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

J. Zico Kolter
Carnegie Mellon University

Fall 2019

Annoucements

HW1 due tomorrow

HW2 released tomorrow, **due 10/1**

Pinned thread on Diderot for common questions on HW1

Very firm on deadlines for HW (submit well before midnight, any additional time will count as a late day)

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

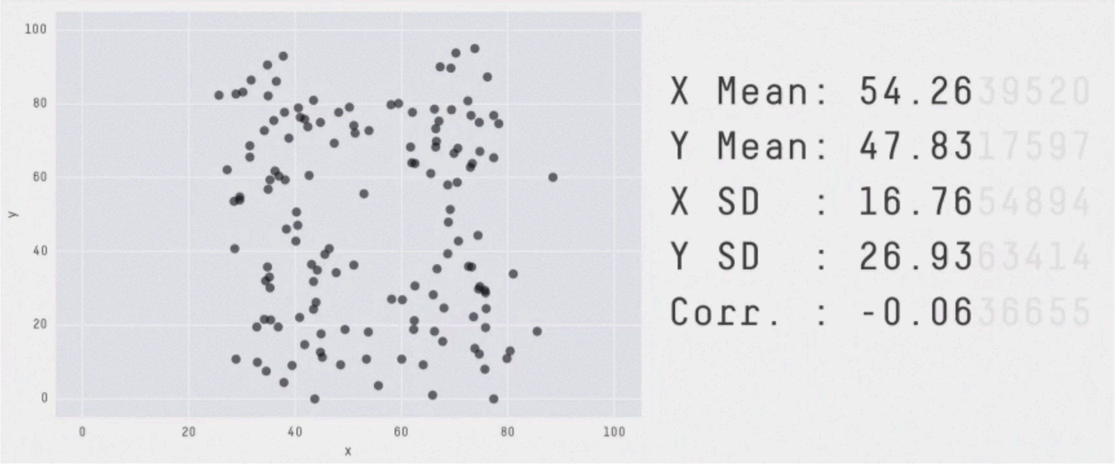
Before you run any analysis, build any machine learning system, etc, always visualize your data

If you can’t identify a trend or make a prediction for your dataset, neither will an automated algorithm

> 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)



```
X Mean: 54.2639520
Y Mean: 47.8317597
X SD  : 16.7654894
Y SD  : 26.9363414
Corr. : -0.0636655
```

[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 has no fixed meaning

Example – Temperature Fahrenheit

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

Ratio: numerical data, zero has special meaning

Example – Temperature Kelvin

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

Poll: Nominal and ordinal values

Which of the following questions that may be asked on a survey would be considered *ordinal*? (unchecked ones are *nominal*)

1. Gender: {male, female, other, prefer not to disclose}

2. Yearly income: {<$18k, $18-40k, $40-75k, >$75k}

3. Reaction to question: {Strongly disagree, slightly disagree, neutral, slightly agree, strongly agree}

4. May we add you to our mailing list: {No, Yes}

Poll: Interval and ratio values

Which of the following quantities would be considered *ratio*? (unchecked values are *interval*)

1. Length (meters)

2. Length (feet)

3. Velocity (meters/second)

4. IQ Score

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

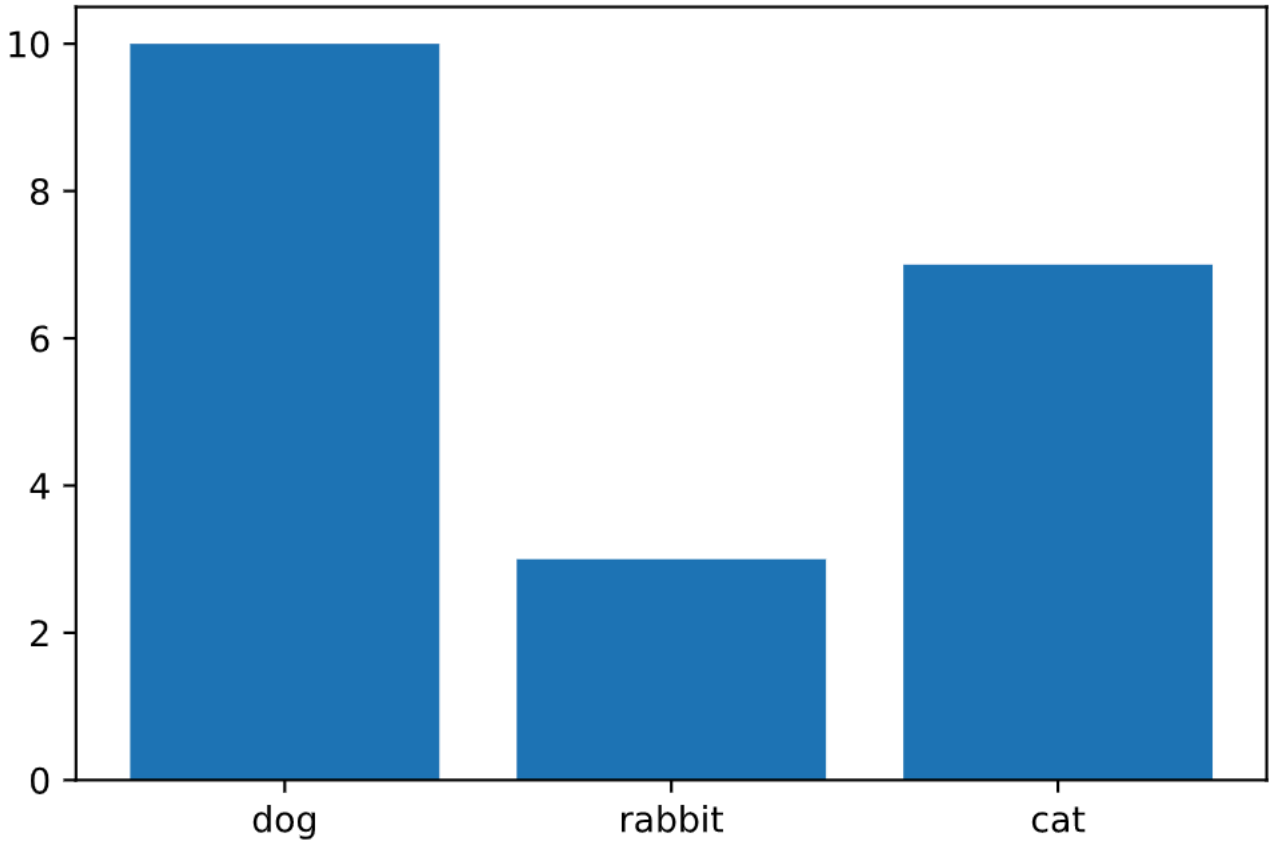
1D DATA

Bar chart

	Data
Nominal	
Ordinal	
Interval	✗
Ratio	✗

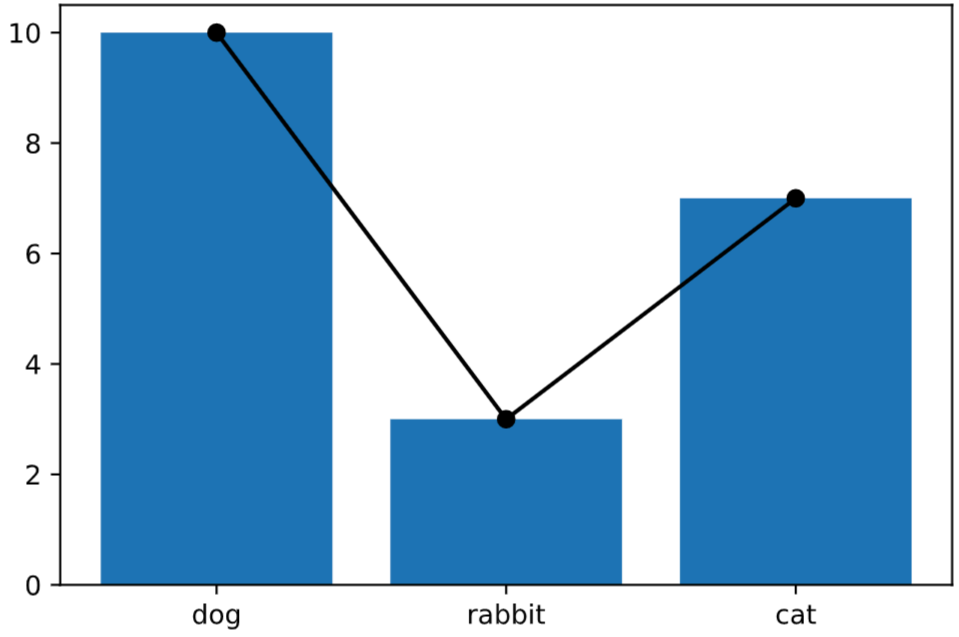
↑

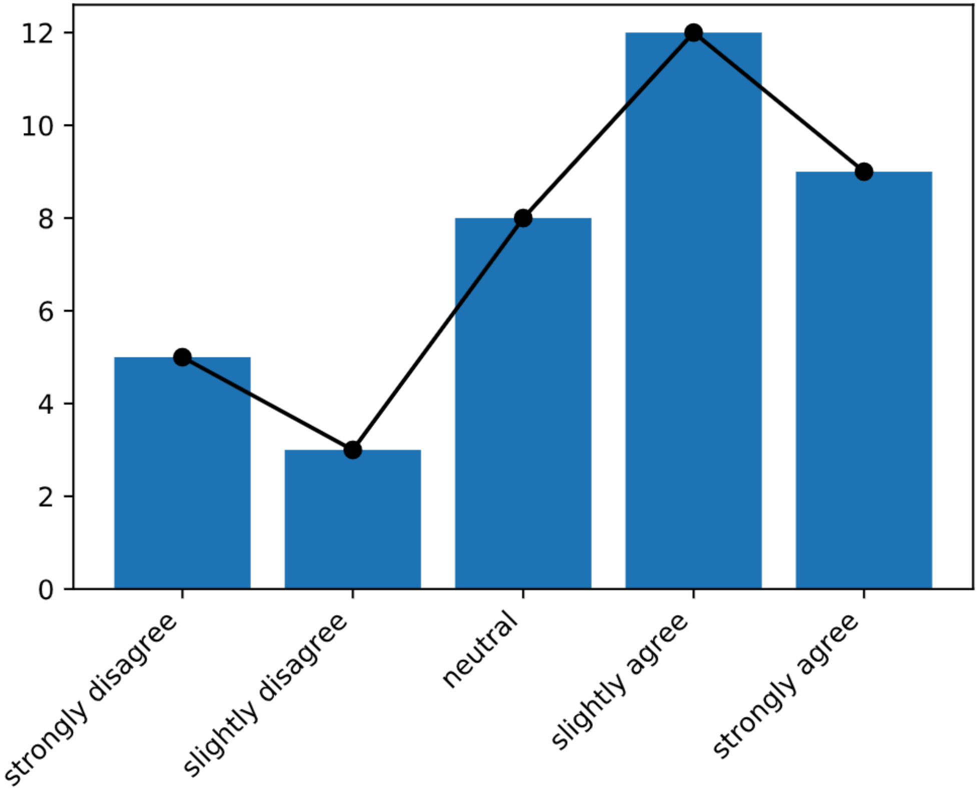
Suggestions, not rules



Bar chart (bad)

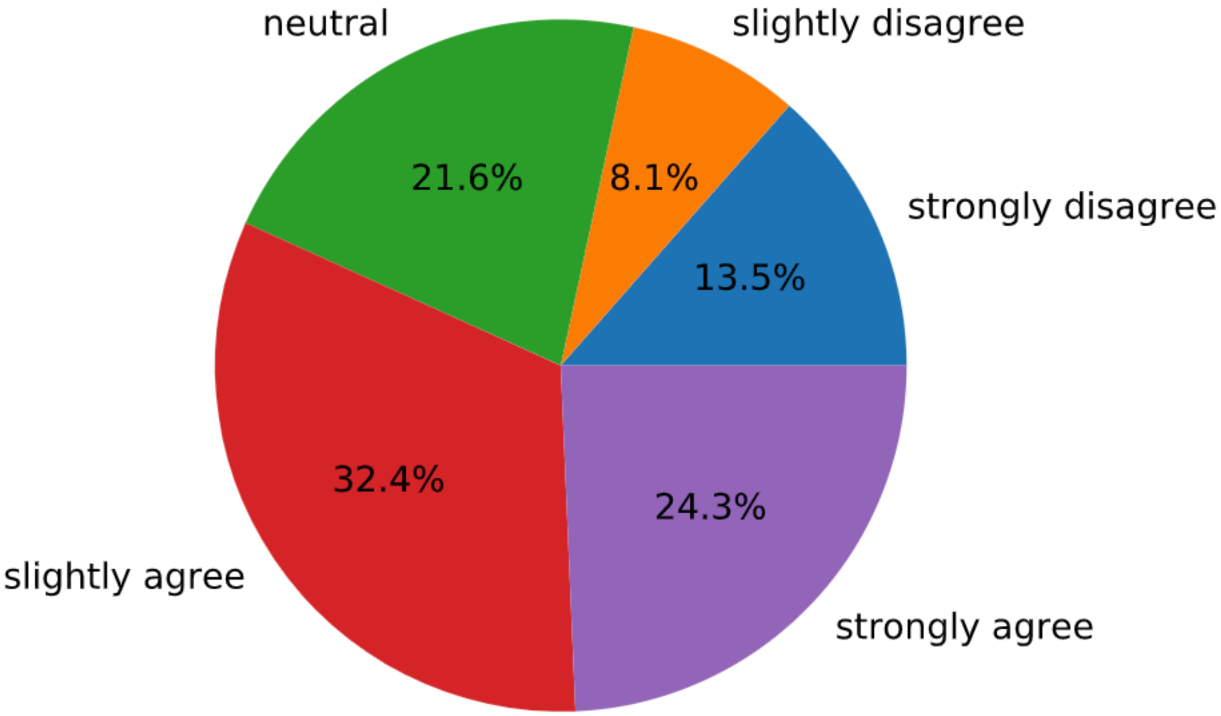
Don't use lines within a bar chart for categorial or ordinal features!





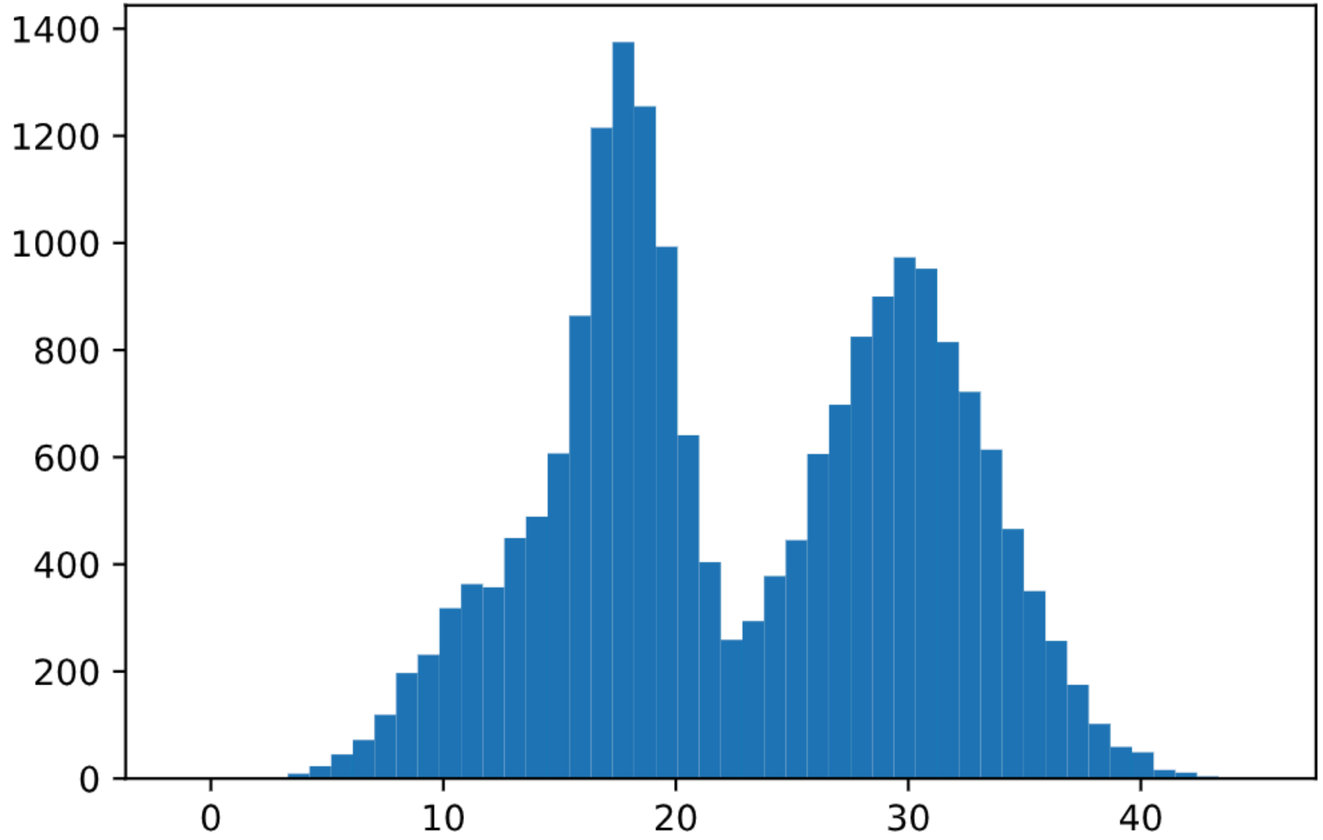
Pie chart

	Data
Nominal	✗
Ordinal	✗
Interval	✗
Ratio	✗



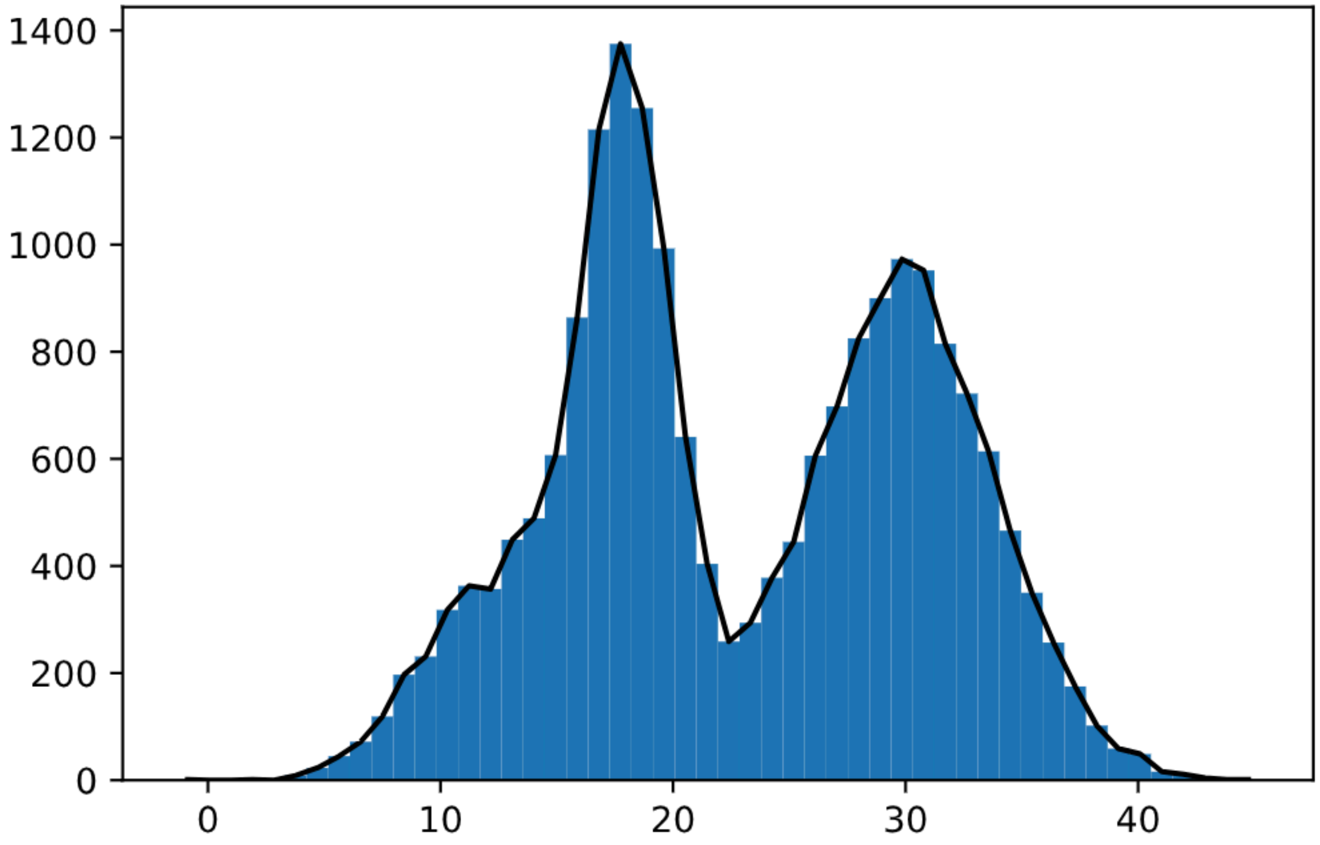
Histogram

	Data
Nominal	✗
Ordinal	✗
Interval	
Ratio	



Histogram

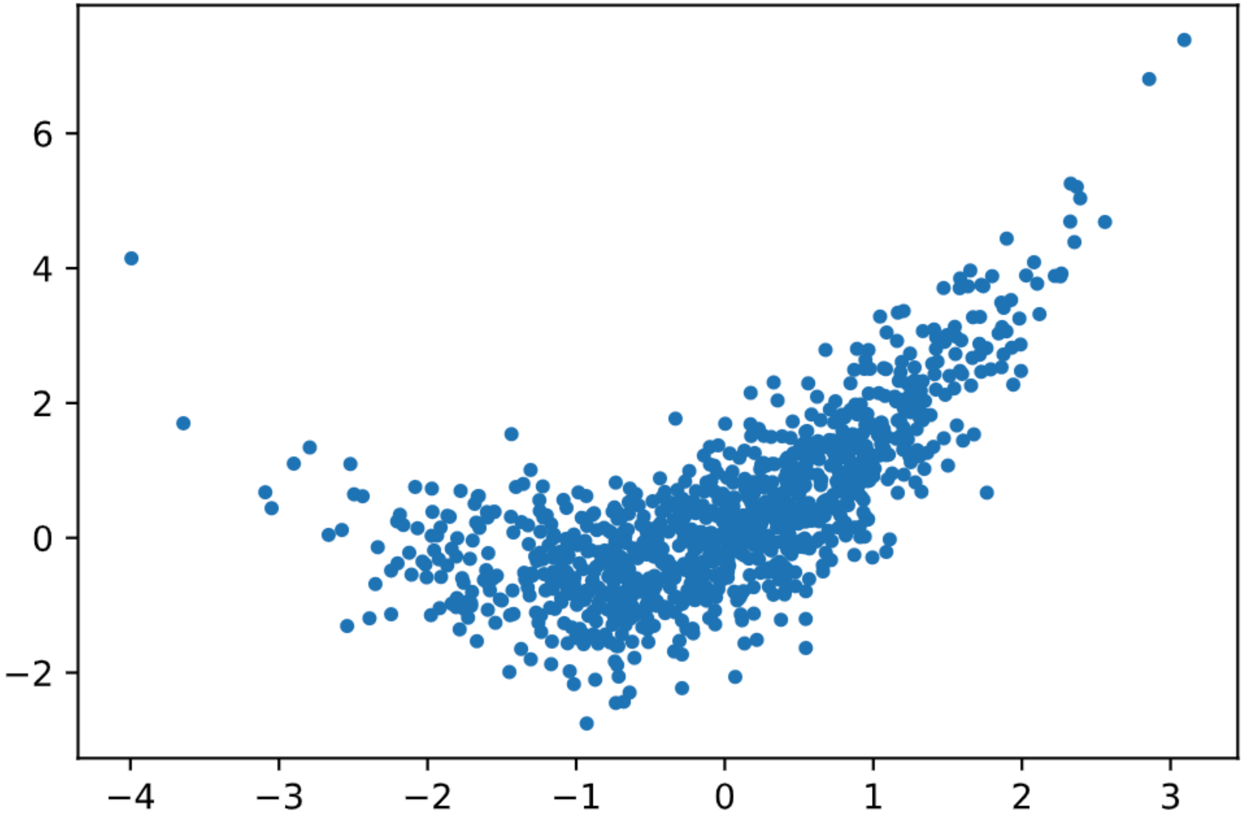
OK to use lines within a histogram (but not very informative)



2D DATA

Scatter plot

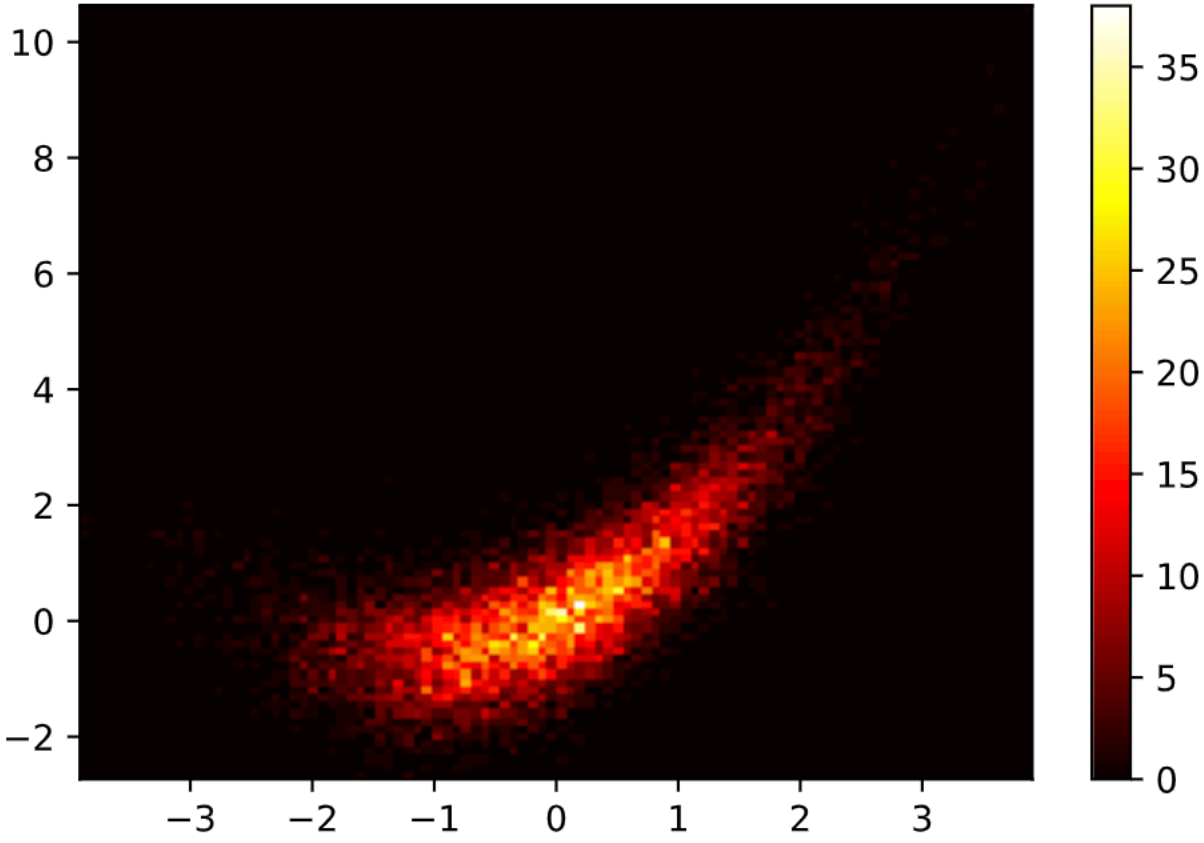
	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval		
Ratio		



Why not ordinal data in first dimension?

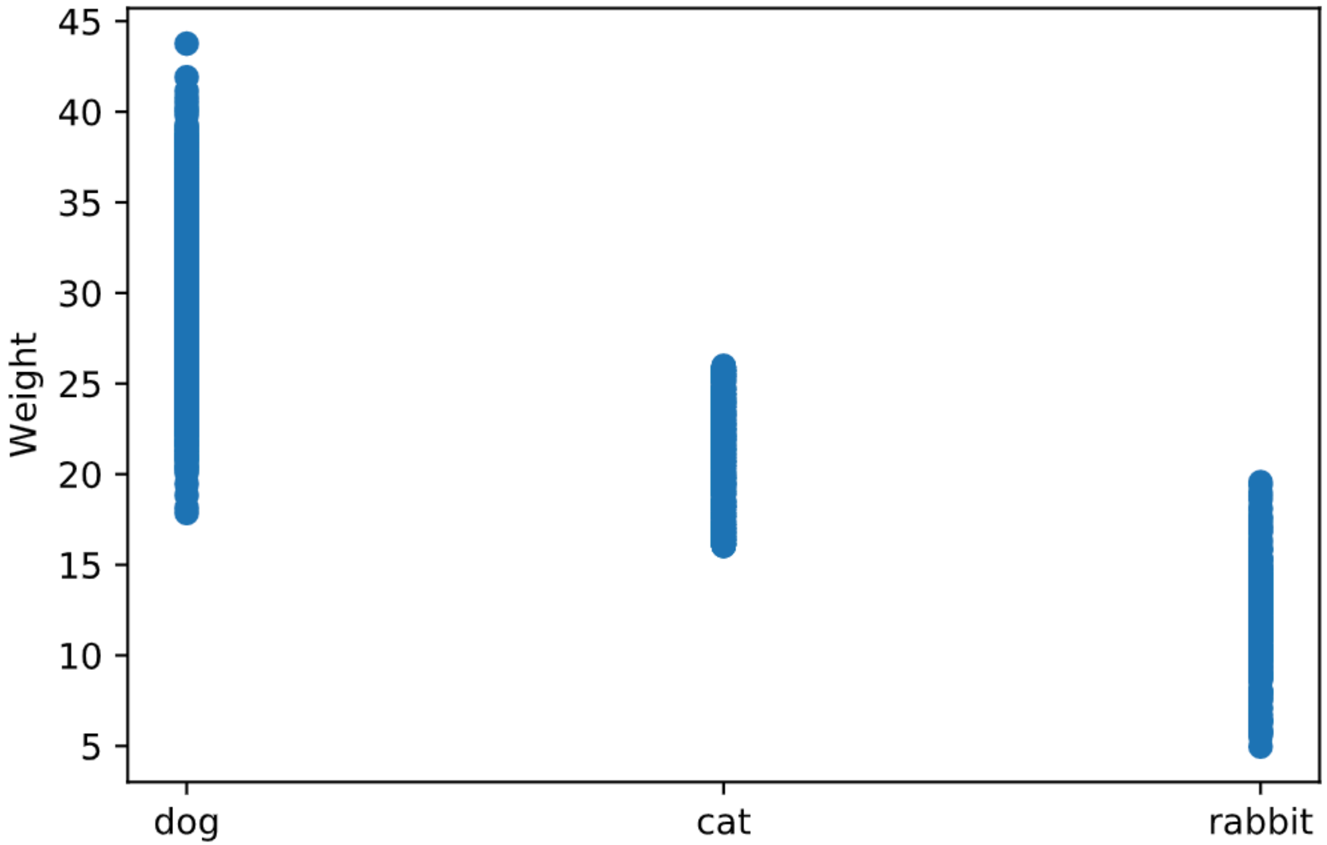
Heatmap (density, or 2D histogram)

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



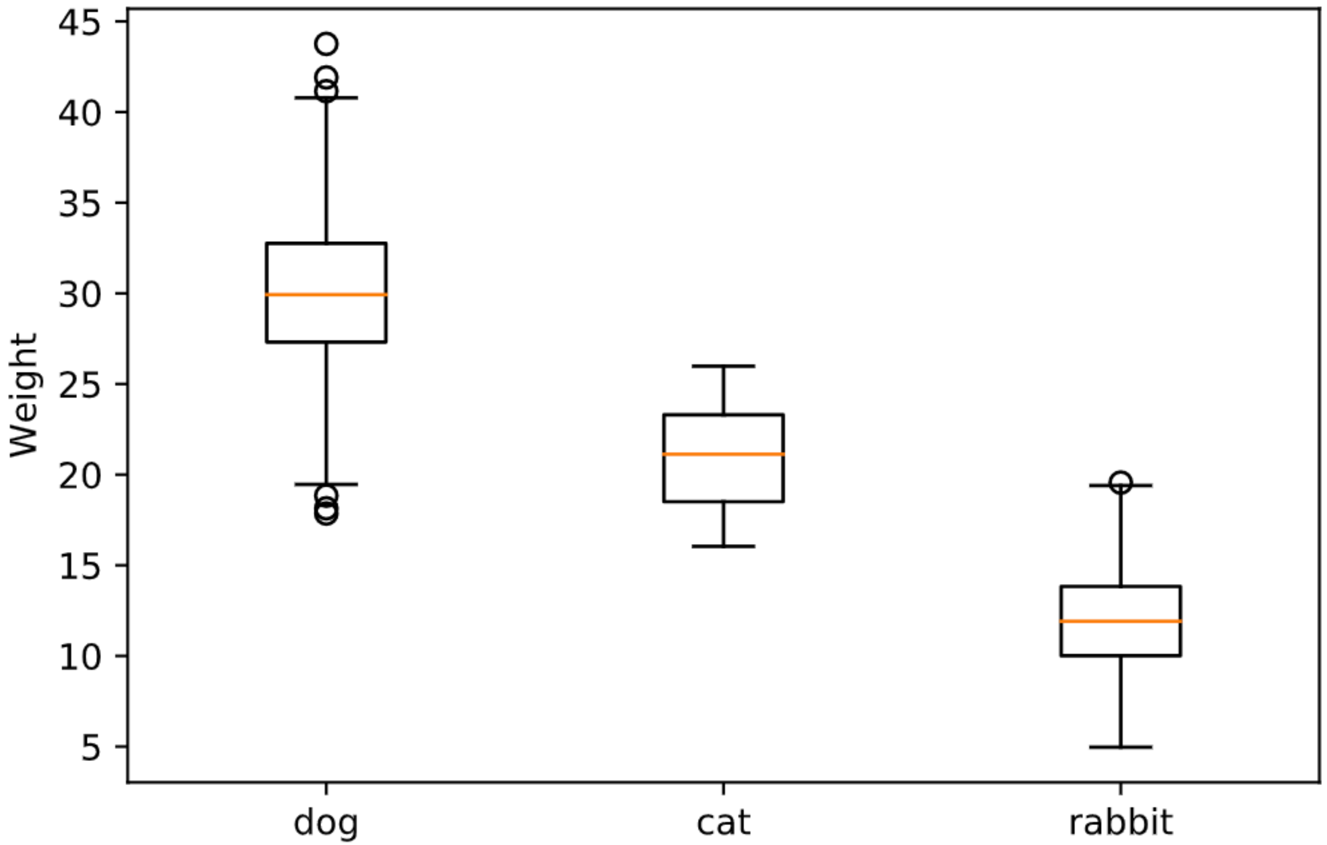
Scatter plot (bad)

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



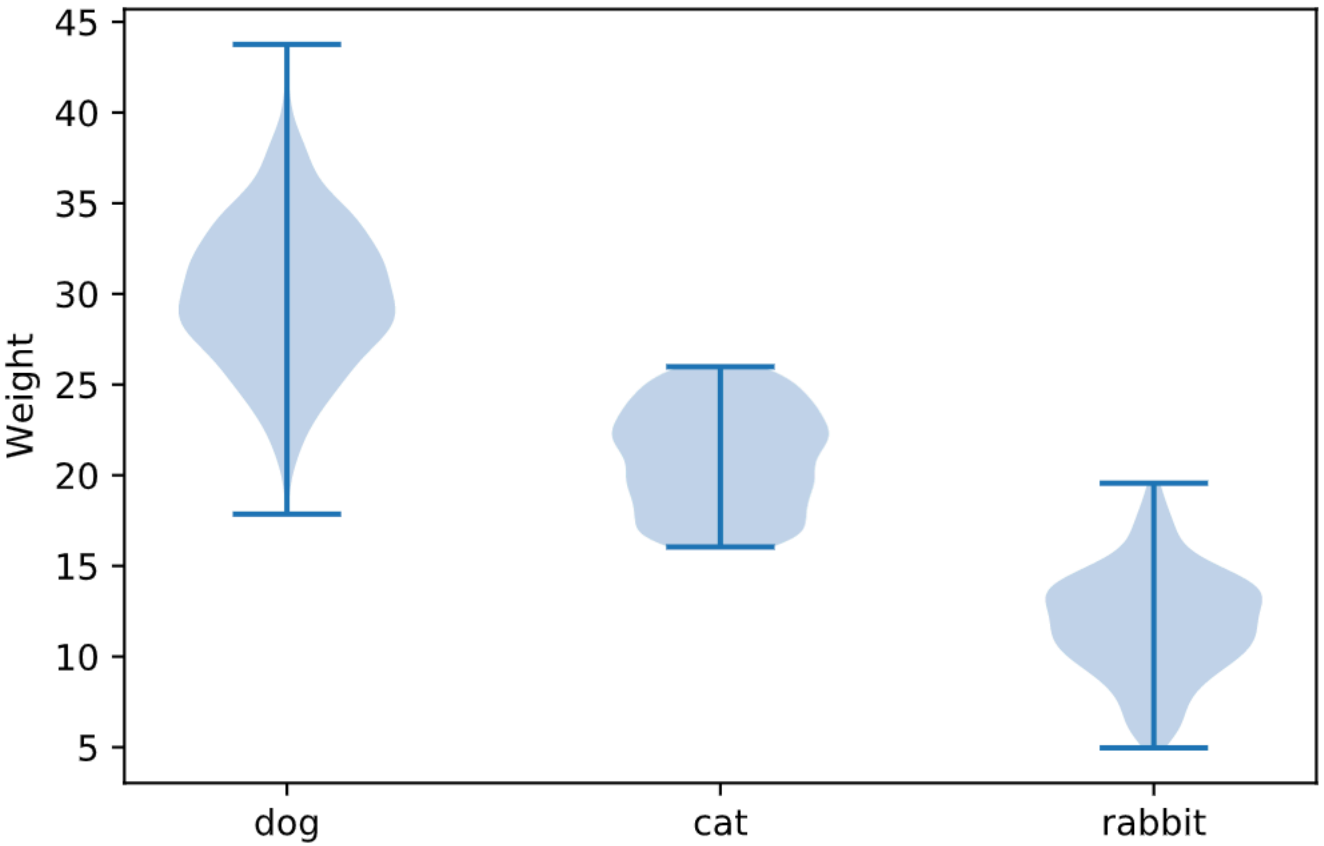
Box and whiskers

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



Violin plot

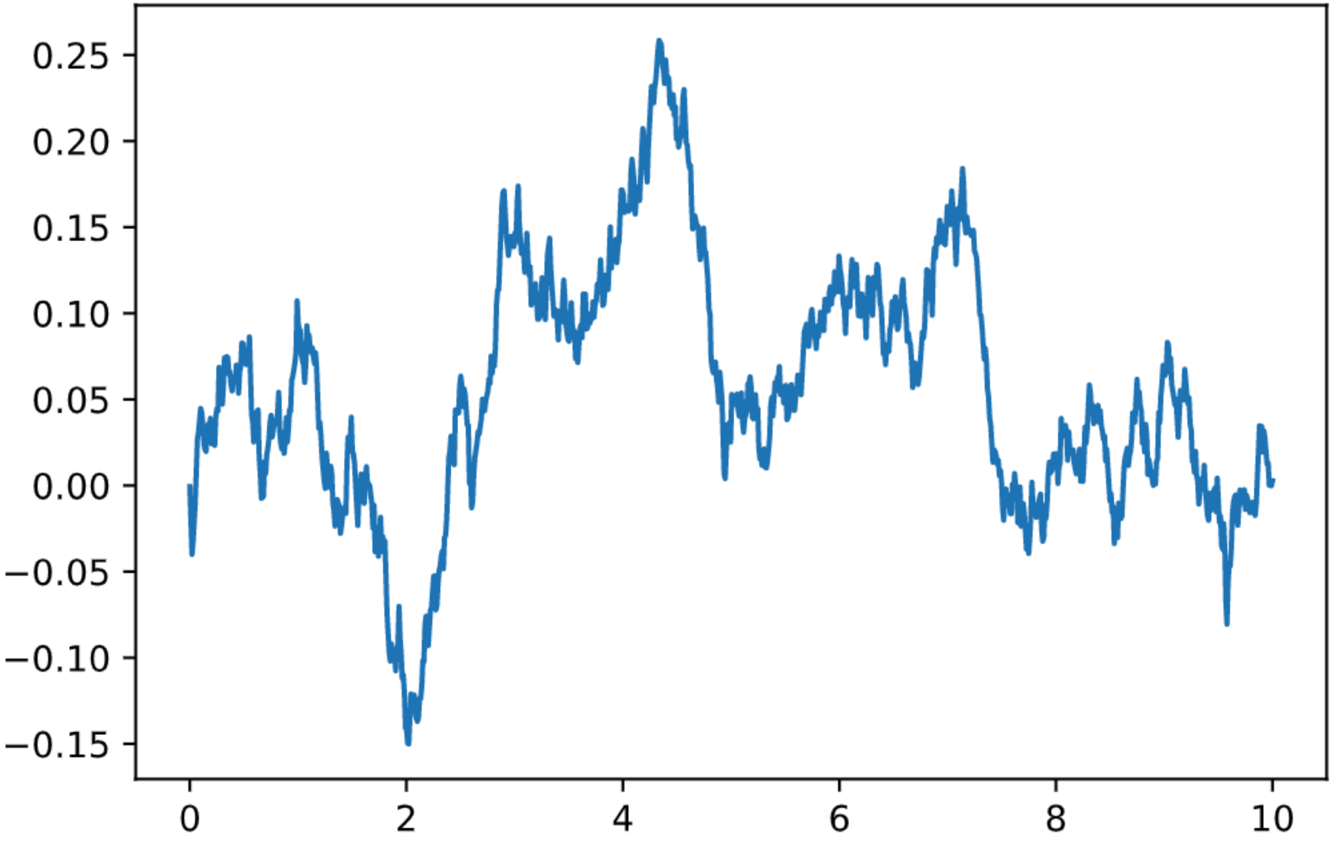
	Dim 1	Dim 2
Nominal		✗
Ordinal		✗
Interval	✗	
Ratio	✗	



Line plot

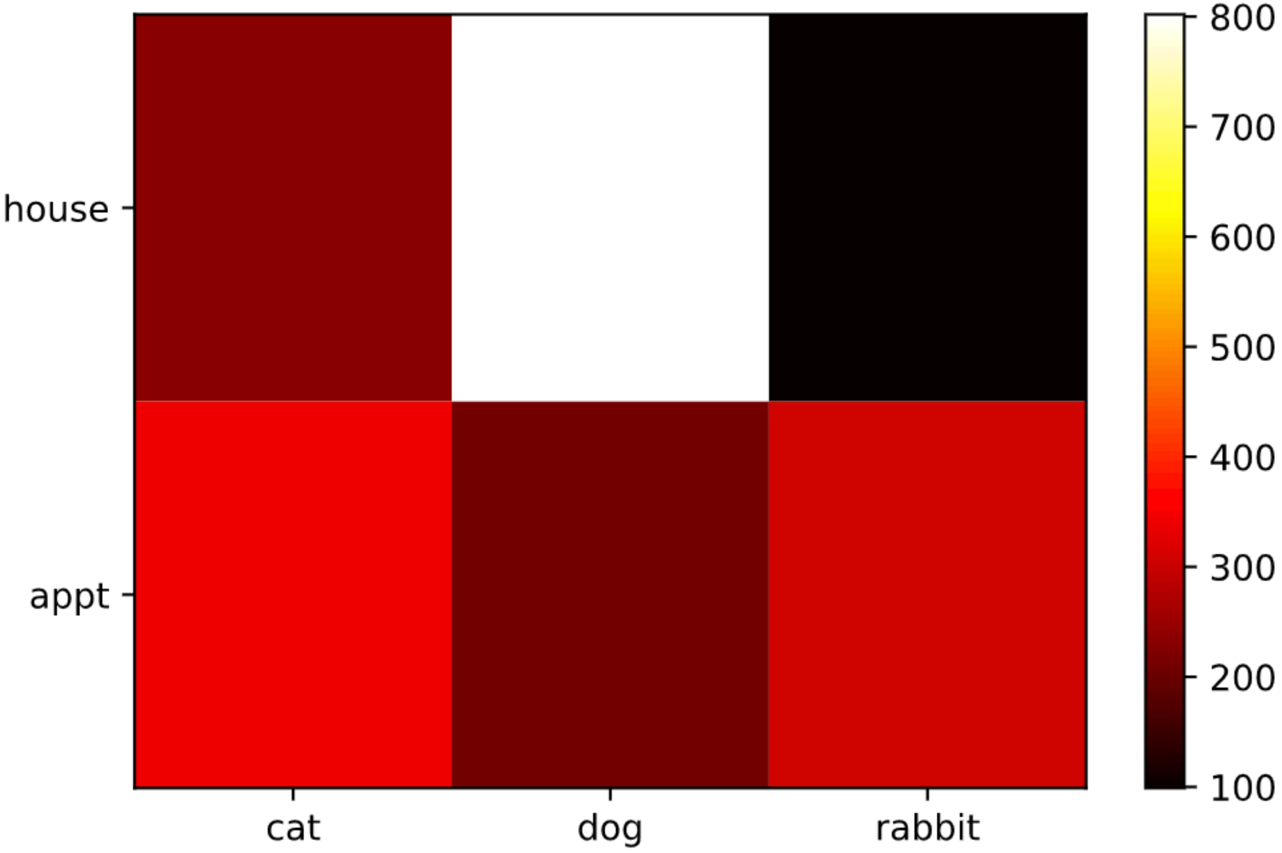
	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval		
Ratio		

Why not ordinal data in first dimension?



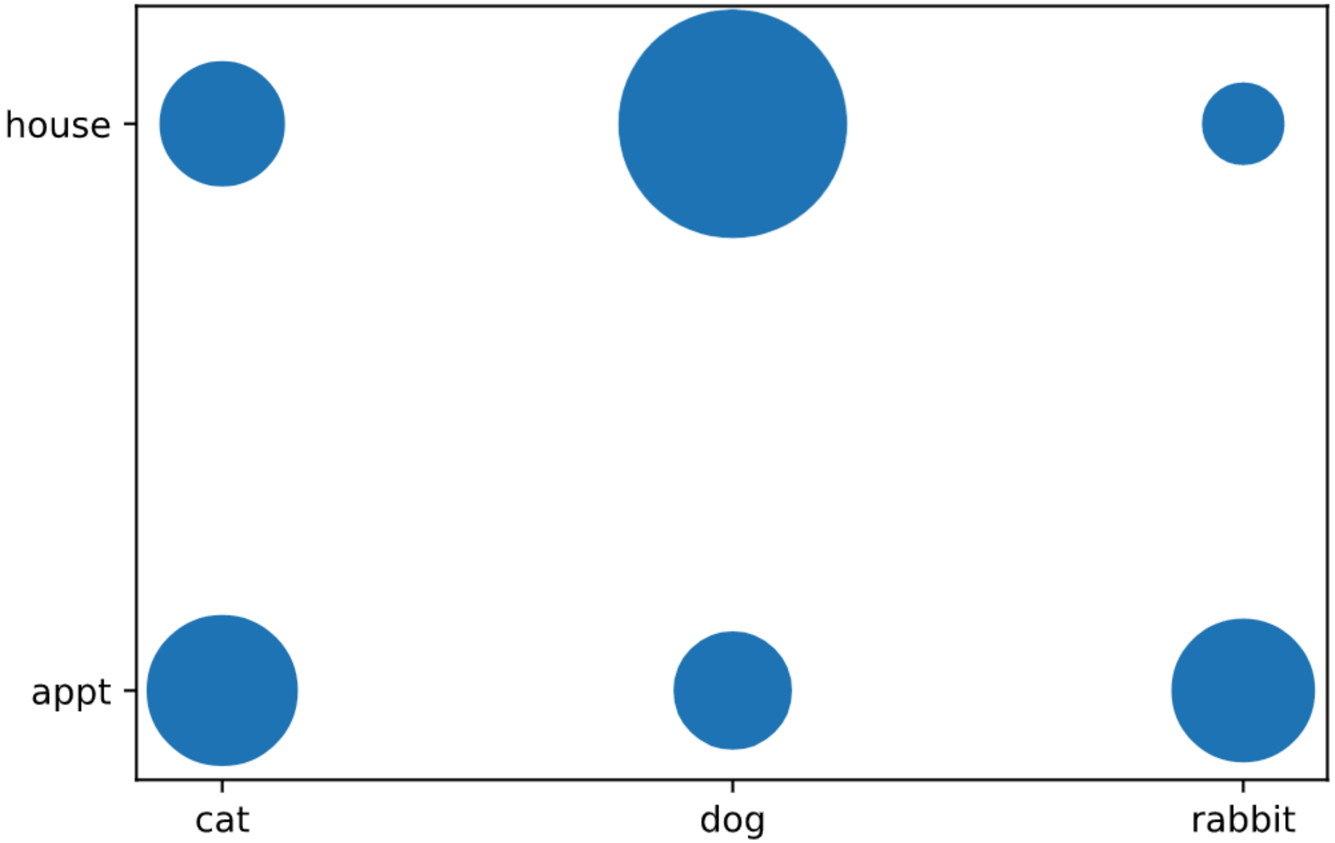
Heatmap (matrix)

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



Bubble plot

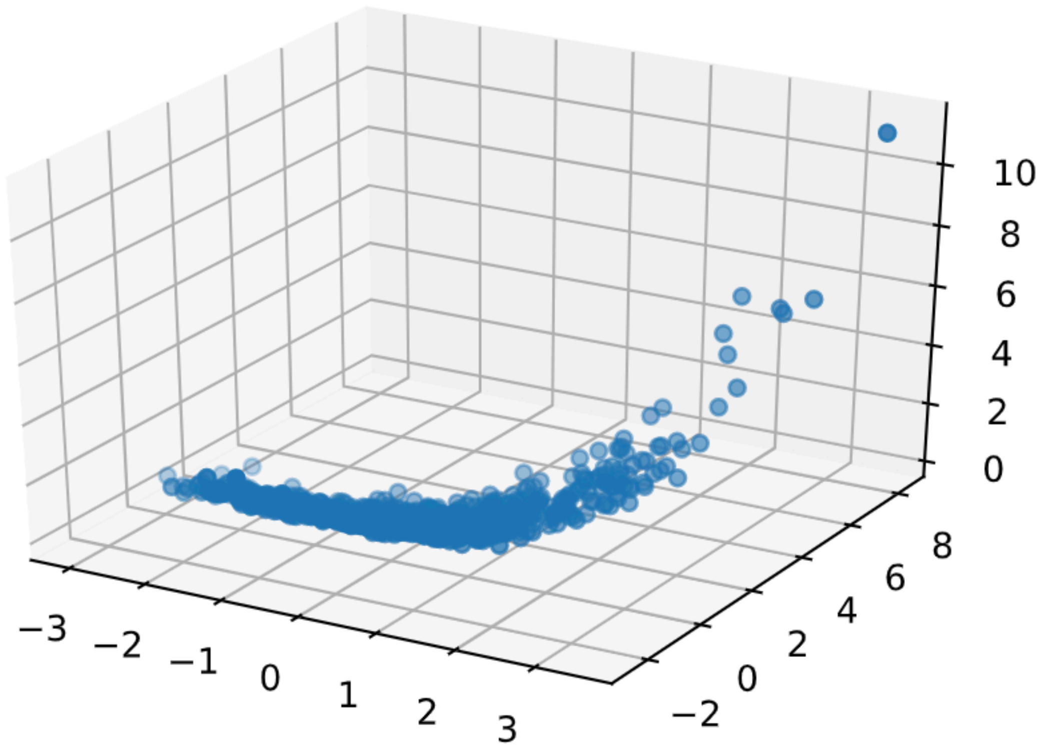
	Dim 1	Dim 2
Nominal		
Ordinal		
Interval	✗	✗
Ratio	✗	✗



3D+ DATA

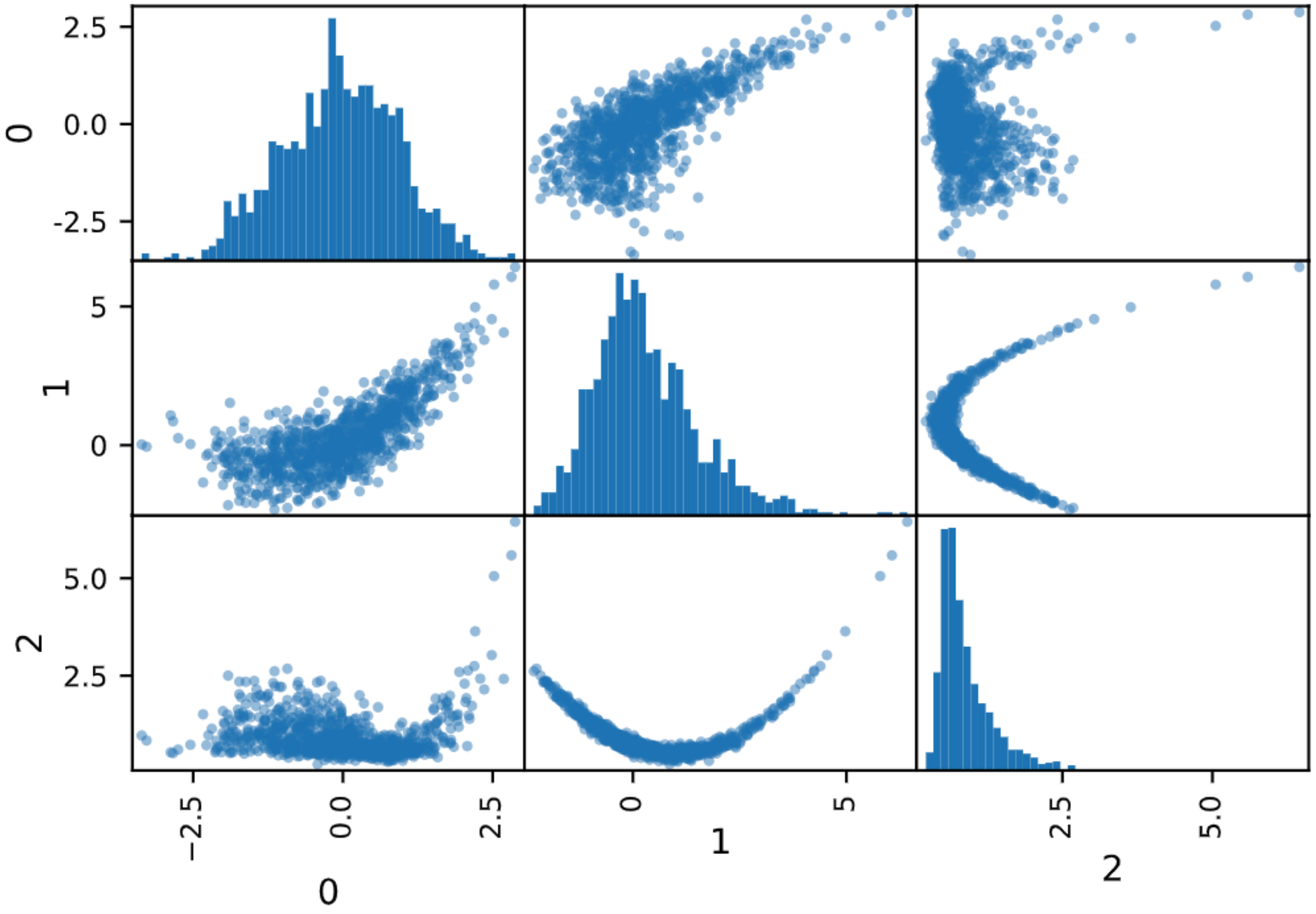
3D scatter plot

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



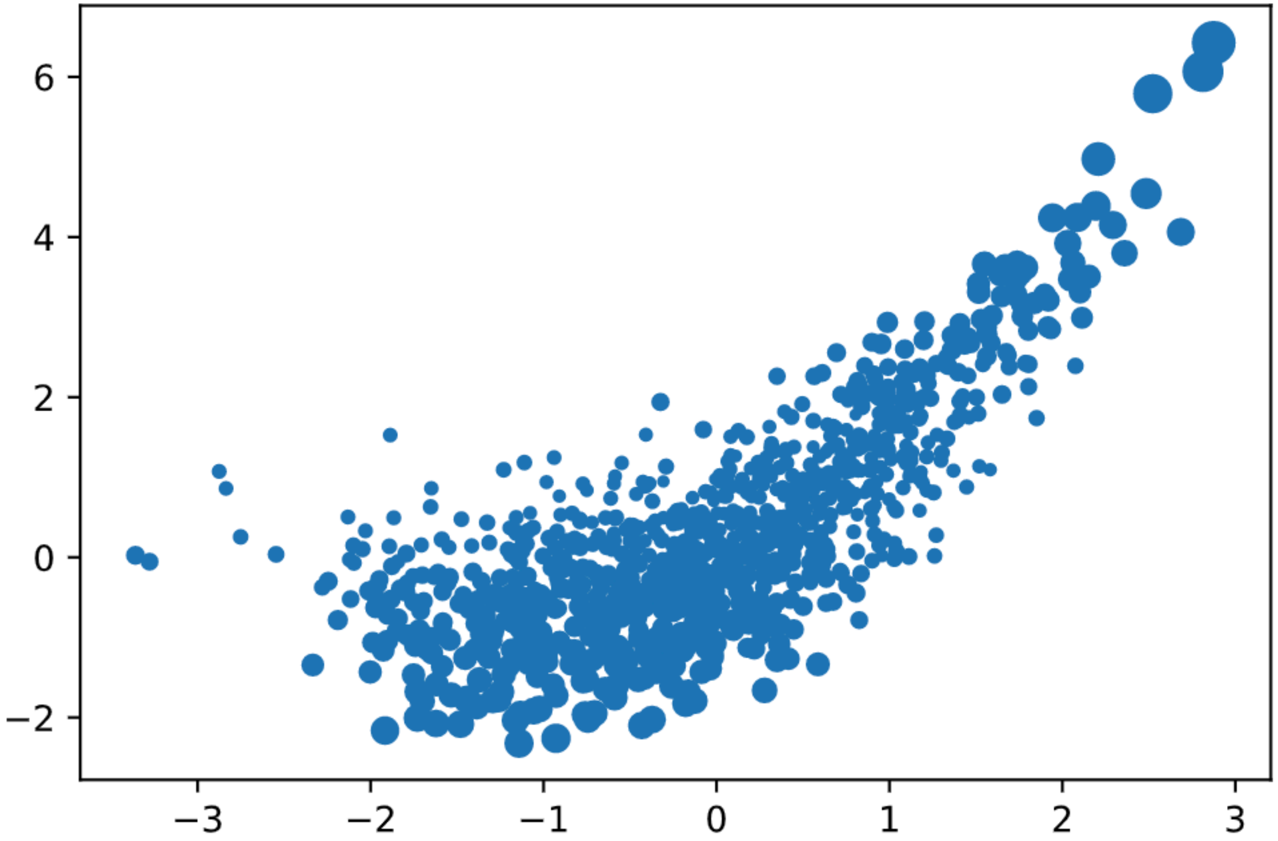
Scatter plot matrix

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



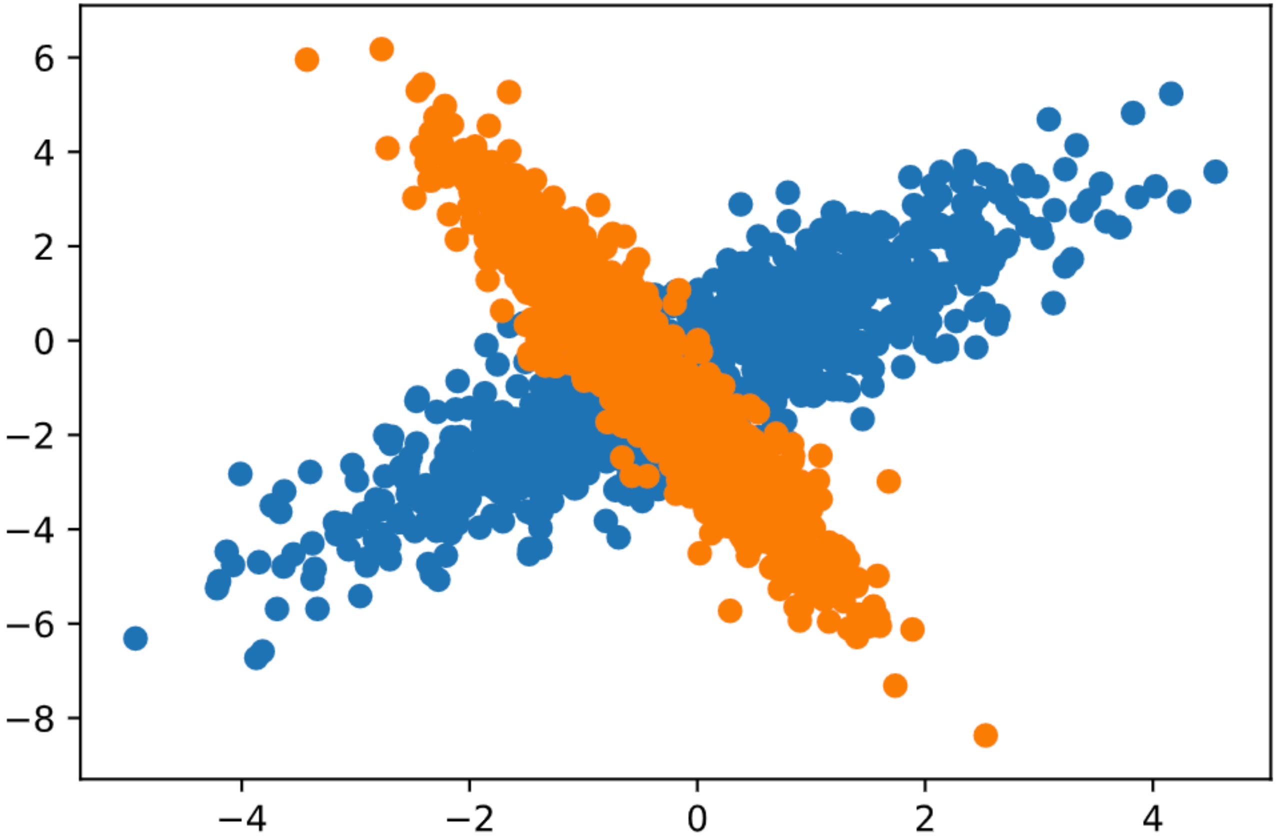
Bubble plot

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



Color scatter plot

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



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