NC STATE
UNIVERSITY

Clinical Data Visualization in R and Python

Wednesday, Nov 5
Workshop: noon to 1p
Open coding session: 1p to 2p

Outline

- Welcome and introduction
  - Sample dataset overview
- Introduction to visualization
- Types of and motivations for plots
- Exploratory data analysis with visualization
- Communicating time series and longitudinal perspectives using visualization

E

Welcome!

Who we are:

- Emily Griffith
  Professor of the Practice, Department of Statistics
  Director of Consulting, Data Science and AI Academy, NC State

- John Slankas
  Senior Research Scholar, Laboratory for Analytic Sciences, NC State

We manage a subaward from NC TraCS to help provide support for clinical research using data science techniques.

E,J

Today’s goals

- We will review
  - Introduction to visualization
  - Types of and motivations for plots
  - Exploratory data analysis with visualization
  - Communicating time series and longitudinal perspectives using visualization
- At 1p, we will have open coding hours for you to stay and work if you’d like
- We are using the same dataset for all examples, and code is available for you in R and Python

E

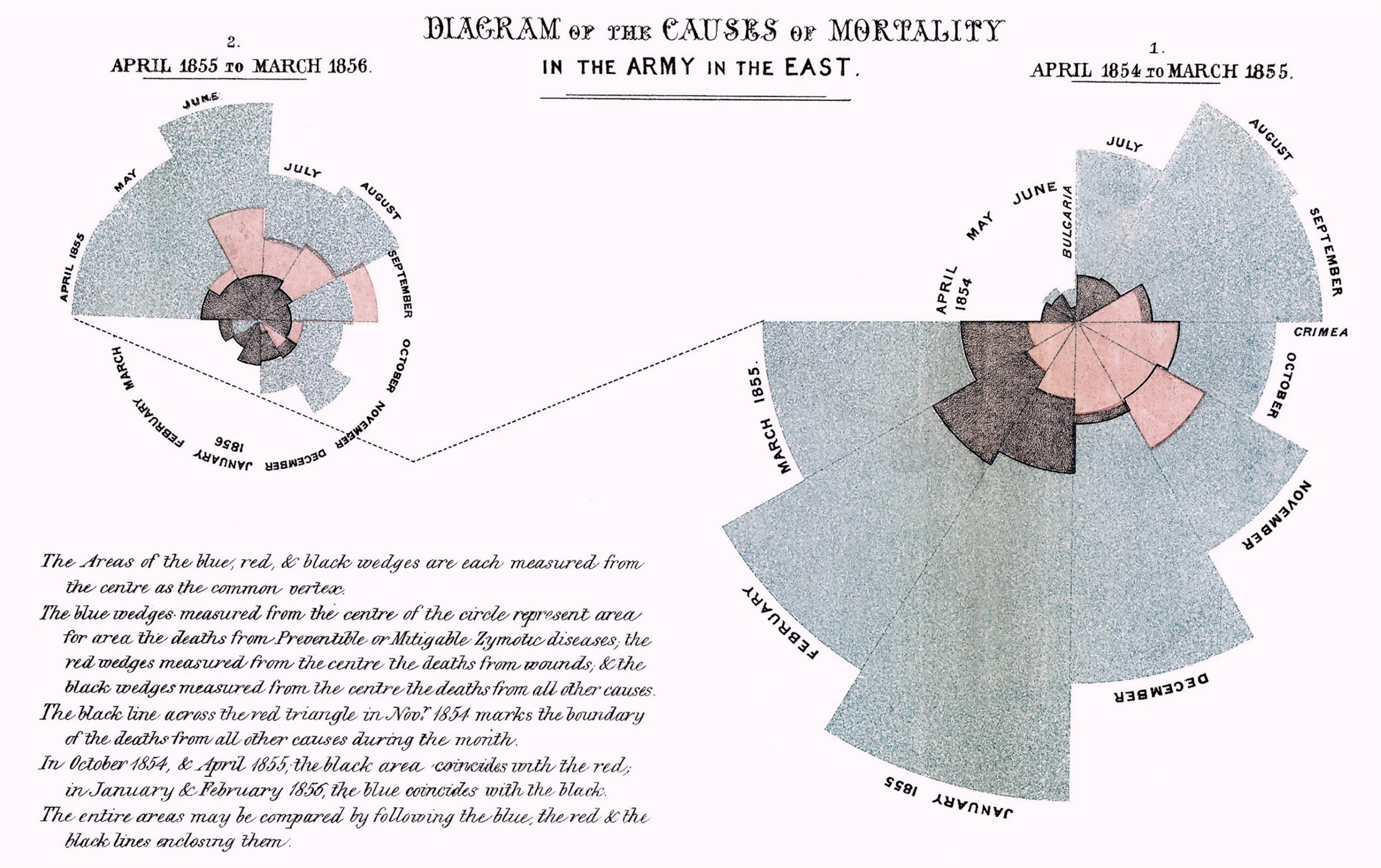
Sample data set

We are using a sample data set for all coding exercises so we can all look at the same examples.

- Length of inpatient and non-outpatient hospital stay
- Blood pressure (diastolic and systolic)
- Sex
- Oxygen Saturation
- Respiratory Rate
- Age at visit
- And more!

What is data visualization?

- Graphical representation of information and data
- Helps you see and understand trends, outliers, and patterns in data
- Often supplemental to statistical analysis



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr 1854 marks the boundary of the deaths from all other causes during the month.

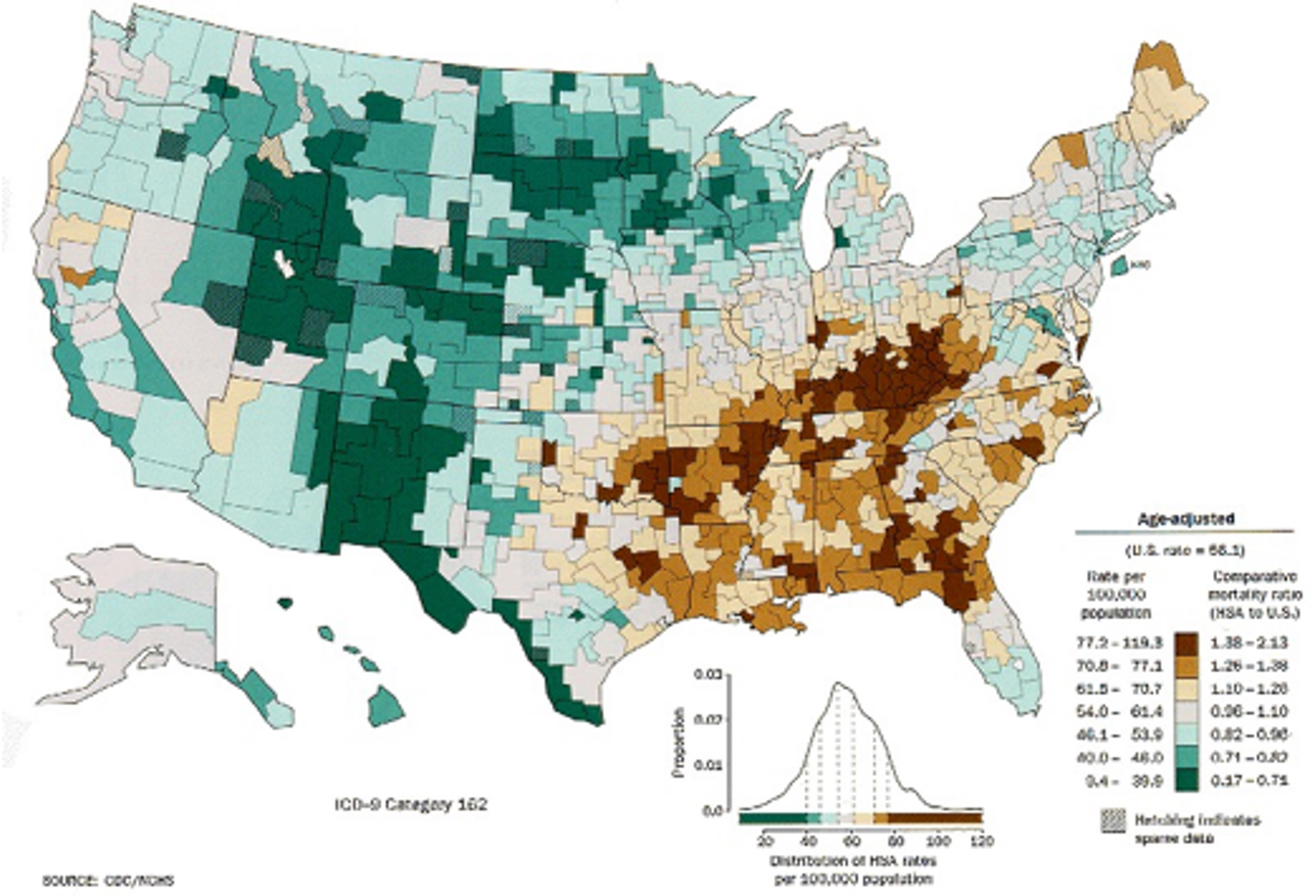
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

NC STATE UNIVERSITY

Why visualize your data?

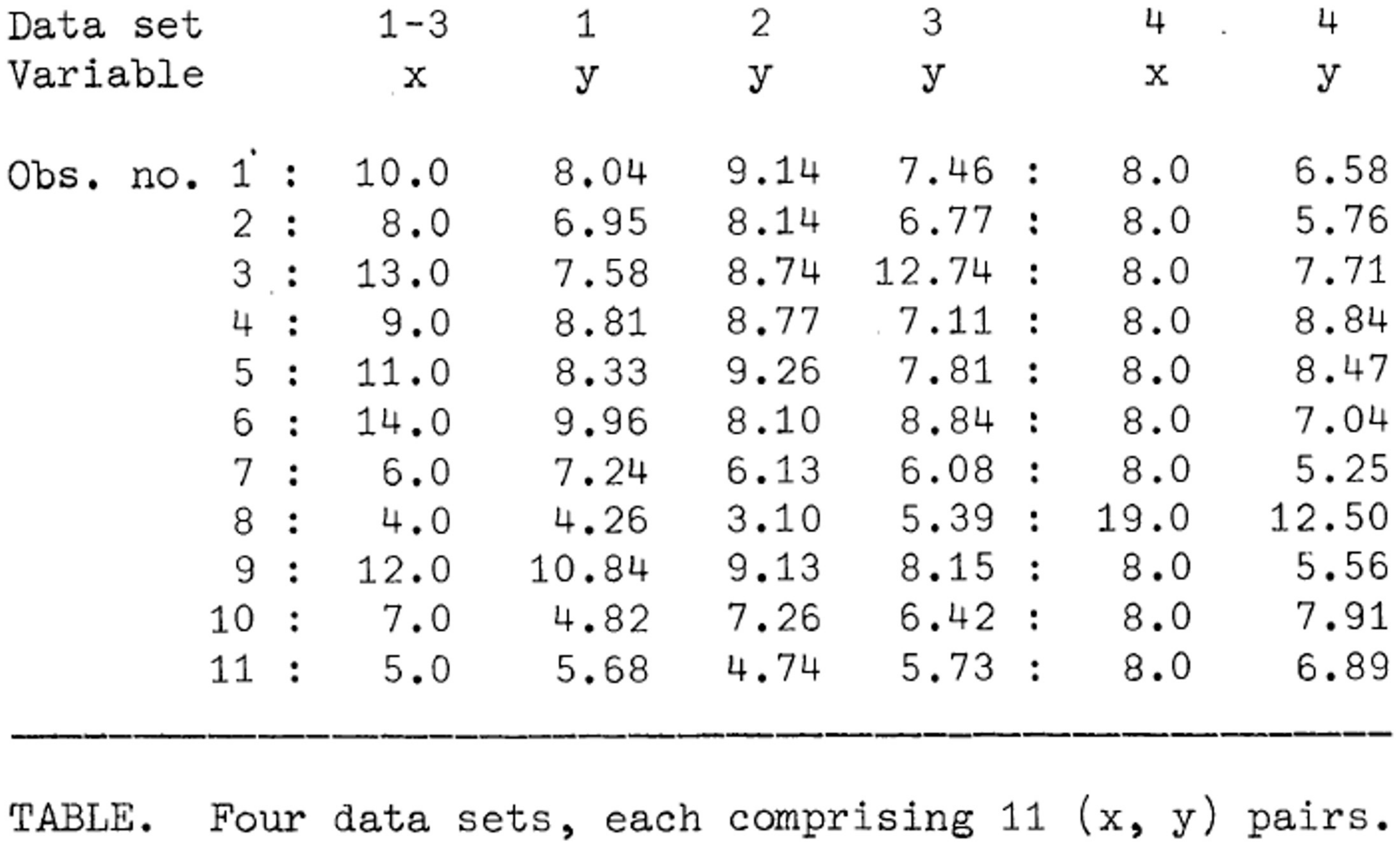
- Answer questions
- Make decisions
- Find patterns
- Present argument
- Tell a story
- Inspire



https://www.cdc.gov/nchs/products/other/atlas/lcwm.htm

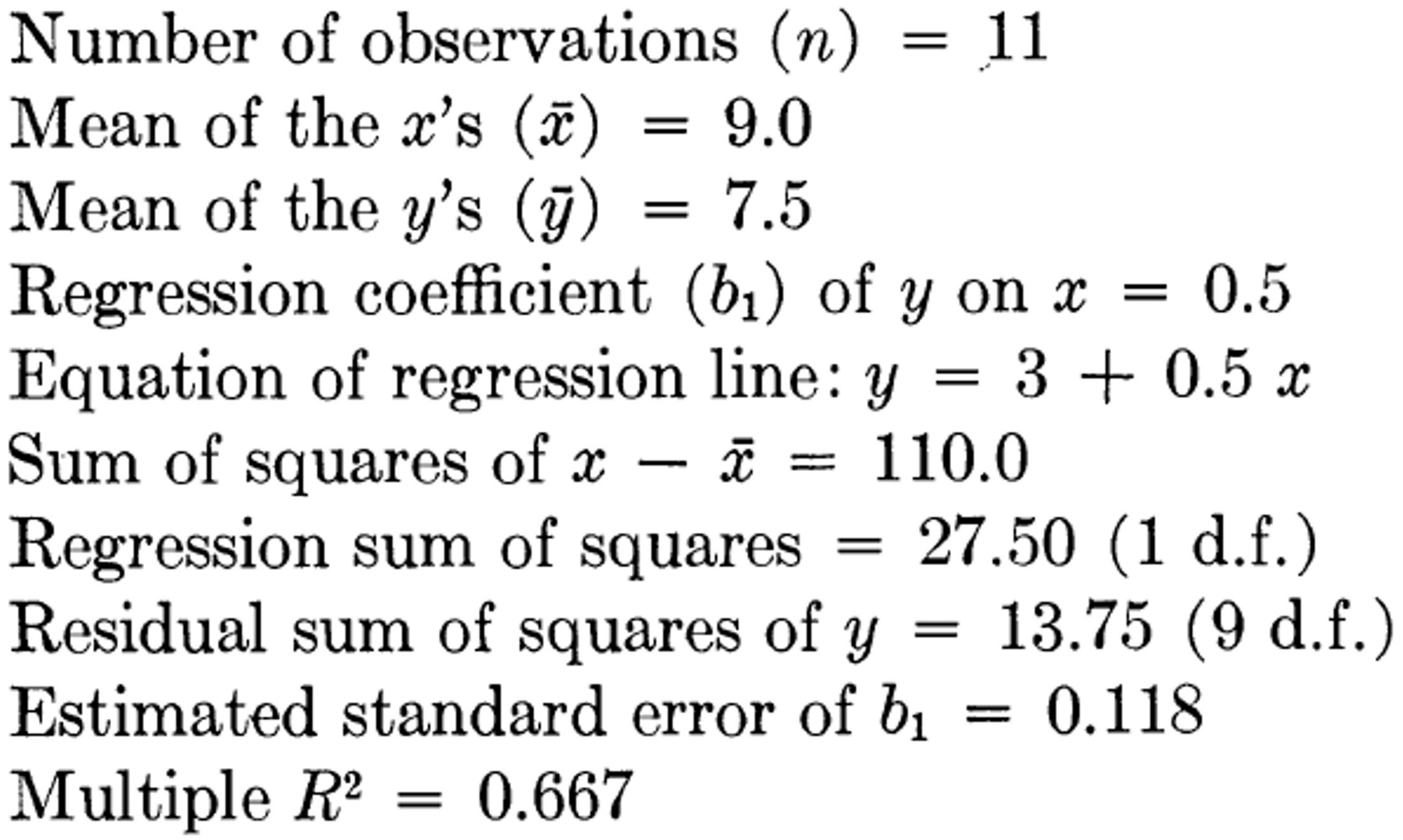
J

Why visualize your data? Anscombe 1973



Data set Variable		1-3 x	1 y	2 y	3 y	4 x	4 y
Obs. no.	1	10.0	8.04	9.14	7.46	8.0	6.58
	2	8.0	6.95	8.14	6.77	8.0	5.76
	3	13.0	7.58	8.74	12.74	8.0	7.71
	4	9.0	8.81	8.77	7.11	8.0	8.84
	5	11.0	8.33	9.26	7.81	8.0	8.47
	6	14.0	9.96	8.10	8.84	8.0	7.04
	7	6.0	7.24	6.13	6.08	8.0	5.25
	8	4.0	4.26	3.10	5.39	19.0	12.50
	9	12.0	10.84	9.13	8.15	8.0	5.56
	10	7.0	4.82	7.26	6.42	8.0	7.91
	11	5.0	5.68	4.74	5.73	8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.



Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

These four data sets have the same summary statistics!

Why visualize your data? Anscombe 1973

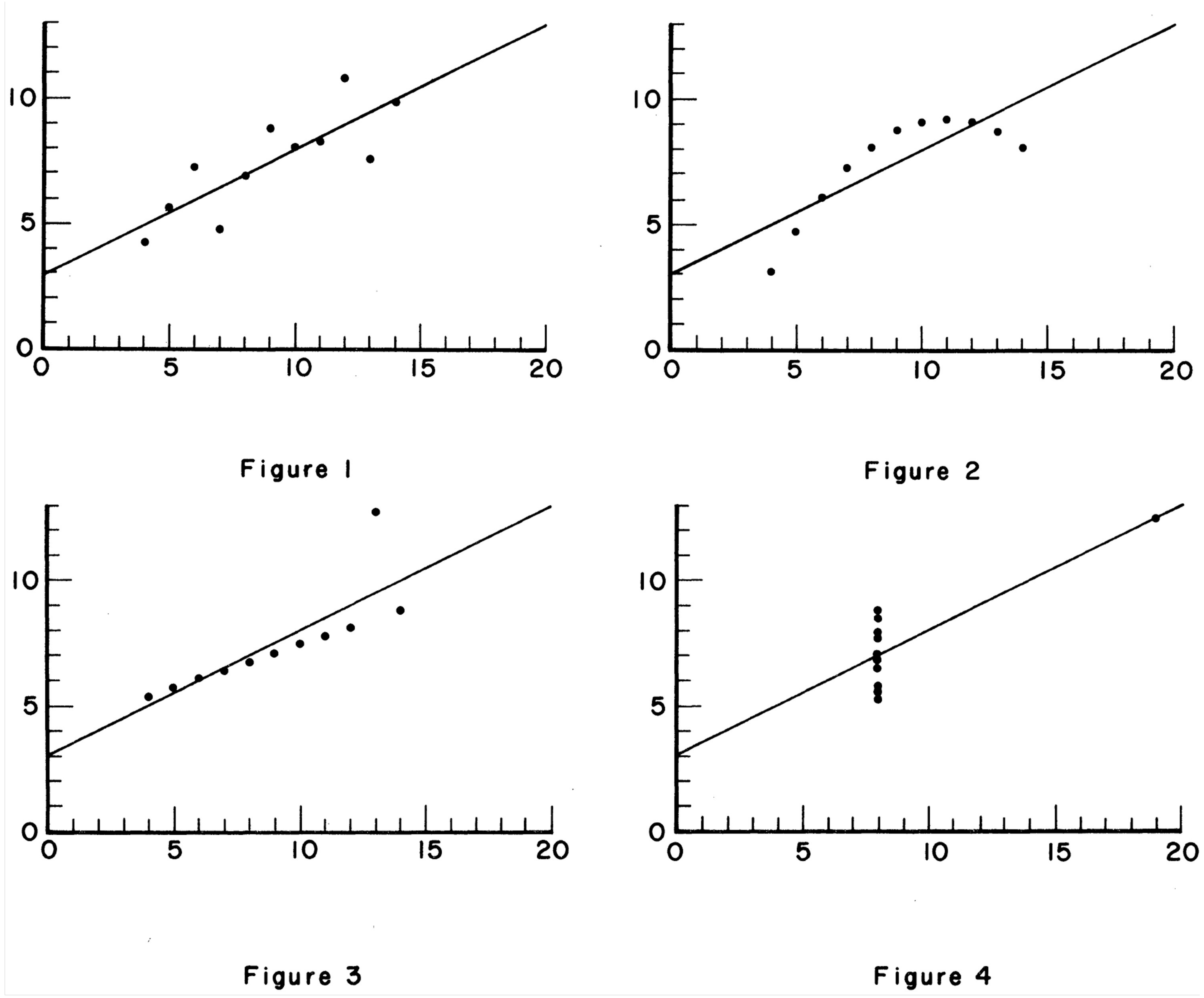


Figure 1

Figure 2

Figure 3

Figure 4

The differences between datasets are easy to see when visualized!

Motivation for Visualization: Exploratory vs. Explanatory

Exploratory

- Explore the data
- Assess a situation
- Determine how to proceed
- Decide what to do

Explanatory

- Present data and ideas
- Explain and inform
- Provide evidence and support
- Influence and persuade

Quick aside: What is Exploratory Data Analysis?

- Important first step in ANY data analysis project!
- Used to answer these questions:
    - What do your data look like?
    - Do the values make sense?
        - Can you look into the values that don't make sense?
    - Do you see any patterns?
    - Is there missing data?
    - Are there data in all the categories you have?
        - Do the counts make sense? Do the categories make sense?

NC STATE UNIVERSITY

How to use visualization in EDA?

One variable:

- Bar charts for categories
- Distribution plots for continuous variables
  - Histograms
  - Distributions

Two + variables:

- Scatterplot for correlations
  - Scatterplot matrix for multiple variables
- Heat map for correlations
- Boxplots and violin plots for one continuous variable across levels of one or more categorical variables

E,J

Data Quality Checks using EDA

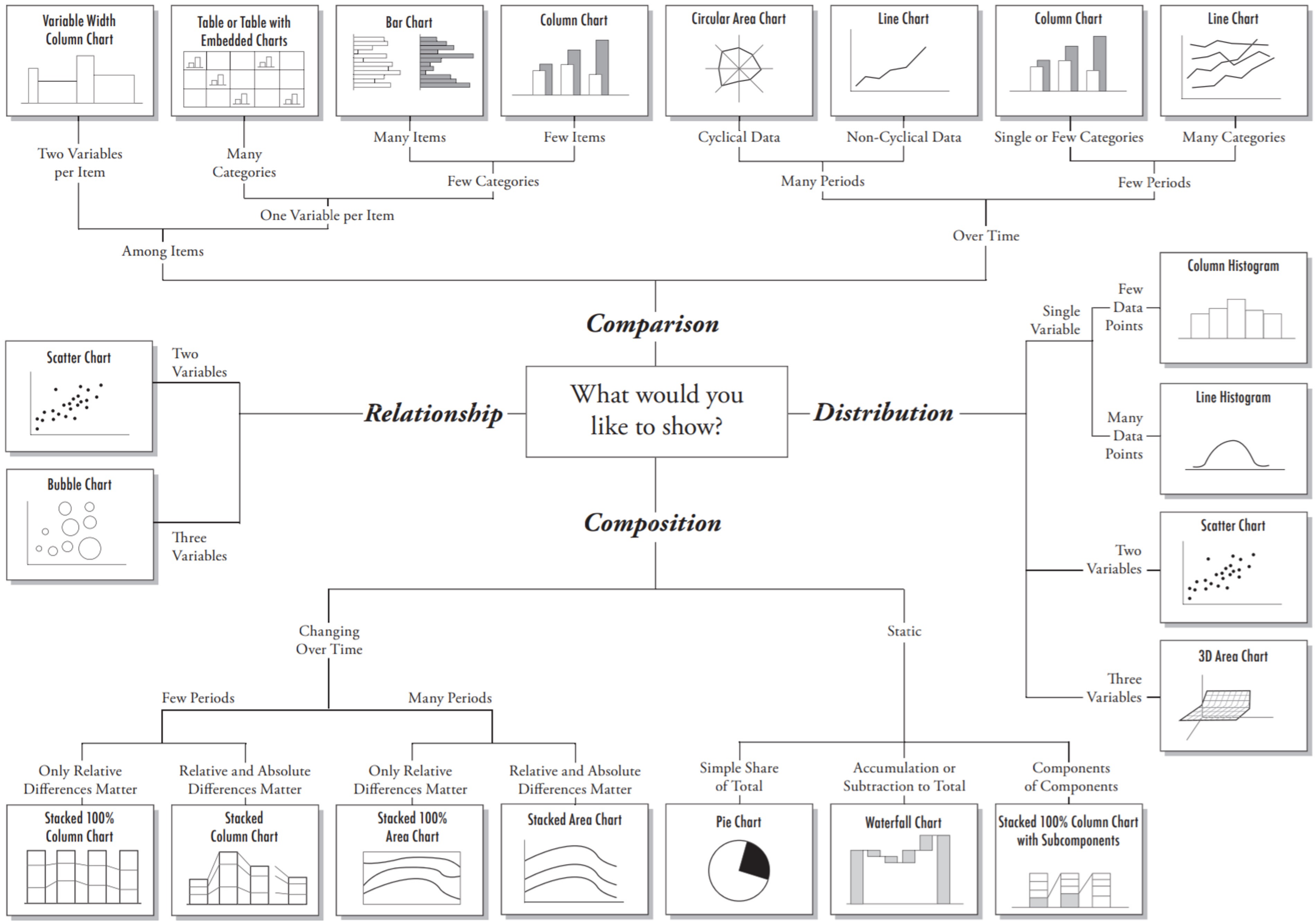
What do you look for in the visualizations from an EDA?

- Do the values shown make sense?
    - If not, which values don’t make sense– look into them! Don’t just delete.
- Does the number of values or points make sense?
- Are there any unusual or unexpected patterns in the data?
- Are there any patterns in the data at all?

Types of Plots

J

Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item — Variable Width Column Chart

One Variable per Item

Many Categories — Table or Table with Embedded Charts

Few Categories

Many Items — Bar Chart

Few Items — Column Chart

Over Time

Many Periods

Cyclical Data — Circular Area Chart

Non-Cyclical Data — Line Chart

Few Periods

Single or Few Categories — Column Chart

Many Categories — Line Chart

Relationship

Two Variables — Scatter Chart

Three Variables — Bubble Chart

Distribution

Single Variable

Few Data Points — Column Histogram

Many Data Points — Line Histogram

Two Variables — Scatter Chart

Three Variables — 3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter — Stacked 100% Column Chart

Relative and Absolute Differences Matter — Stacked Column Chart

Many Periods

Only Relative Differences Matter — Stacked 100% Area Chart

Relative and Absolute Differences Matter — Stacked Area Chart

Static

Simple Share of Total — Pie Chart

Accumulation or Subtraction to Total — Waterfall Chart

Components of Components — Stacked 100% Column Chart with Subcomponents

https://extremepresentation.com/wp-content/uploads/choosing-a-good-chart-09-1.pdf

Distribution
Comparison
Comparison over Time
Relationship
Composition

- We will review examples in Python first, then share R code

Conveying Information: Aesthetics

- **Position (x, y)** – usually the most accurate way to encode values
- **Length/size** – bar length, point size, area, line width
- **Orientation/angle** – less precise; used in some charts
- **Shape** – categorical differences in scatterplots
- **Texture/line style** – solid vs dashed, etc.
- **Area / Volume** – encode magnitude in a given region / shape
- **Color** – hue for categories; lightness/saturation for numeric magnitude

Communicating Time Series and Longitudinal Perspectives using Visualization

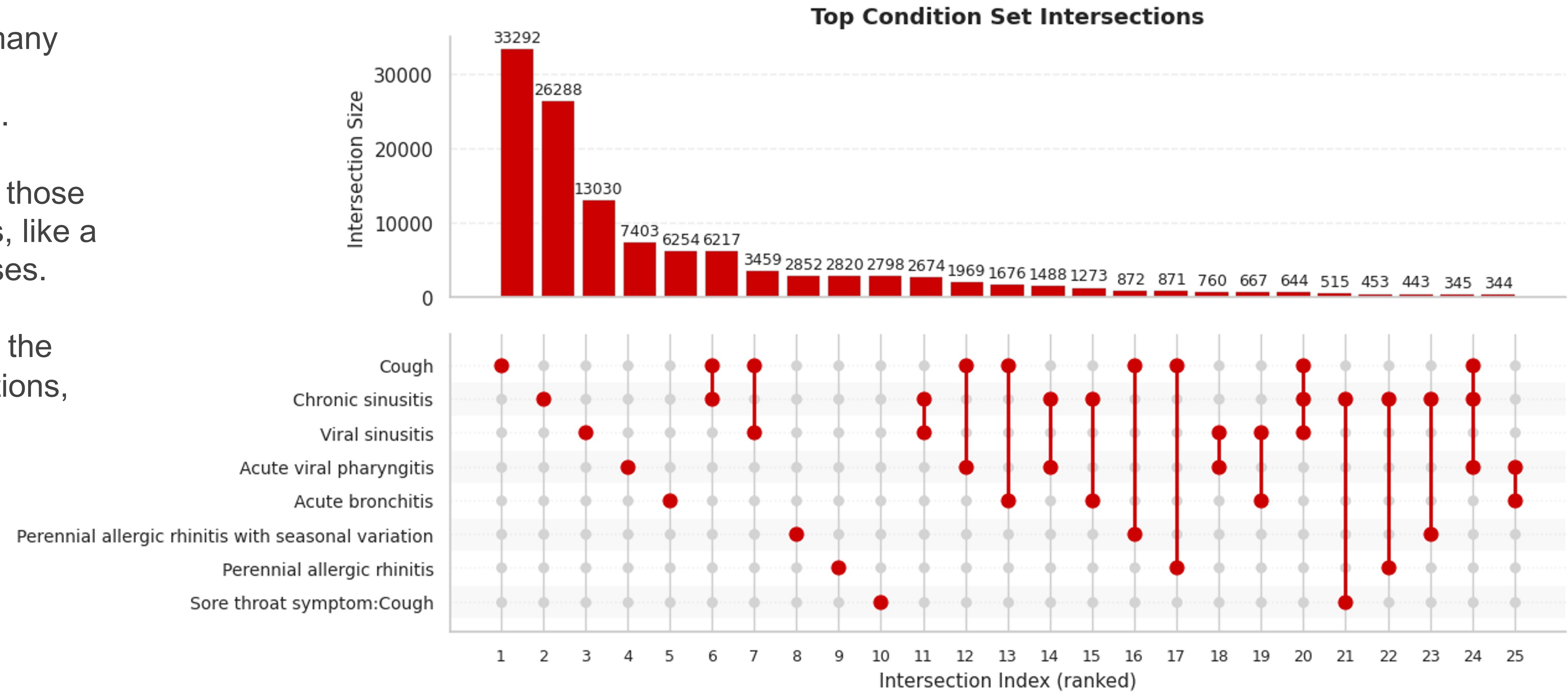
- Time ALWAYS goes on the x-axis
- Think carefully about what makes sense
- Keep in mind that data from one individual is not independent
  - Be sure to show any dependence that you have in your data!

We’ve provided a few examples in R and Python for you to try.

Finally, have fun with it, too!

Let the data tell its story visually.

- Each bar shows how many patients share certain condition combinations.
- Each dot line connects those co-occurring conditions, like a “friend group” of diseases.
- The bold red highlights the most frequent intersections, clear and confident.



Works Best with:

- Categorical datasets with multiple labels or attributes per record
- Overlapping groups or sets (e.g., co-occurring diseases, shared features)
- Medium-to-large samples where intersections reveal meaningful patterns

E

NC STATE UNIVERSITY

References and Suggested Resources

- *Communicating with Data: The Art of Writing for Data Science* by D Nolan and S Stoudt
- *Effective Data Storytelling: How to Drive Change with Data, Narrative, and Visuals* by Brent Dykes
- *Fundamentals of Data Visualization* by Claus Wilke
- *Storytelling with Data: A Data Visualization Guide for Business Professionals* by Cole Nussbaumer Knaflic
- *Good Charts: The HBR Guide to Making Smarter, More Persuasive Data Visualizations* by Scott Bernator
- *The Visual Display of Quantitative Information* by Edward Tufte
- *The Functional Art: An introduction to information graphics and visualization* by Alberto Cairo
- Data Visualization Catalog: https://datavizcatalogue.com/
- Flowing Data: https://flowingdata.com/
- Information is Beautiful: https://informationisbeautiful.net/
- For R: ggplot2 documentation: https://ggplot2.tidyverse.org/
- For Python: Python Graph Gallery: https://python-graph-gallery.com/

J,E

Please help us plan future workshops

- We’d appreciate your feedback
- Thanks to NC TraCS for supporting this work
- Thank you for attending!