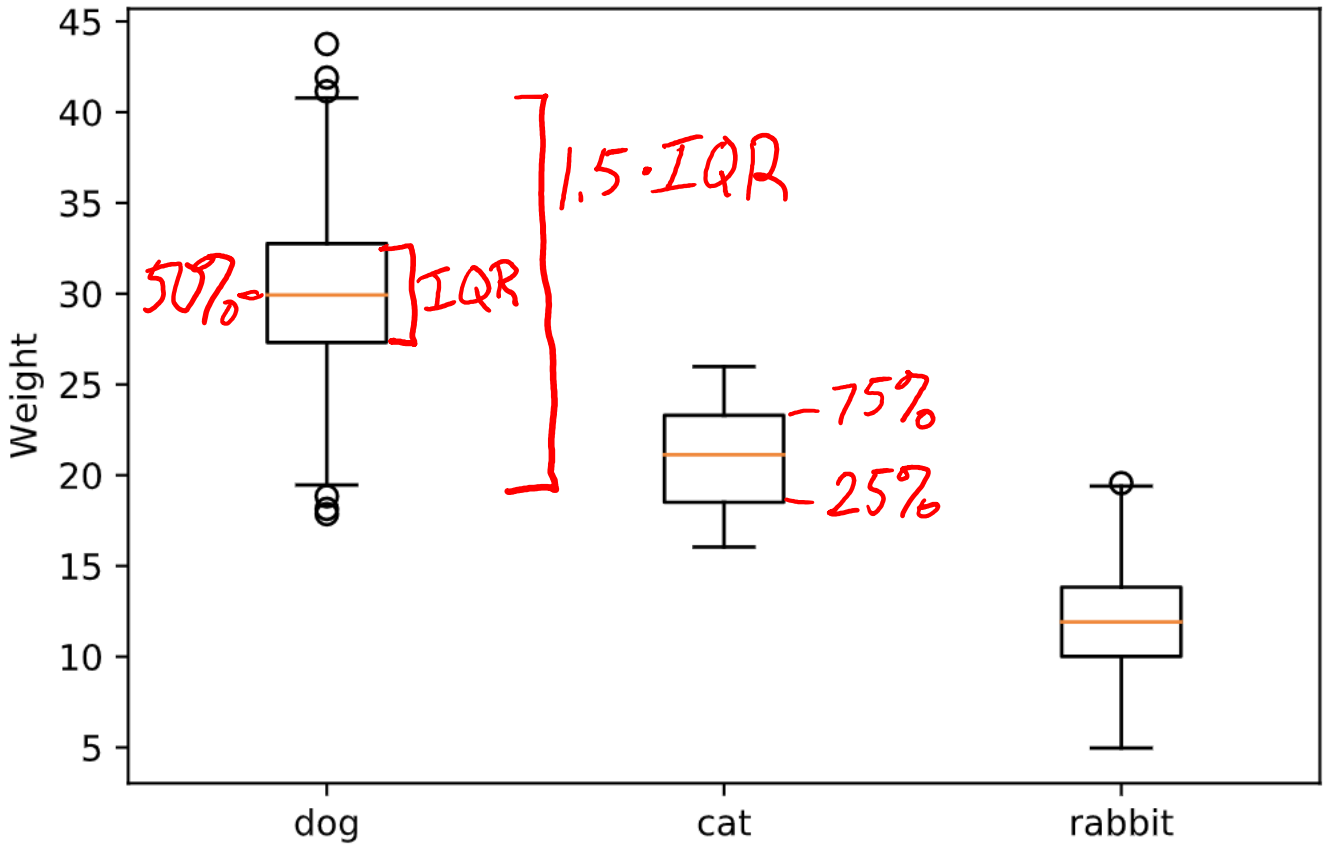
Scatter plot (bad)

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓



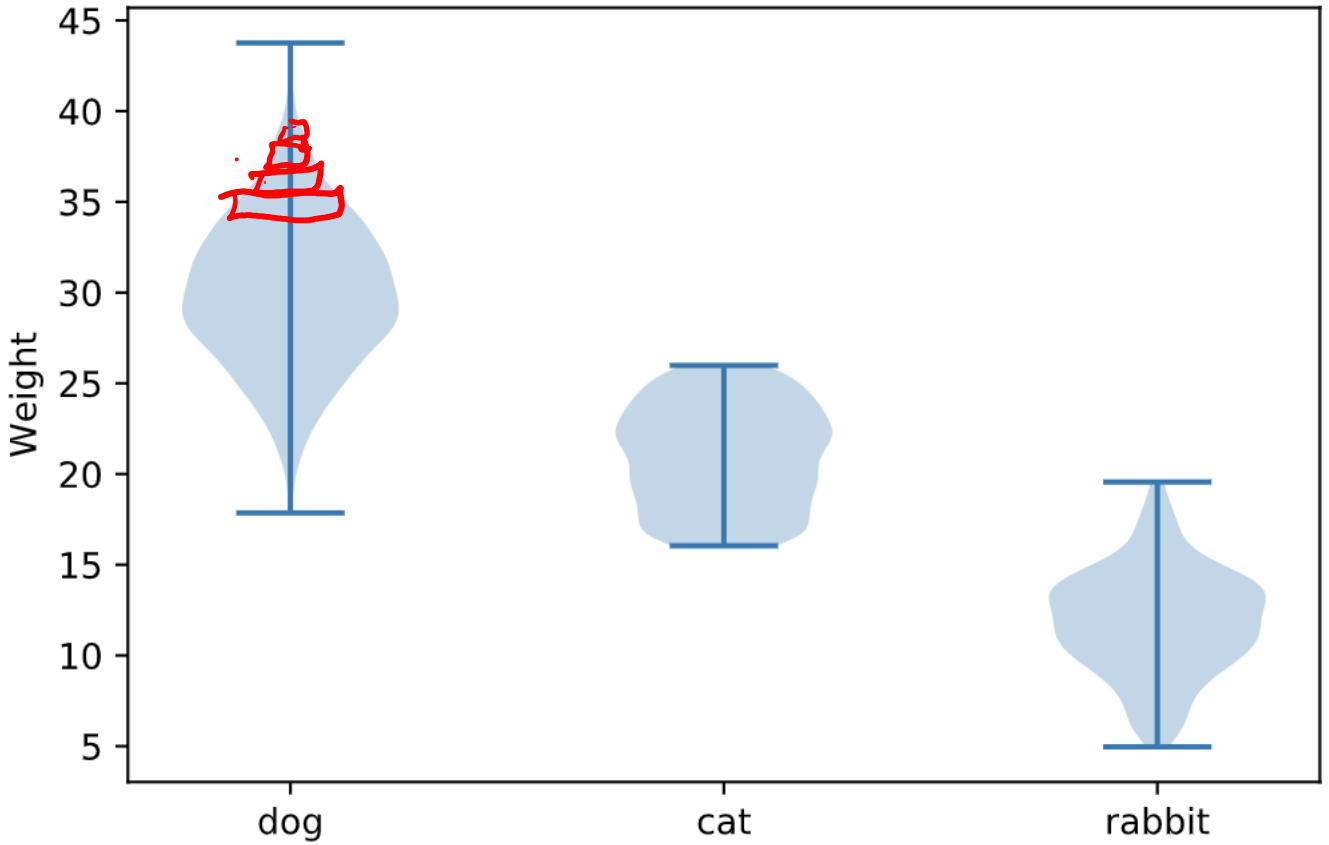
Box and whiskers

	Dim 1	Dim 2
Nominal	✔	✗
Ordinal	✔	✗
Interval	✗	✔
Ratio	✗	✔



Violin plot

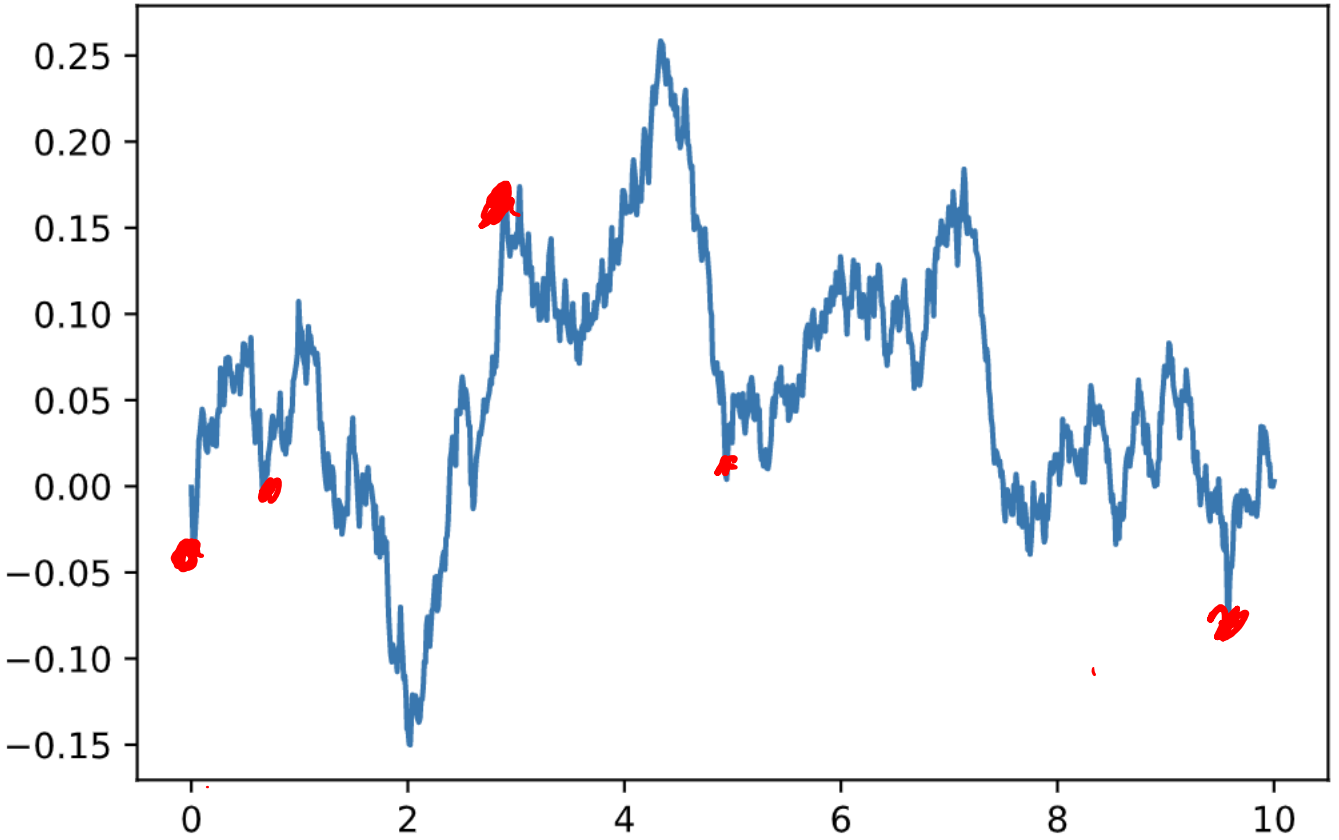
	Dim 1	Dim 2
Nominal	✔	✗
Ordinal	✔	✗
Interval	✗	✔
Ratio	✗	✔



Line plot

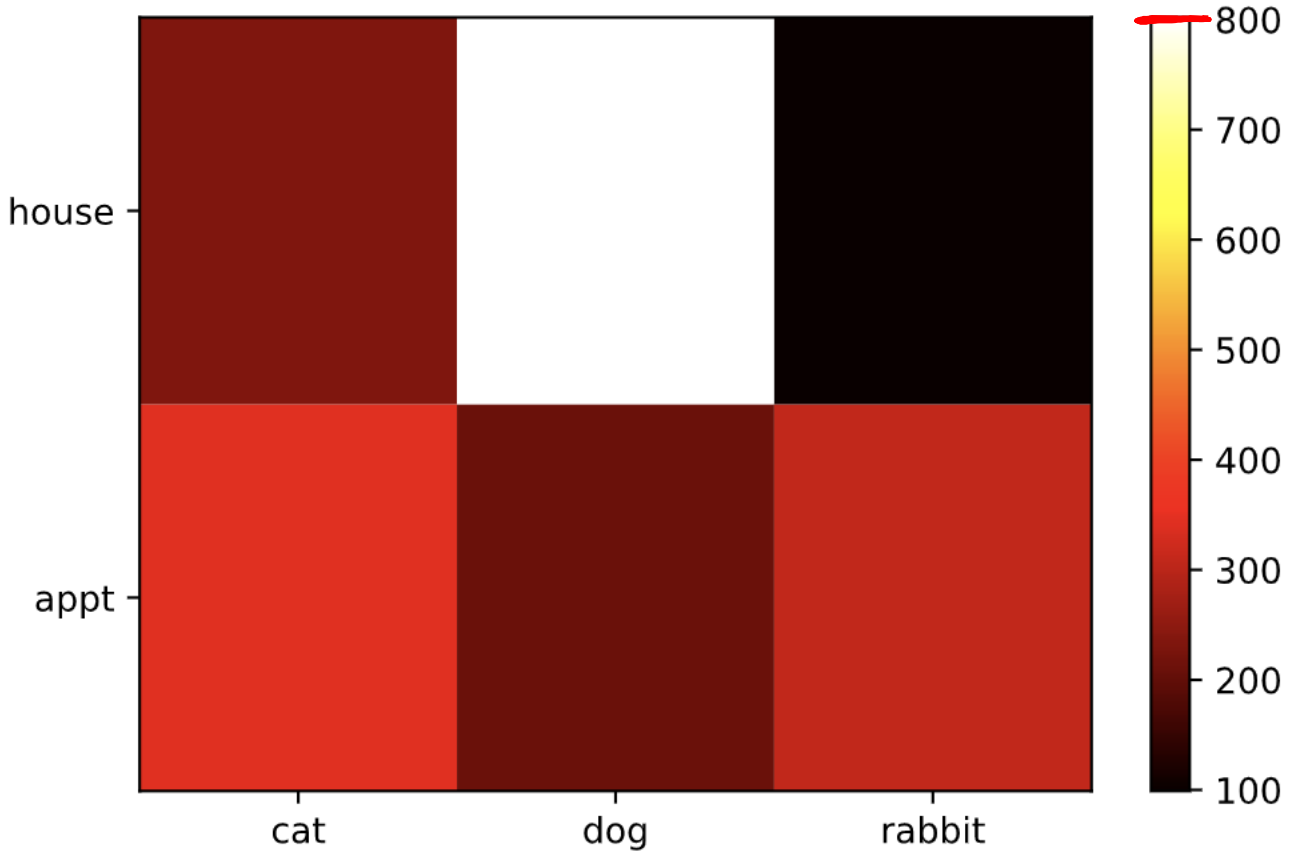
	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓

Why not ordinal data in first dimension?



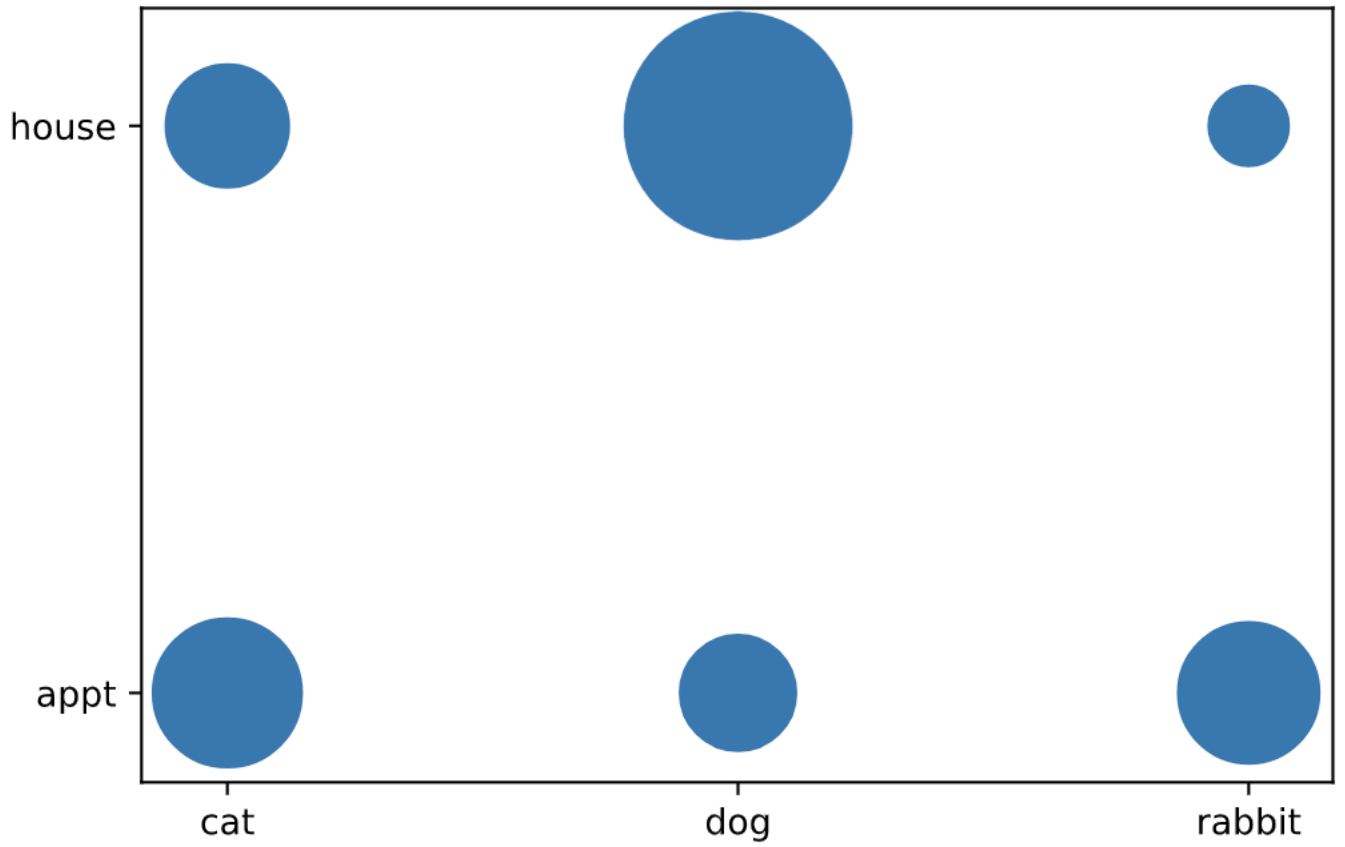
Heatmap (matrix)

	Dim 1	Dim 2
Nominal	✔	✔
Ordinal	✔	✔
Interval	✗	✗
Ratio	✗	✗



Bubble plot

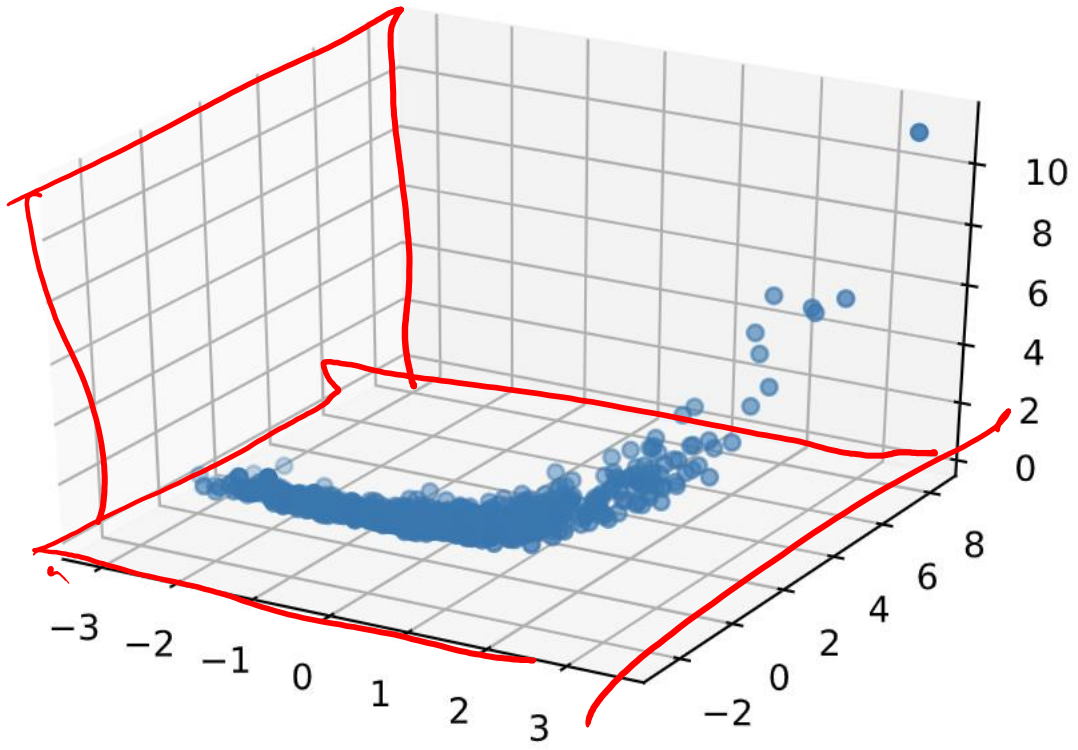
	Dim 1	Dim 2
Nominal	✔	✔
Ordinal	✔	✔
Interval	✗	✗
Ratio	✗	✗



3D+ DATA

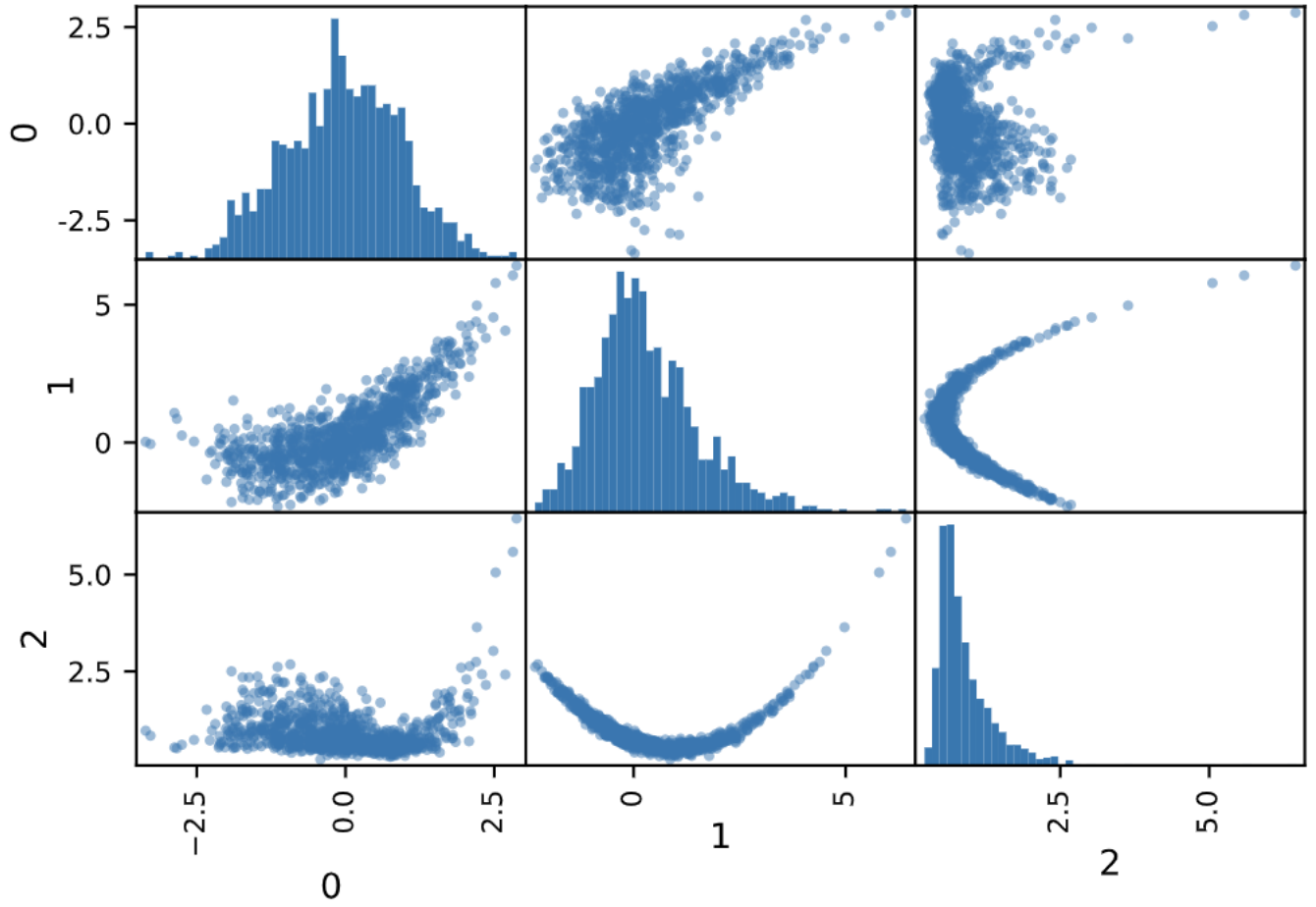
3D scatter plot

	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✗
Ordinal	✗	✗	✗
Interval	✗	✗	✗
Ratio	✗	✗	✗



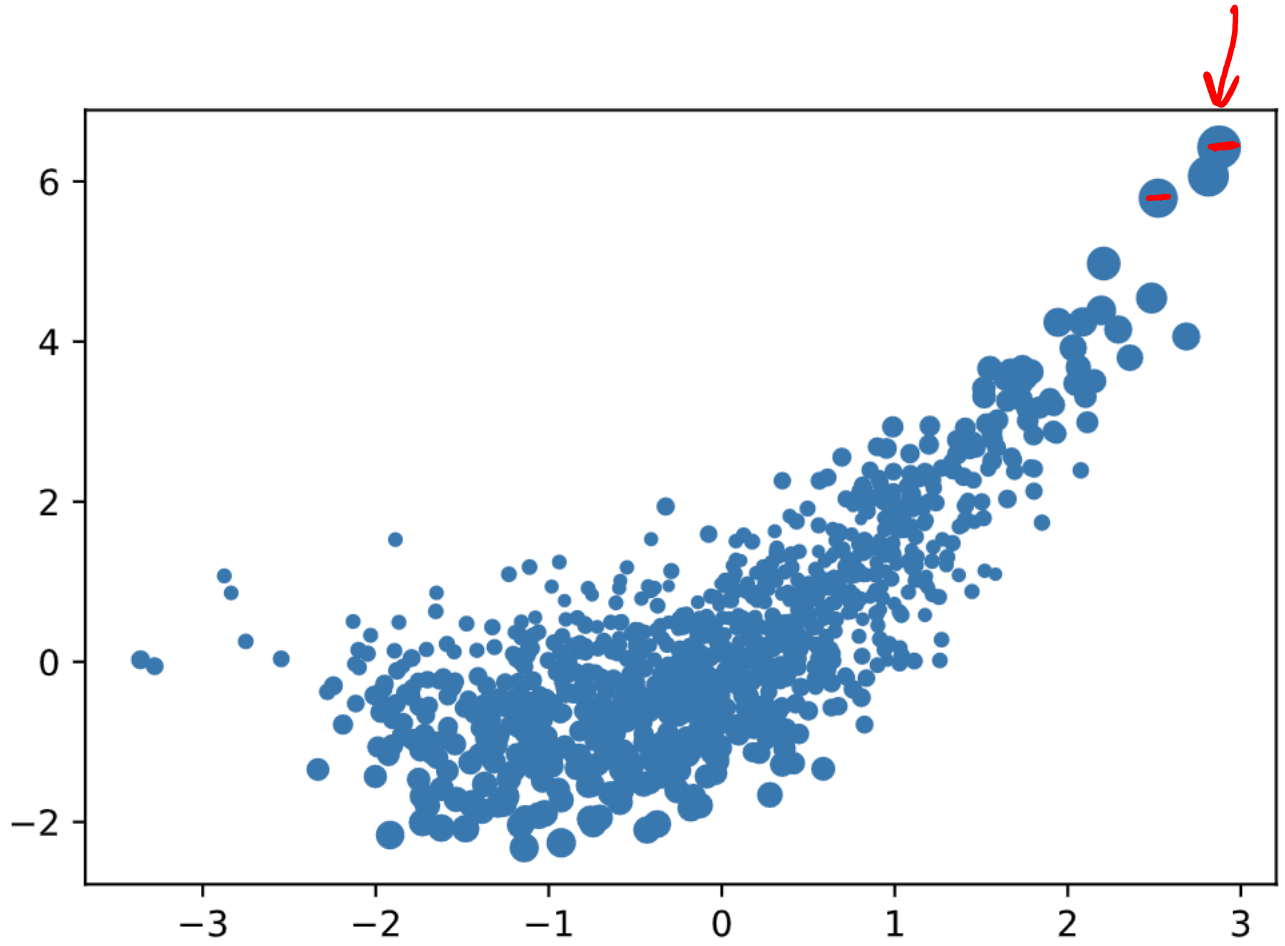
Scatter plot matrix

	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✗
Ordinal	✗	✗	✗
Interval	✓	✓	✓
Ratio	✓	✓	✓



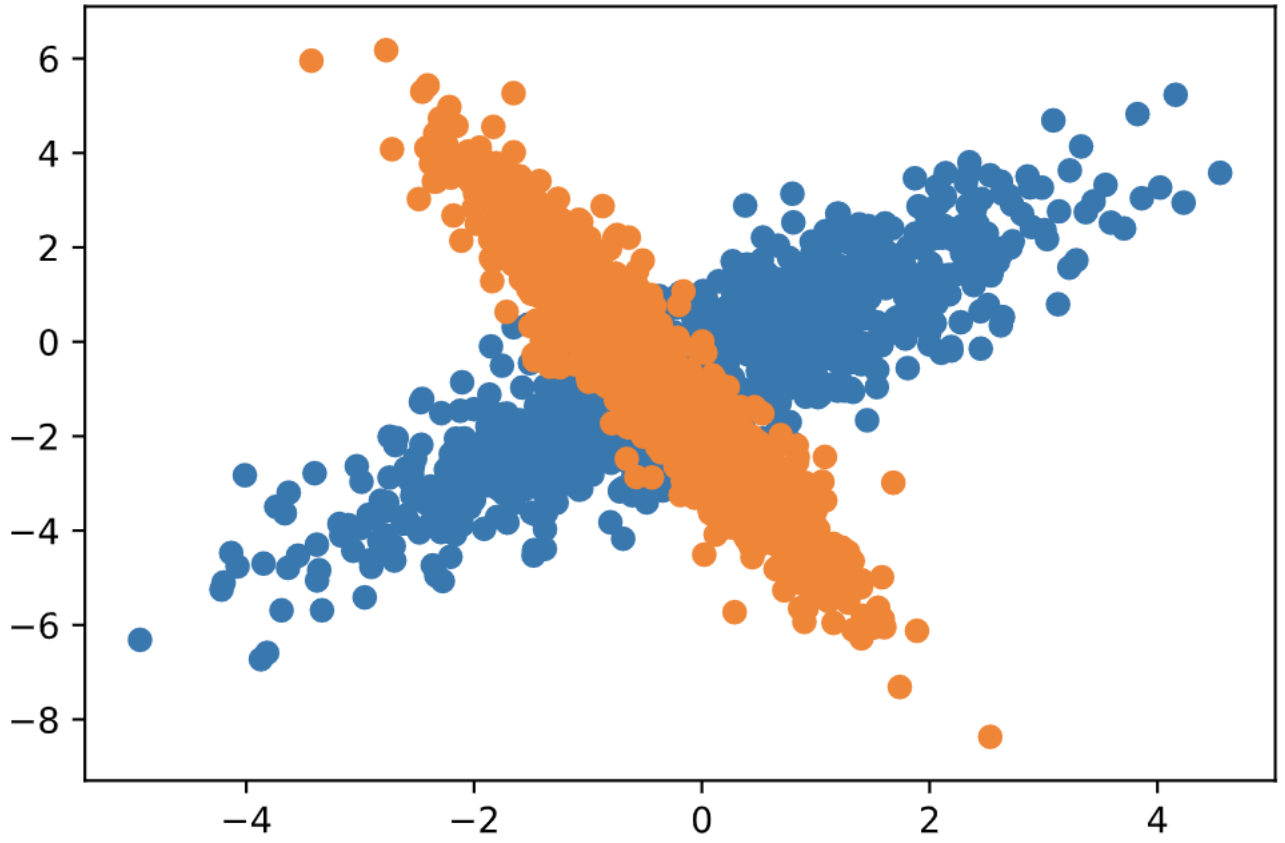
Bubble plot

	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✗
Ordinal	✗	✗	✗
Interval	✓	✓	✓
Ratio	✓	✓	✓



Color scatter plot

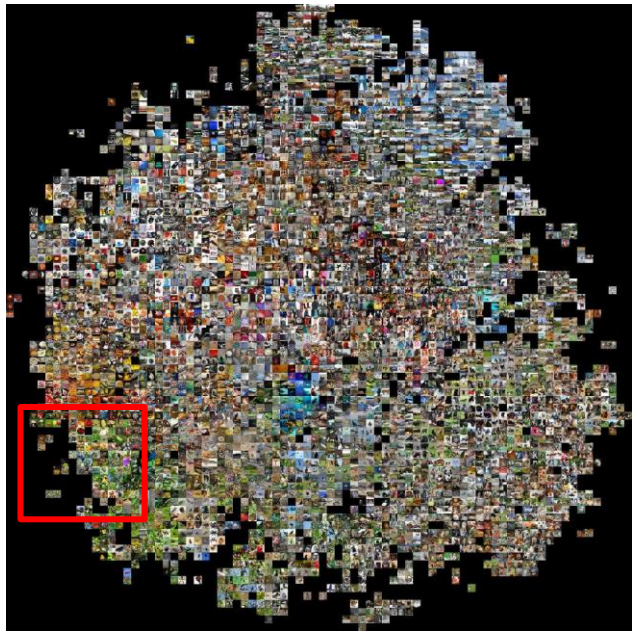
	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✓
Ordinal	✗	✗	✓
Interval	✓	✓	✗
Ratio	✓	✓	✗



t-SNE (be careful!)

t-SNE (T-distributed Stochastic Neighbor Embedding) works to embed high dimensional data in smaller (often 2 or 3) dimensions. Van der Maaten, L., & Hinton, G. (2008).

Easy to misread : https://distill.pub/2016/misread-tsne/



Images: https://cs.stanford.edu/people/karpathy/cnnembed/, 4069-dimensional vectors from CNN embedding

Outline

Basics of visualization

Data types and visualization types

Software plotting libraries

Matplotlib

Matplotlib is the standard for plotting in Python / Jupyter Notebook

Matplotlib used to generate fairly ugly plots by default, but in recent versions this is no longer the case, so minimal need for additional libraries

It is aimed at generating static plots, not very good for interacting with data (with a few exceptions)

A number of additional libraries provide some level of interactive plot (and static plots), but matplotlib is enough of a standard that we'll use it here

Courses