

Teaching Visualization: Integrating Theory + Practice

Michael Freeman
University of Washington
@mf_viz

Integrating theory (*concepts*)
and practice (*implementation*)
enhances visualization skills
and literacy.

Integrating theory (*design*)
and practice (*programming*)
enhances visualization skills
and literacy.

Today's Discussion

Brief Introduction

Teaching through the integration of theory and practice:

- Selecting visual layouts
- Choosing graphical encodings
- Implementing visualizations

W Are you able to get Poll Everywhere to work?

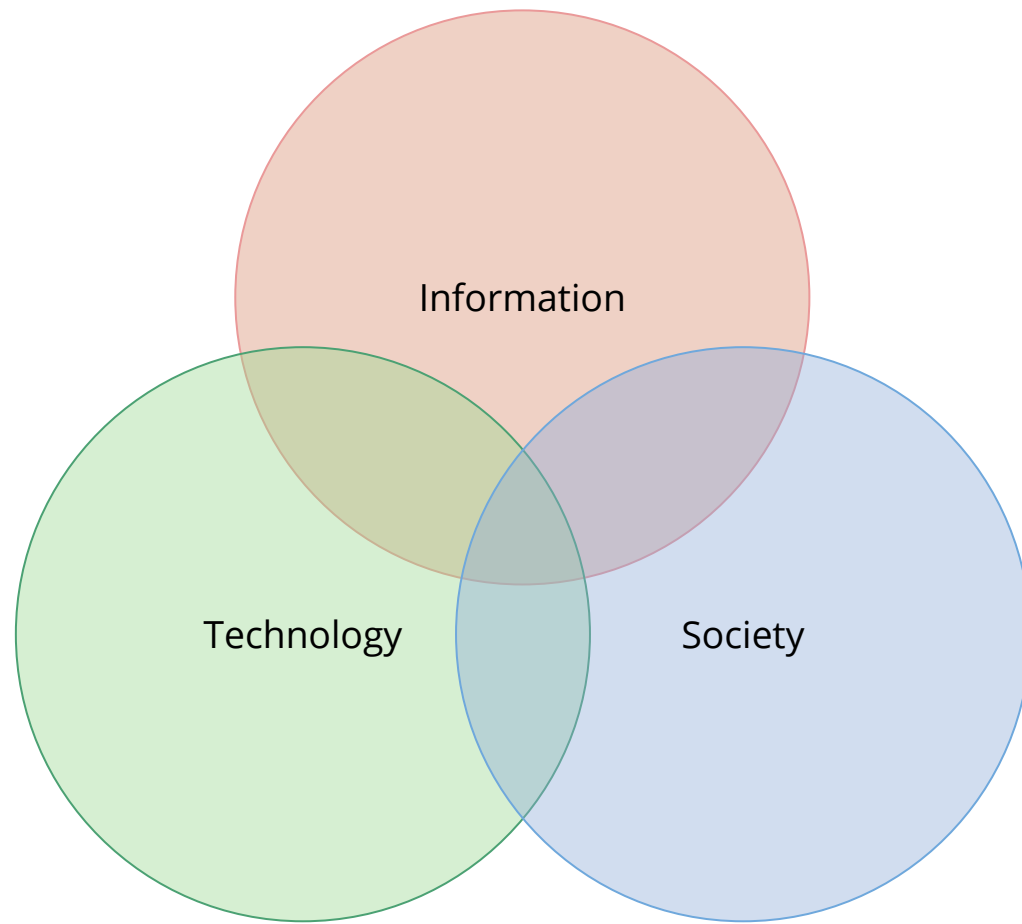
Yes

No
(?)

Introduction



Faculty member at the UW Information School



Courses I Teach

INFO 200: Intellectual Foundations of Informatics

INFO 201: Technical Foundations of Informatics

INFO 340: Client-side Development

INFO 370: Core Methods in Data Science

INFO 474: Interactive Data Visualization

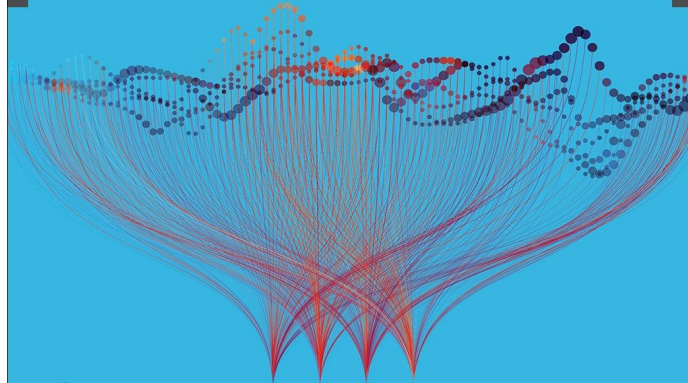
INFO 478: Population Health Informatics

ADDISON WESLEY DATA & ANALYTICS SERIES



PROGRAMMING SKILLS FOR DATA SCIENCE

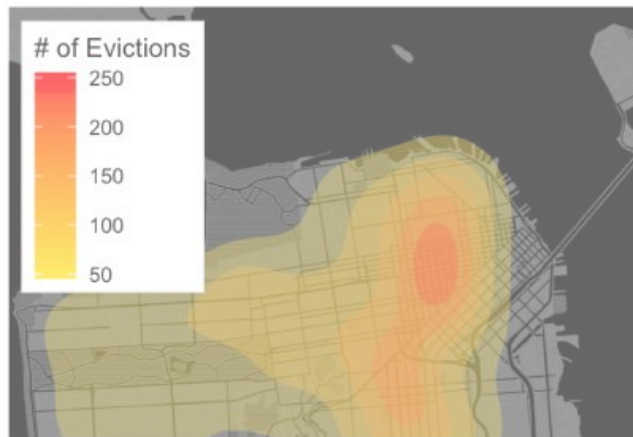
Start Writing Code to Wrangle,
Analyze, and Visualize Data with R



MICHAEL FREEMAN | JOEL ROSS



Recently published book (bit.ly/ps4ds), available on [safaribooksonline](https://www.safaribooksonline.com)



Fatal Police Shootings

Level of Analysis

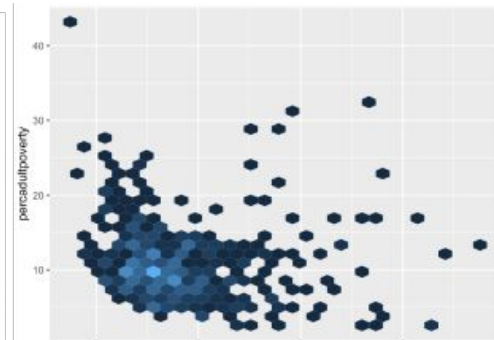
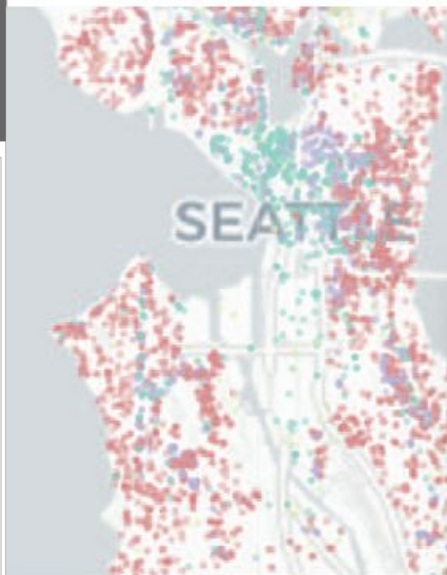
race

gender

race

body_camera

threat_level



Country

Change in Life Exp.

Maldives

39.8 years

Bhutan

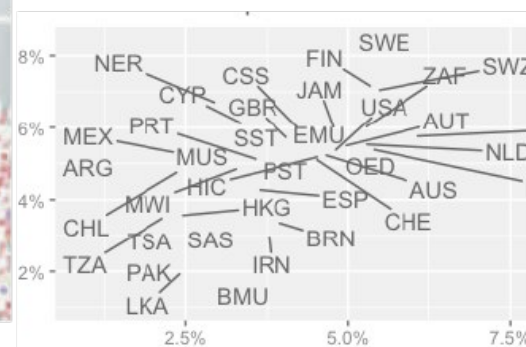
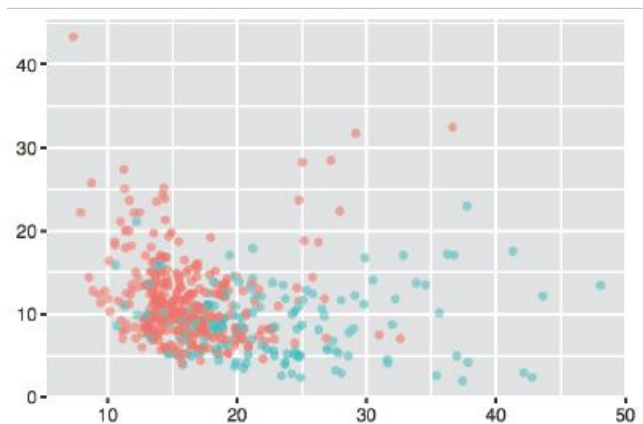
35.3 years

Timor-Leste

34.8 years

Nepal

34.7 years



Exercises + Solutions are available online: <http://bit.ly/ps4ds-code>

Selecting Layouts

Theory provides
foundational
guidance to *start* the
visualization design
process.

<u>Data Type(s)</u>	<u>Question of interest</u>	<u>Visual layouts</u>

Theory: provide guidance on chart type based on **data type** and **question** of interest

<u>Data Type(s)</u>	<u>Question of interest</u>	<u>Visual layouts</u>
1 Continuous	<i>How is my variable distributed?</i>	Histogram, box-plot, violin plot
1 Categorical	<i>How often does each value appear?</i>	Bar chart
1 continuous X 1 categorical	<i>Is the continuous variable similarly distributed within each grouping?</i>	Small multiples of 1 continuous, add a categorical encoding (color)
1 continuous X 1 continuous	<i>Are these variables correlated?</i>	Scatterplot
1 categorical X 1 categorical	<i>How often do these values co-occur?</i>	Heatmap
2+ continuous variables	<i>Is each pair of variables correlated?</i>	Scatterplot matrix
2 continuous X 1 categorical	<i>Is the relationship between variables similar within each group?</i>	Small multiples of 2 continuous, add a categorical encoding (color)

Theory: provide guidance on chart type based on **data type** and **question** of interest

	email	hours	difficulty	assignment
1	iwRmqcOBrQ@uw.edu	2.0	NA	a1
2	fEvjyJlcPy@uw.edu	2.0	3	a1
3	q7GNUealtn@uw.edu	NA	7	a1
4	RpwEsvwHPN@uw.edu	5.0	4	a1
5	IH9Qsy7k2D@uw.edu	1.5	3	a1
6	hvHrAQtMX2@uw.edu	6.0	6	a2
7	El7G0fsBru@uw.edu	4.0	5	a2
8	Zc5CfplHn5@uw.edu	4.0	5	a2
9	RTLYFnMaLz@uw.edu	6.0	5	a2
10	14OXkMCujP@uw.edu	2.0	4	a2
11	28CPcg2LGf@uw.edu	8.0	8	a3
12	gq7Q5Wo1rj@uw.edu	15.0	NA	a3

Practice: applying principles to data of interest (assignment feedback from students)

W

What is the data type you're working with if you want to understand the distribution of the number of hours worked on assignment 1?

	email	hours	difficulty	assignment
1	iwRmqcOBrQ@uw.edu	2.0	NA	a1
2	fEvjyJlcPy@uw.edu	2.0	3	a1
3	q7GNUealtn@uw.edu	NA	7	a1
4	RpwEsvwHPN@uw.edu	5.0	4	a1
5	IH9Qsy7k2D@uw.edu	1.5	3	a1
6	hvHrAQtMX2@uw.edu	6.0	6	a2
7	El7G0fsBru@uw.edu	4.0	5	a2
8	Zc5CfplHn5@uw.edu	4.0	5	a2
9	RTLYFnMaLz@uw.edu	6.0	5	a2
10	14OXkMCujP@uw.edu	2.0	4	a2
11	28CPcg2LGf@uw.edu	8.0	8	a3
12	gg7Q5Wo1rj@uw.edu	15.0	NA	a3

Categorical 1

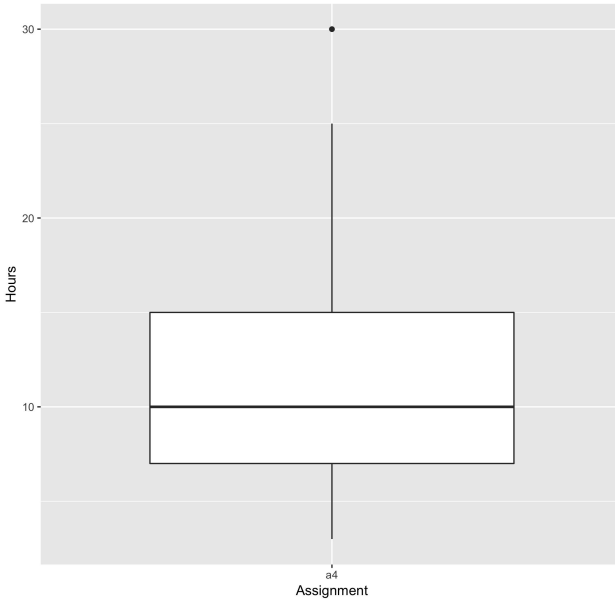
100%

Continuous 2

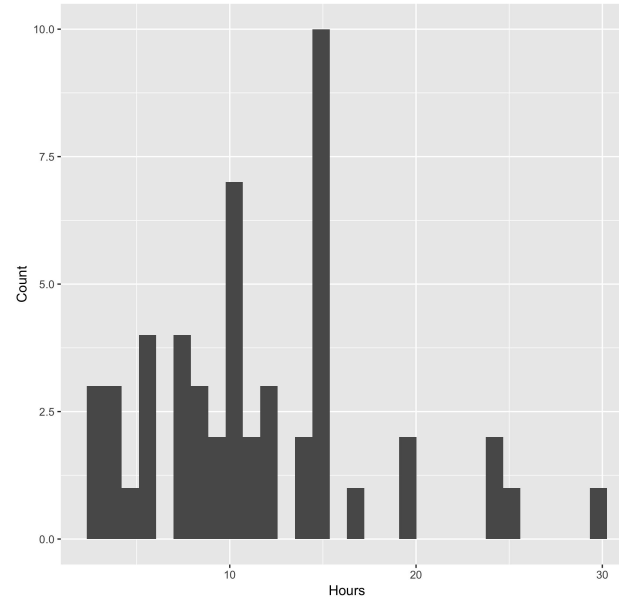
0% 20% 40% 60% 80% 100%

<u>Variable Type(s)</u>	<u>Question of interest</u>	<u>Visual Layout</u>
1 Continuous	<i>How is my variable distributed?</i>	Histogram, box-plot, violin plot

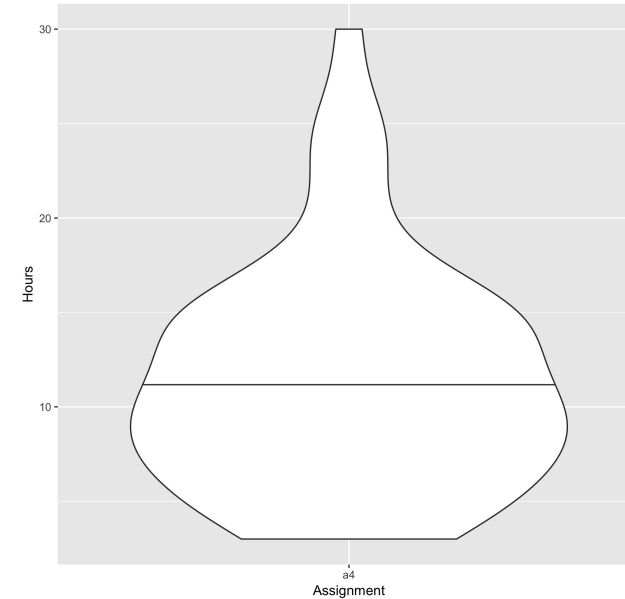
Distribution of Hours Spent per Assignment



Distribution of Reported Hours per Assignment



Distribution of Hours Spent per Assignment



Practice: how can you visualize the distribution of hours spent on assignment 4?

What is the data type you're working with if you want to understand the distribution of the number of hours worked *each* assignment?

	email	hours	difficulty	assignment
1	iwRmqcOBrQ@uw.edu	2.0	NA	a1
2	fEvjyJlcPy@uw.edu	2.0	3	a1
3	q7GNUealtn@uw.edu	NA	7	a1
4	RpwEsvwHPN@uw.edu	5.0	4	a1
5	IH9Qsy7k2D@uw.edu	1.5	3	a1
6	hvhHrAQtMX2@uw.edu	6.0	6	a2
7	El7G0fsBru@uw.edu	4.0	5	a2
8	Zc5CfplHn5@uw.edu	4.0	5	a2
9	RTLYFnMaLz@uw.edu	6.0	5	a2
10	14OXkMCujP@uw.edu	2.0	4	a2
11	28CPcg2LGf@uw.edu	8.0	8	a3
12	gq7Q5Wo1rj@uw.edu	15.0	NA	a3

2 categorical
variables

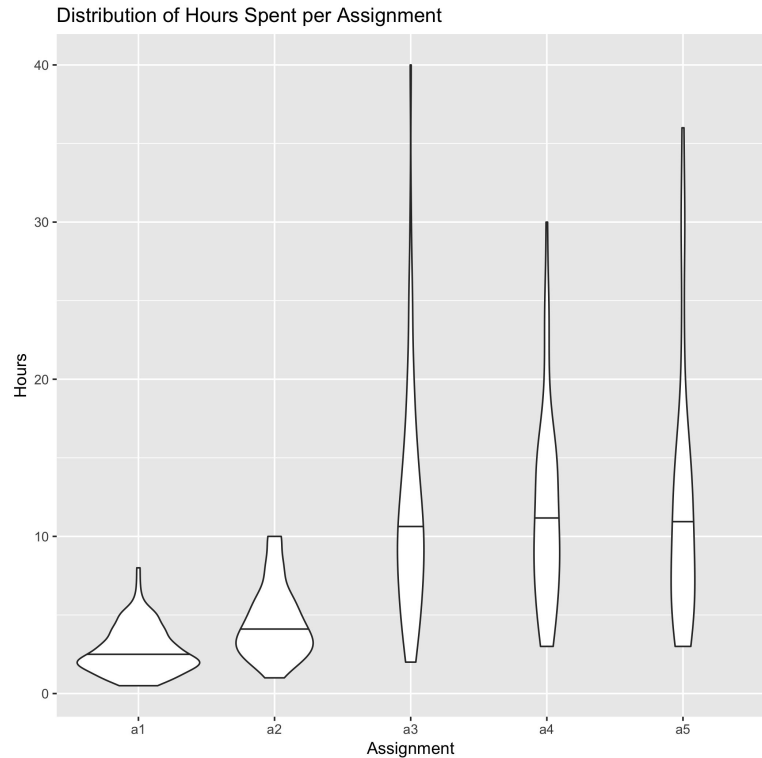
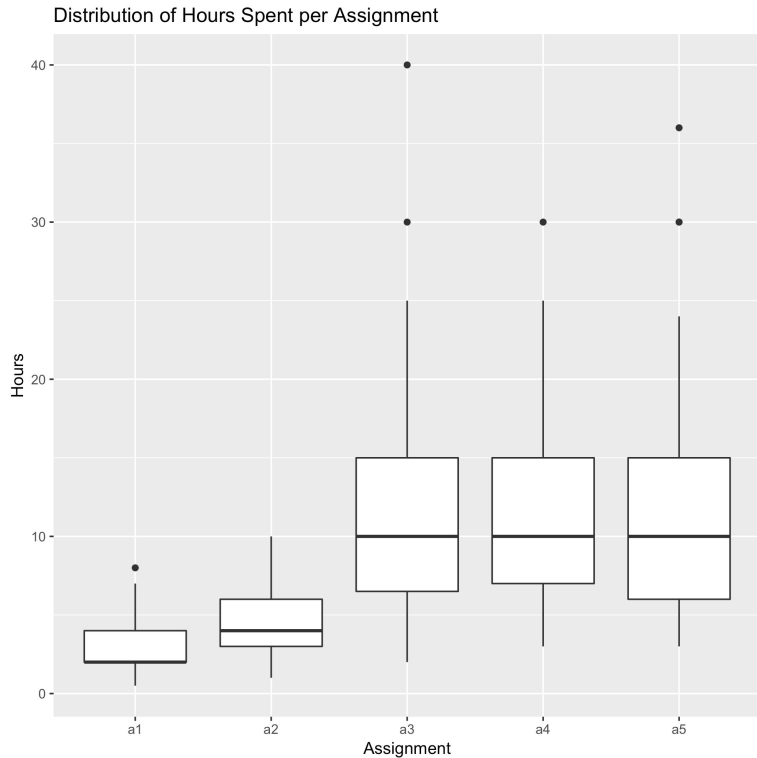
1 categorical and 1
continuous variable

2 continuous
variables

100%

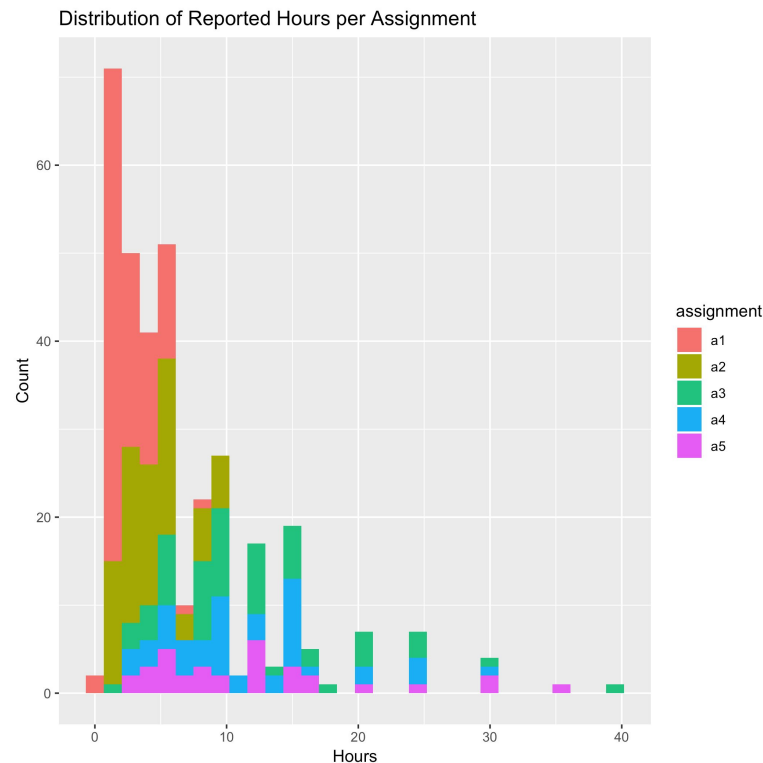
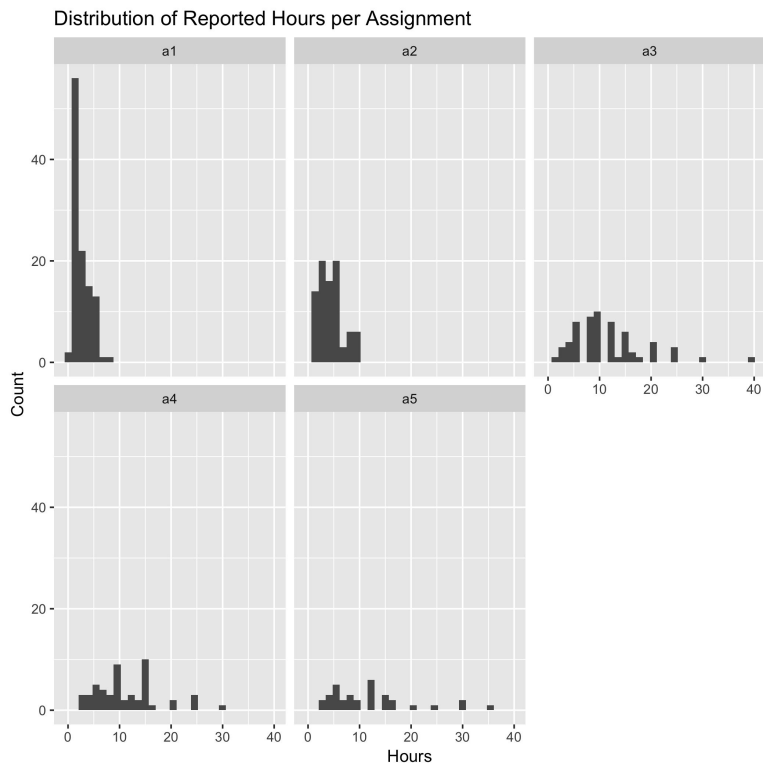
0% 20% 40% 60% 80% 100%

<u>Variable Type(s)</u>	<u>Question of interest</u>	<u>Visual Layout</u>
1 continuous X 1 categorical	<i>Is the continuous variable similarly distributed within each grouping?</i>	Small multiples of 1 continuous, add a categorical encoding (color)



Practice: how can you visualize the distribution of hours spent **each** assignment?

<u>Variable Type(s)</u>	<u>Question of interest</u>	<u>Visual Layout</u>
1 continuous X 1 categorical	<i>Is the continuous variable similarly distributed within each grouping?</i>	Small multiples of 1 continuous, add a categorical encoding (color)



Practice: how can you visualize the distribution of hours spent **each** assignment?

What are the data types you're working with if you want to assess how the reported difficulty and hours worked are related for assignment 4?

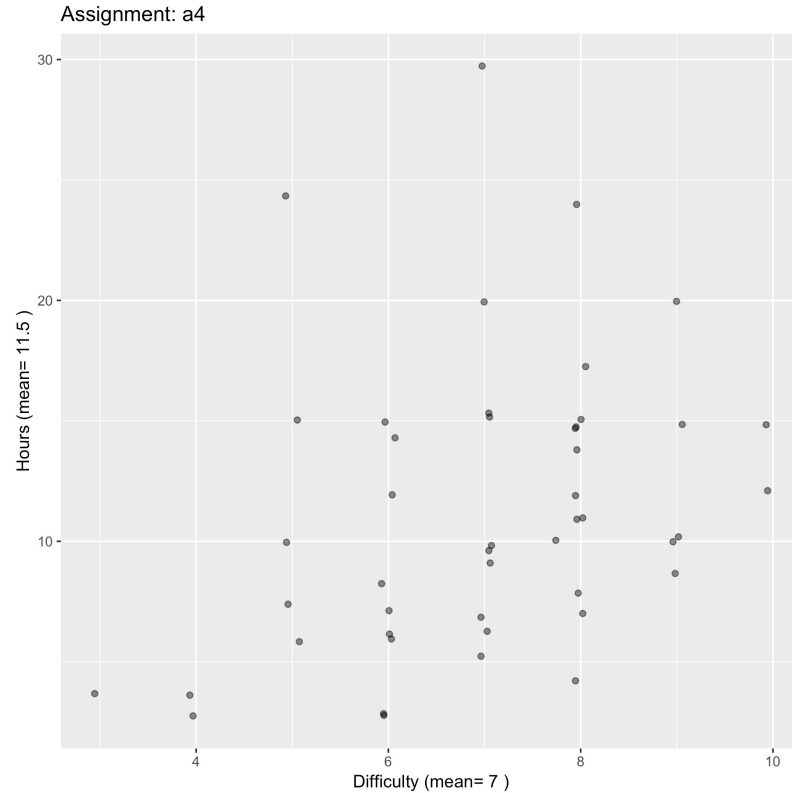
	email	hours	difficulty	assignment
1	iwRmqcOBrQ@uw.edu	2.0	NA	a1
2	fEvjyJlcPy@uw.edu	2.0	3	a1
3	q7GNUealtn@uw.edu	NA	7	a1
4	RpwEsvwHPN@uw.edu	5.0	4	a1
5	IH9Qsy7k2D@uw.edu	1.5	3	a1
6	hVHrAQtMX2@uw.edu	6.0	6	a2
7	El7G0fsBru@uw.edu	4.0	5	a2
8	Zc5CfplHn5@uw.edu	4.0	5	a2
9	RTLYFnMaLz@uw.edu	6.0	5	a2
10	14OXkMCujP@uw.edu	2.0	4	a2
11	28CPcg2LGf@uw.edu	8.0	8	a3
12	gq7Q5Wo1rj@uw.edu	15.0	NA	a3

2 categorical
variables

1 categorical and 1
continuous variable

2 continuous
variables

<u>Variable Type(s)</u>	<u>Question of interest</u>	<u>Visual Layout</u>
1 continuous X 1 continuous	<i>Are these variables correlated?</i>	Scatterplot



Practice: how can you visualize hours v.s. difficulty for assignment 4?

What are the data types you're working with if you want to assess how the reported difficulty and hours worked are related for *each assignment*?

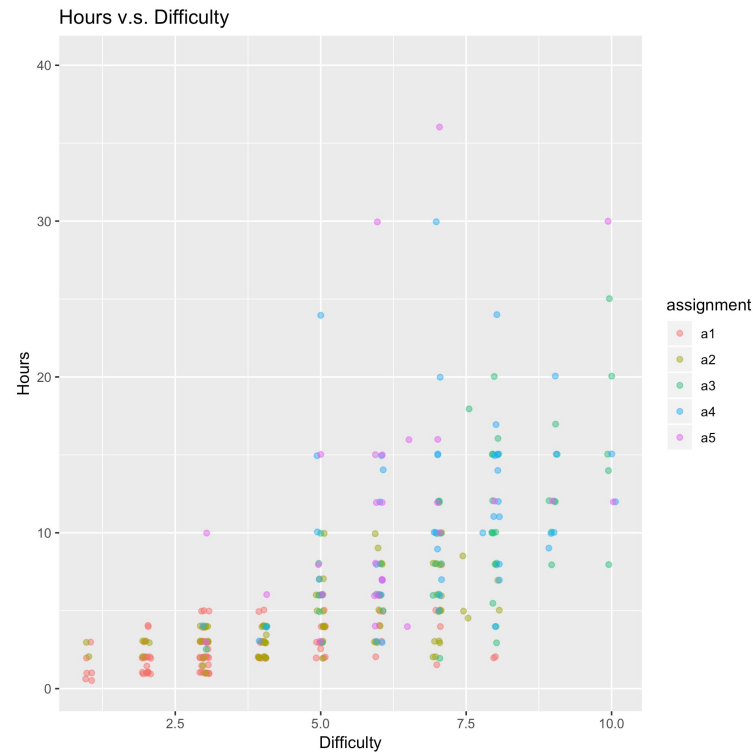
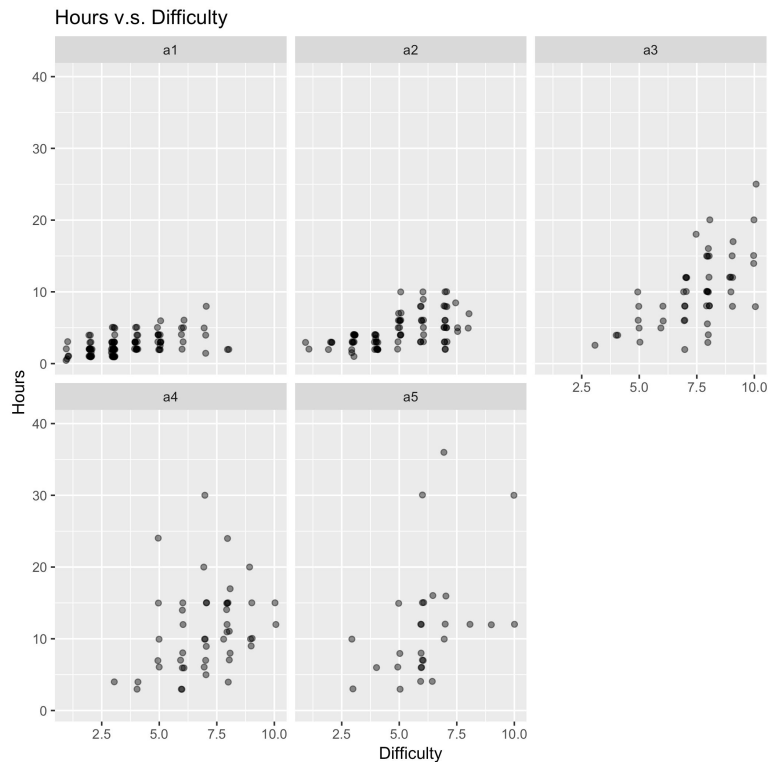
	email	hours	difficulty	assignment
1	iwRmqcOBrQ@uw.edu	2.0	NA	a1
2	fEvjyJlcPy@uw.edu	2.0	3	a1
3	q7GNUealtn@uw.edu	NA	7	a1
4	RpwEsvwHPN@uw.edu	5.0	4	a1
5	IH9Qsy7k2D@uw.edu	1.5	3	a1
6	hvHrAQtMX2@uw.edu	6.0	6	a2
7	El7G0fsBru@uw.edu	4.0	5	a2
8	Zc5CfplHn5@uw.edu	4.0	5	a2
9	RTLYFnMaLz@uw.edu	6.0	5	a2
10	14OXkMCujP@uw.edu	2.0	4	a2
11	28CPcg2LGf@uw.edu	8.0	8	a3
12	gq7Q5Wo1rj@uw.edu	15.0	NA	a3

3 continuous
variables

2 categorical
variables and 1
continuous variable

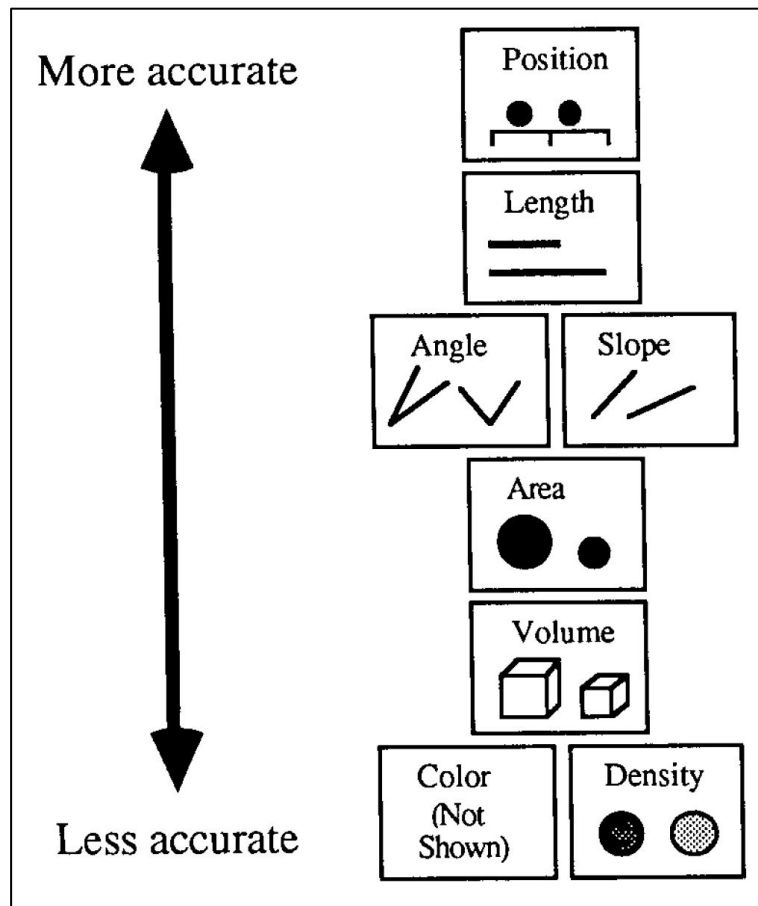
2 continuous
variables and 1
categorical variable

<u>Variable Type(s)</u>	<u>Question of interest</u>	<u>Visual Layout</u>
2 continuous X 1 categorical	<i>Is the relationship between variables similar within each group?</i>	Small multiples of 2 continuous, add a categorical encoding (color)



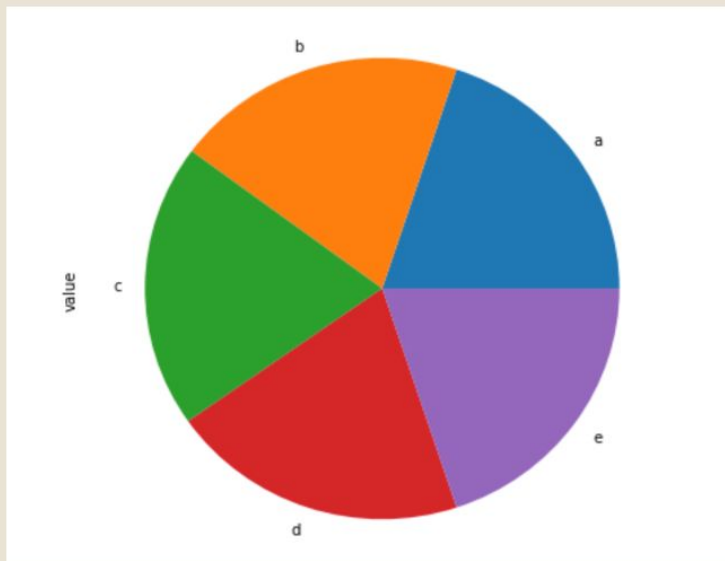
Practice: how can you visualize hours v.s. difficulty for **each** assignment ?

Selecting Graphical Encodings



Theory: choosing graphical encodings

W Which slice of the pie is largest?



a

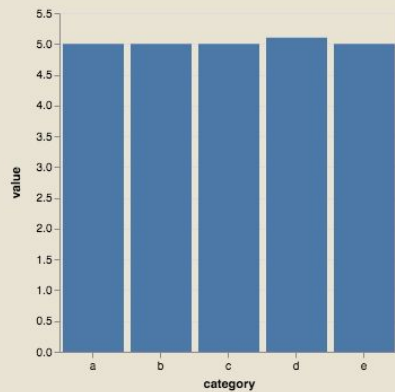
b

c

d

e

W Which bar is tallest?



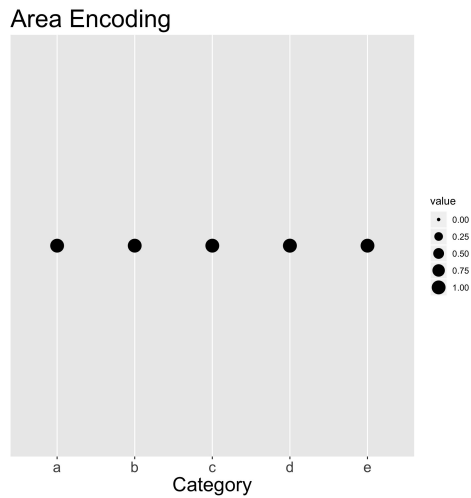
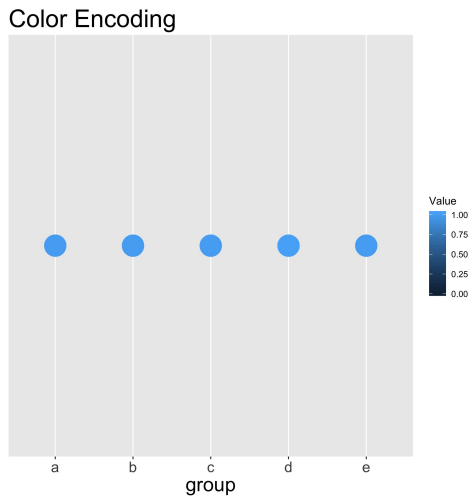
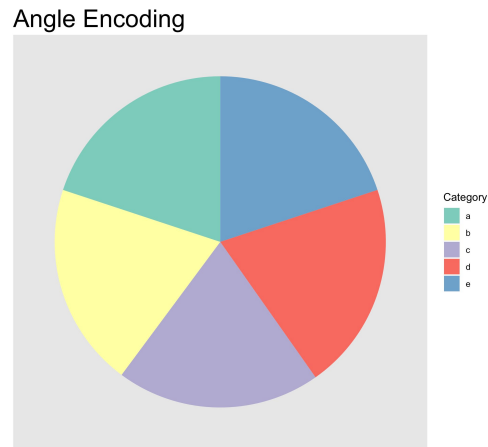
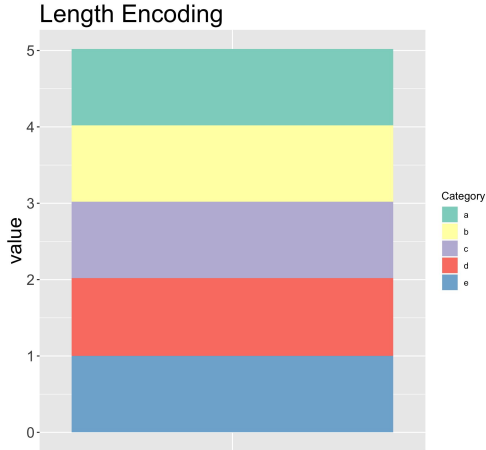
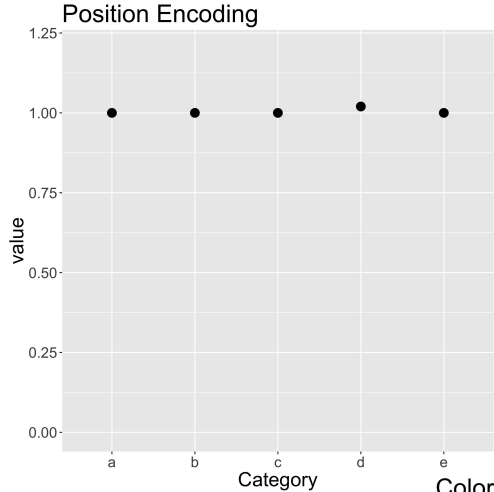
a

b

c

d

e



Theory: choosing graphical encodings

Programmatically Building Visualizations

Code is an
expression of
design choices.

W Do you use the R package ggplot2?

Yes, very frequently

Everyone once in a while

I've tried it

No, not yet

Grammar of Graphics

Provides a **consistent vocabulary** for the design choices we make:

- **Data** to be shown in the plot
- **Geometric** objects we wish to display
- **Aesthetic** mappings between our data values and their graphical encodings
- **Statistical** transformations to be performed on the data
- **Scales** of values to be applied to our aesthetics
- **Coordinate** system to organize our geometries
- **Facets** (groups) of our data to show in different plots (small multiples)

Expressing choices with ggplot2

Create a drawing canvas using the **ggplot()** function, then add layers of visual elements using the grammar

The **aes** function describes *which aesthetics* (x position, color, etc.) should be driven by *which data*

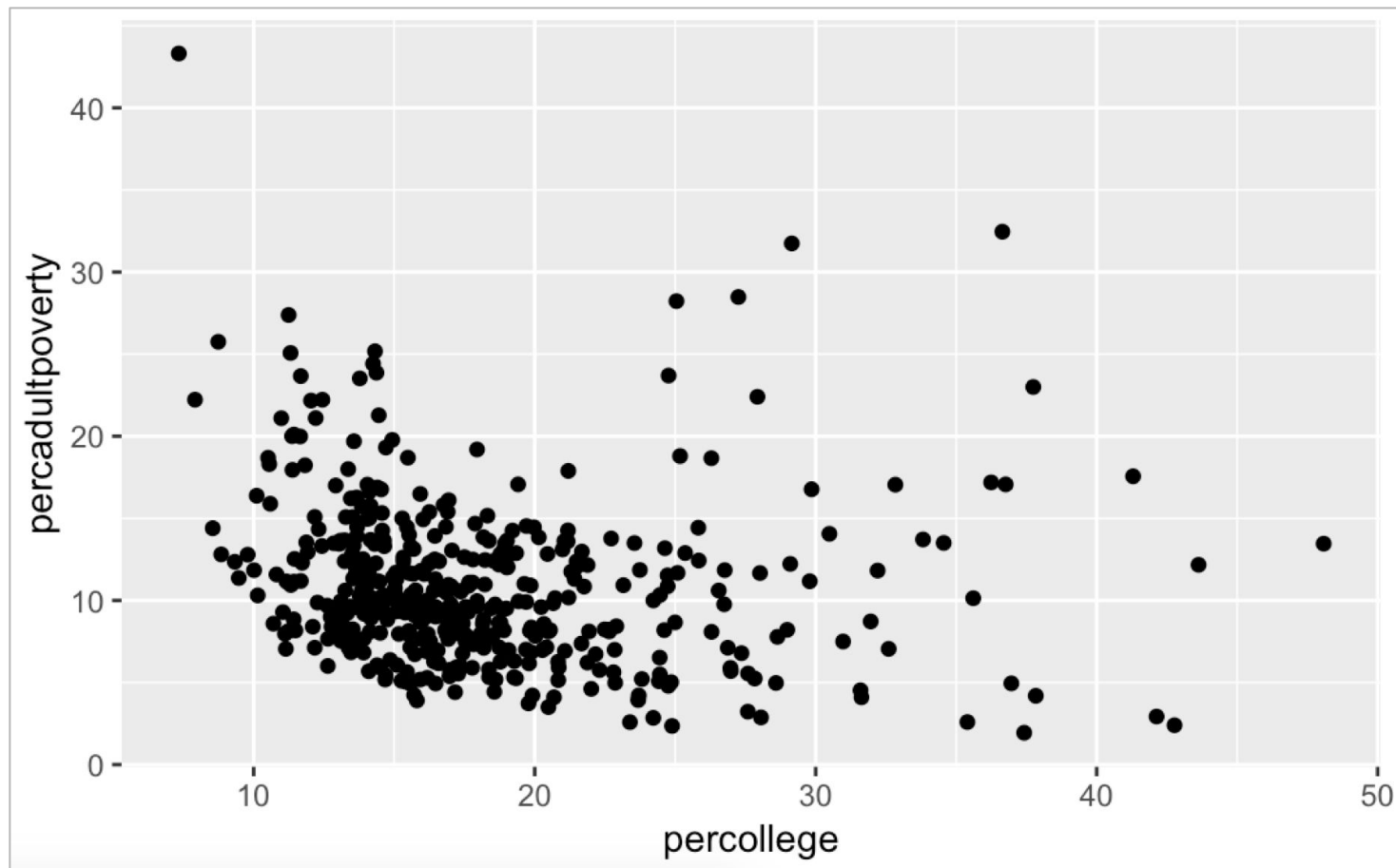
```
ggplot(data = midwest) +  
geom_point(mapping = aes(x = percollege, y = percadultpoverty))
```

Data to plot

Add a geometry

Geometry to add (circles)

Describe aesthetic mapping from data space to a visual space



ggplot2 Scatter Plot

W Do you use the JavaScript package D3.js?

Yes, very frequently

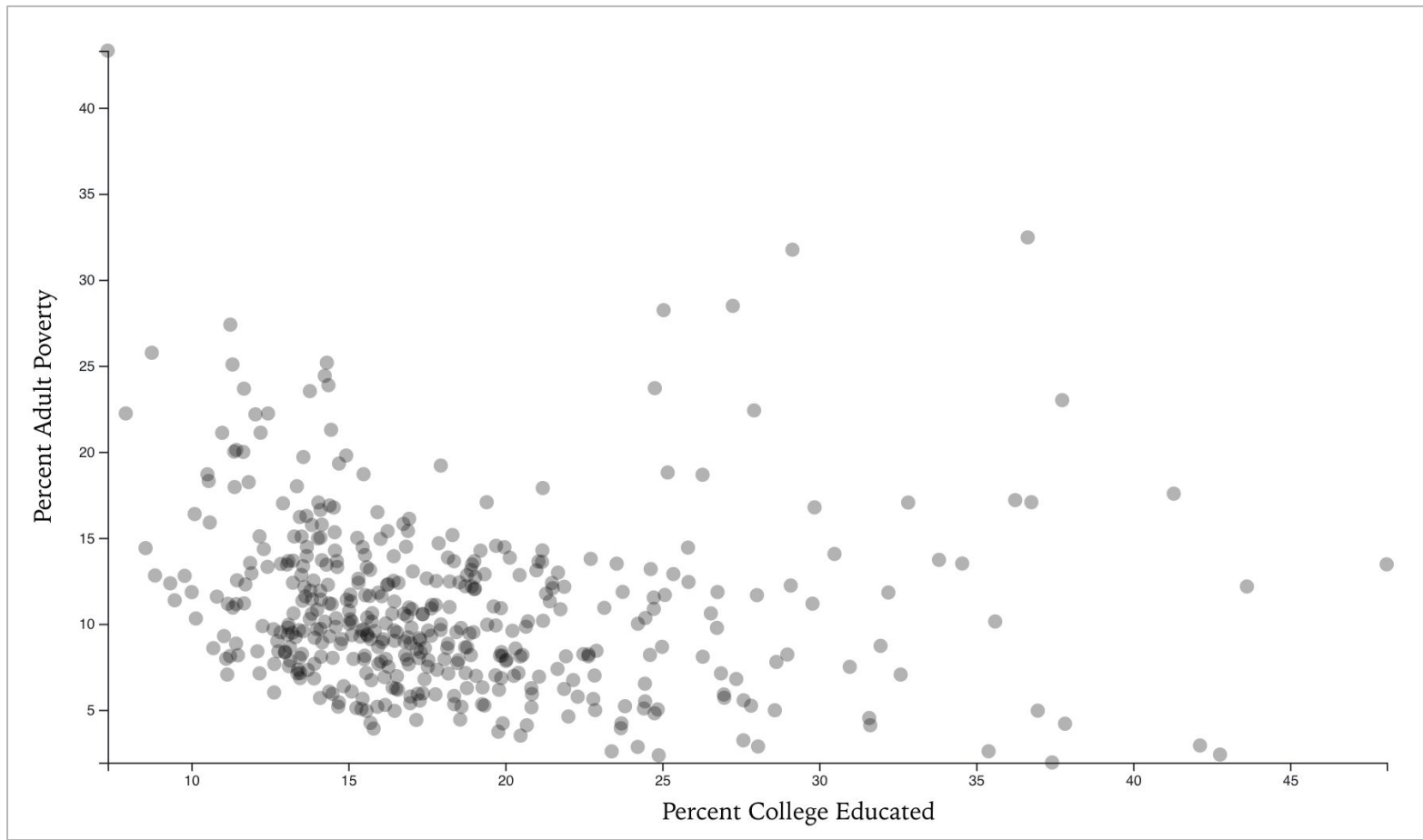
Everyone once in a
while

I've tried it

No, not yet

D3 is not a charting
library.

D3 is a library for
mapping from data
elements to visual
elements.



Building a scatterplot in D3

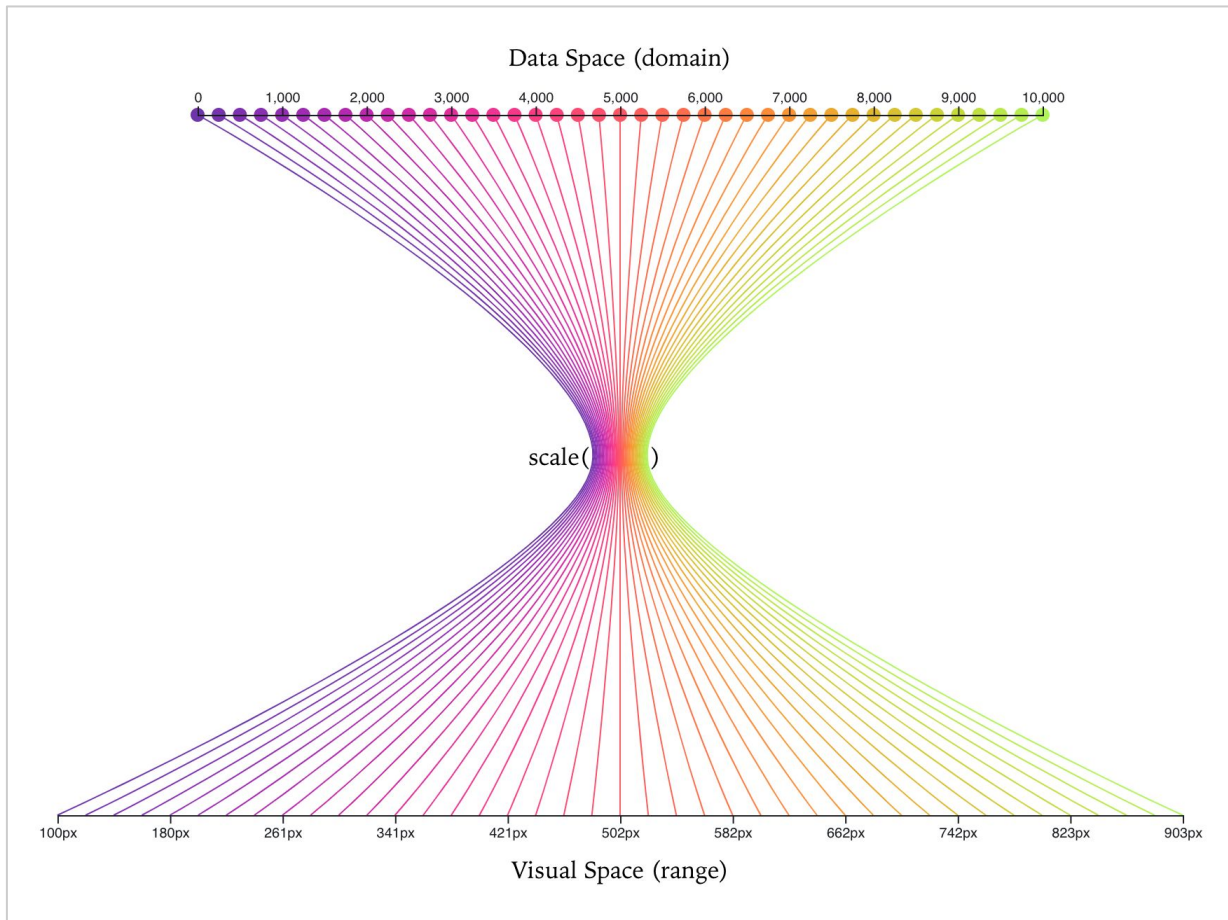
Scaling your data in D3

Define data minimum / maximum

```
// Determine min/max x and y values
let xMin = d3.min(data, (d) => d.x);
let xMax = d3.max(data, (d) => d.x);
```

Create a **function** to map from your data *domain* to a visual *range*

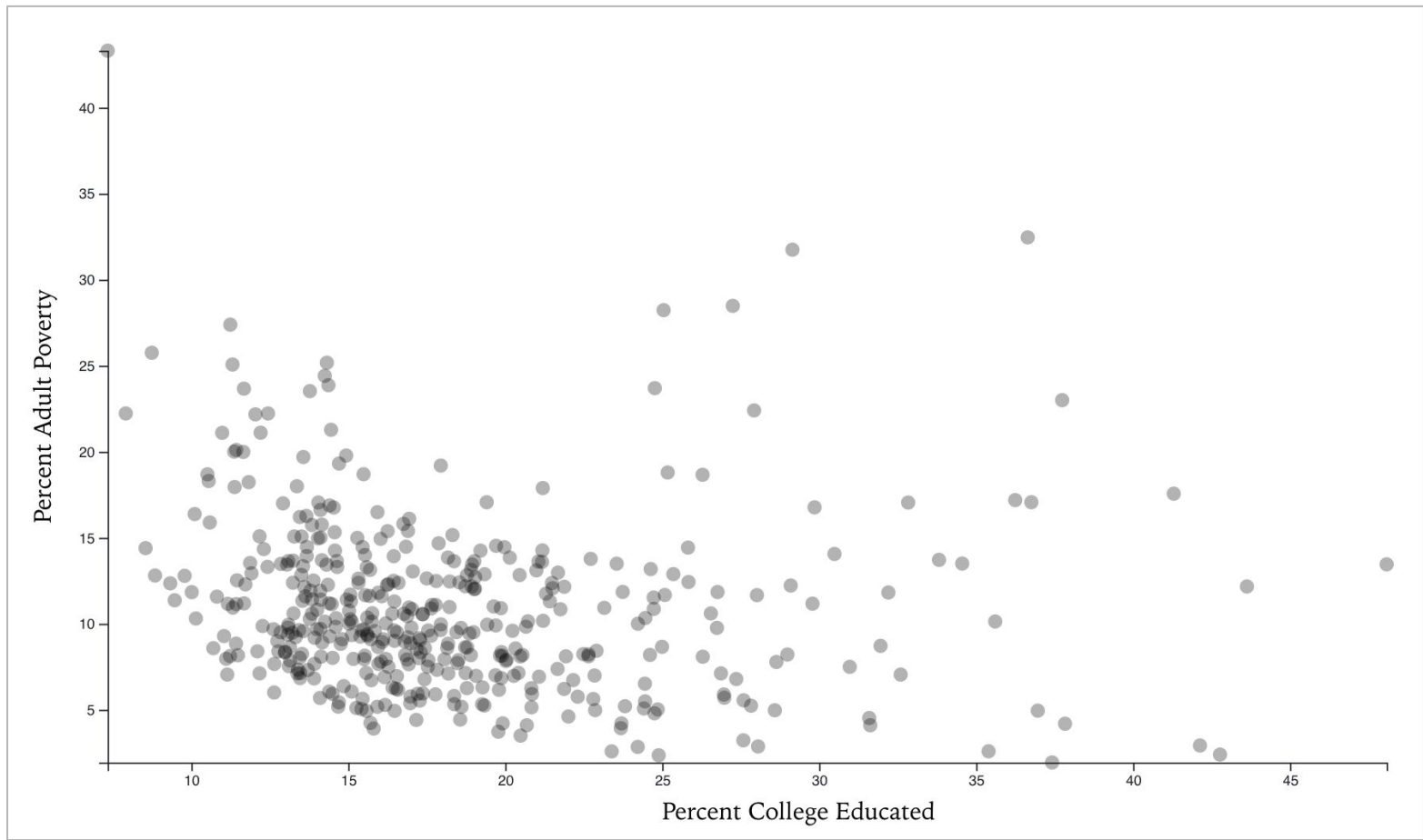
```
// Create scales using the domain of your data, and your visual range
let xScale = d3.scaleLinear()
  .domain([xMin, xMax])
  .range([0, width]);
```



Bind data to visual elements

Explicitly create a relationship between each data item and it's visual representation

```
// Bind data to the selection of elements and append circles, positioning w/scales
let circles = g.selectAll("circle")
  .data(data, (d) => d.id)
  .enter()
  .append("circle")
  .attr("cx", (d) => scales.x(d.x))
  .attr("cy", (d) => scales.y(d.y))
  .attr("r", 5);
```



Scatterplot in D3 ([link](#))

Integrating theory (*concepts*)
and practice (*implementation*)
enhances visualization skills
and literacy.

Integrating theory (*design*)
and practice (*programming*)
enhances visualization skills
and literacy.

Thank you

Twitter: @mf_viz

Book: bit.ly/ps4ds

Book-exercises: <http://bit.ly/ps4ds-code>

Slides: bit.ly/freeman_sdss

