Data Visualization

Understanding how to tell a story with your data the appropriate way

Rosie Martinez

Data Science Immersive

Learning Objectives

After today's lecture you will:

- Identify the key elements necessary on any graph
- Understand which types of visualizations are good for different data types
- Examine and identify what makes a bad graph
- Use the graphs to support the larger data story

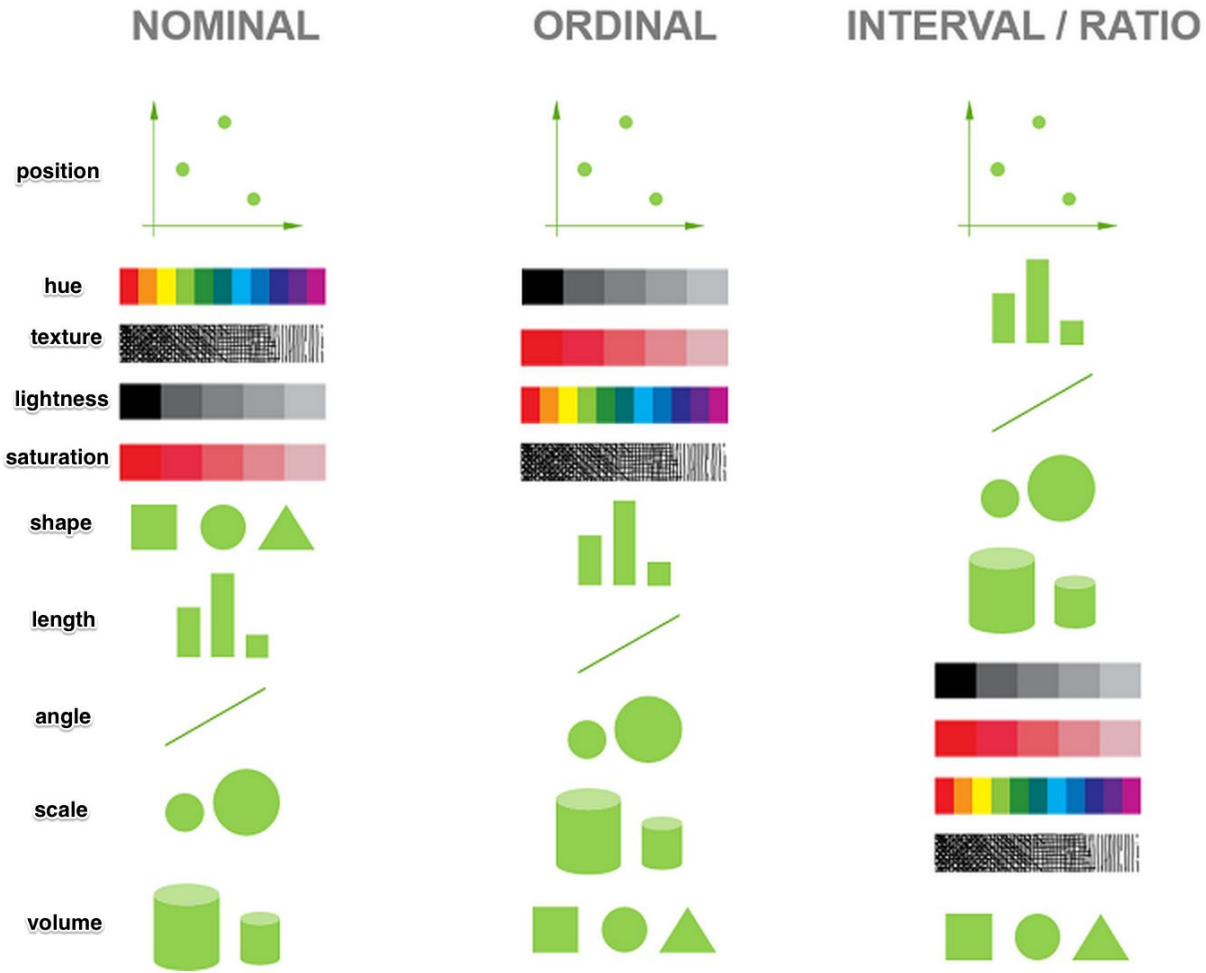
Overview

- Data viz is one of the most creative outlets
- Results need visualizations
- Arguably, the most important tool for communicating results
- Identifies patterns and summaries using graphs and tables

Types of Data

- Qualitative vs Quantitative
  - Qualitative: data is descriptive information (i.e. the bookcase is brown)
  - Quantitative: data is numerical information (i.e. the bookcase is 3 feet tall
- Discrete vs Continuous
  - Discrete: data can only take certain values
  - Continuous: data can take any value
- Other types:
  - Nominal: non-numeric categories (brand)
  - Ordinal: numeric data with non-constant or unknown spacing (t-shirt sizes)
  - Interval: numeric data with uniform spacing (date)
  - Ratio: interval data with a natural zero

Depending on the type of data you are working with, there is an order of "importance" that allows the reader to understand the graph.



Types of Graphs

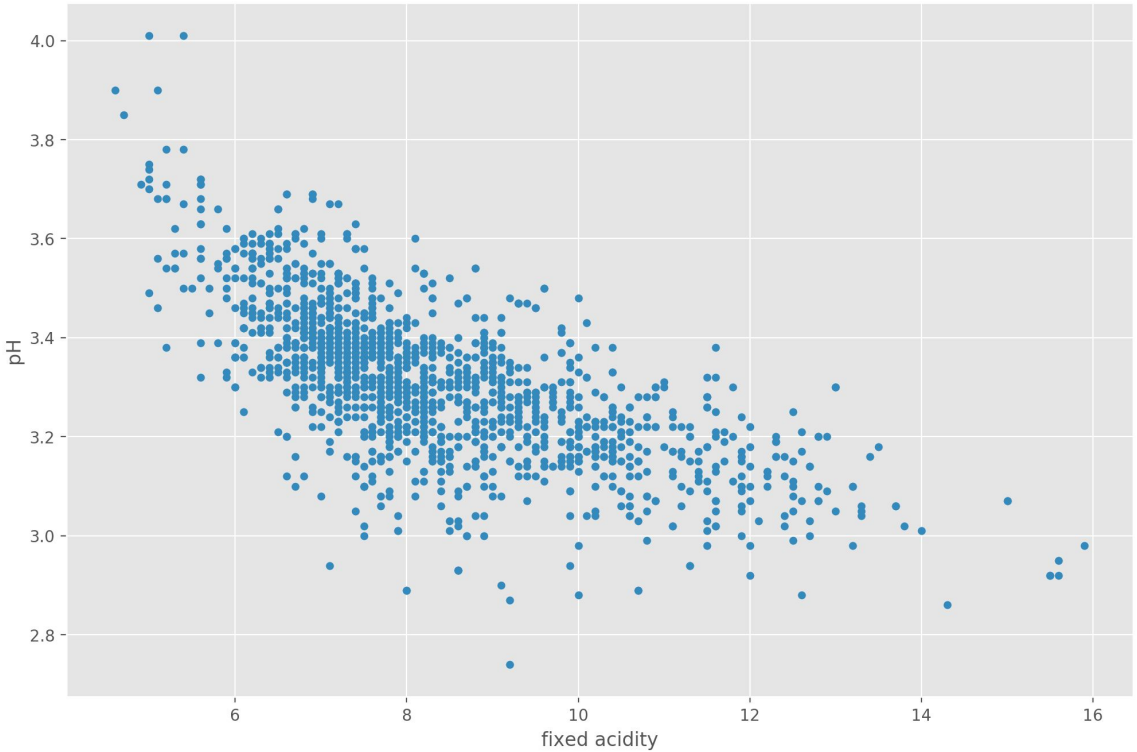
- Scatterplots
- Line Graphs
- Histograms
- Bar Charts
- Box Plots
- Pie Charts
- Heatmaps
- Pairplot
- Plus many more

Scatter Plots

Data Types: Continuous, Quantitative

Comparing an X variable to a Y variable

Used to observe relationships between variables



Line Plots

Data Types: Continuous, Quantitative, Interval, Ordinal

Constructed of lines connecting points called “markers”

X-axis is ordinal or Interval in nature, like a time series

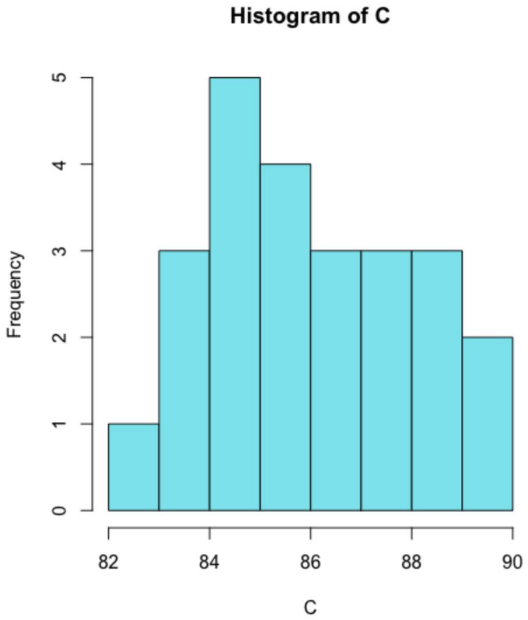
Histograms

$$C = \begin{bmatrix} 89.3 & 84.5 & 85.5 & 83.2 & 86.6 & 88.8 & 84.4 & 90.0 \\ 88.5 & 87.0 & 88.3 & 84.2 & 85.6 & 87.9 & 88.0 & 84.7 \\ 83.2 & 82.2 & 85.9 & 86.3 & 86.5 & 85.5 & 83.9 & 87.8 \end{bmatrix}$$

Data Types: Continuous, Quantitative

Creates “bins” to separate the data, convention says that each ‘bin’ is left-inclusive, right-exclusive

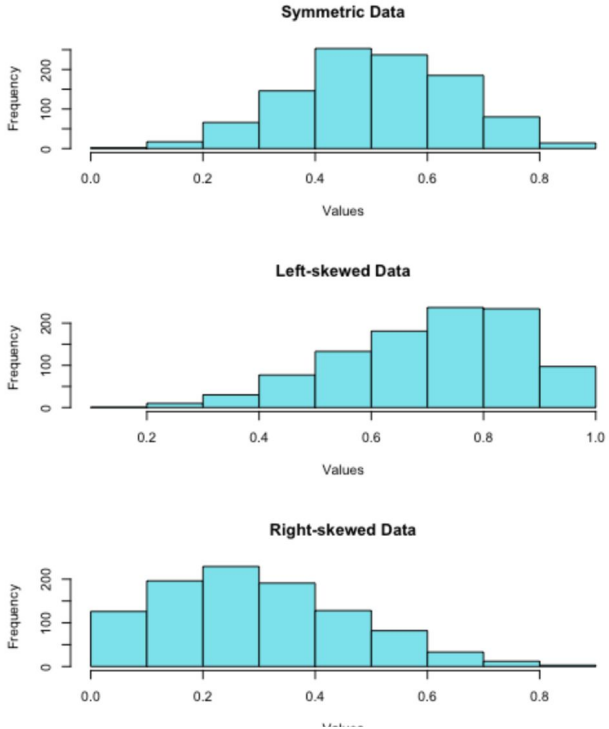
Can show the overall distribution of the data



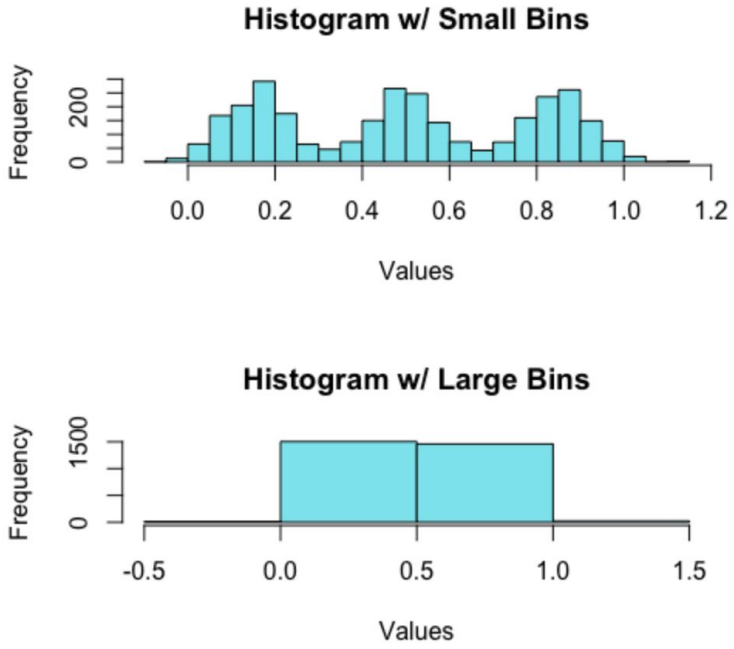
Histograms

Distribution of a dataset:

- Uni-modal
- Symmetry
  - Approximately “centered”
- Left Skewed/Negative skew
  - Tail goes to the left
- Right Skewed/positive skew
  - Tail goes to the right



Histograms



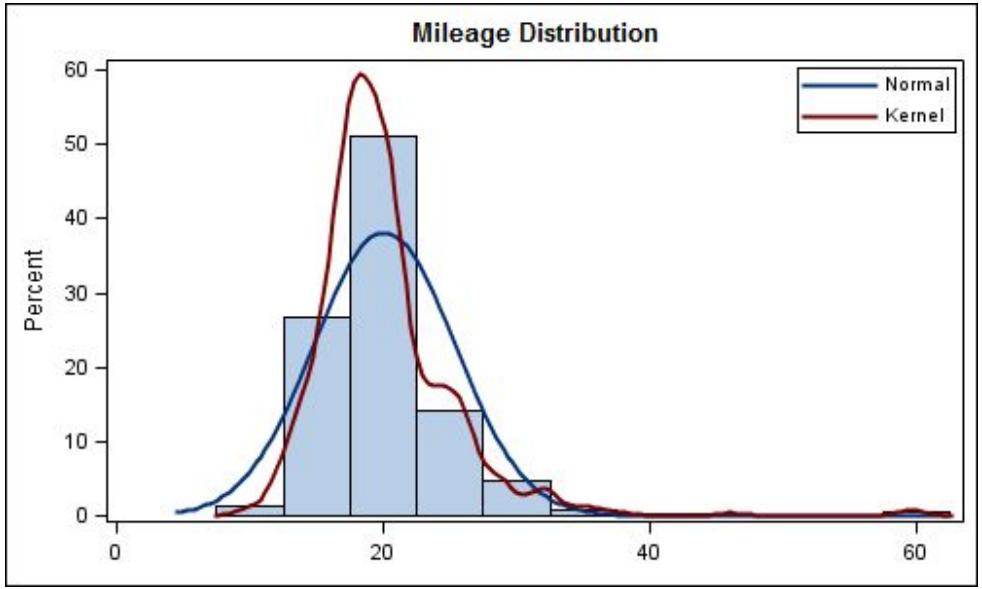
How to determine how many bins?

A common function to calculate the number of bins is:

$$k = \text{sqrt}(n)$$

Where k is the number of bins and n is the sample size

Kernel Density Plot



Data Types: Continuous, Quantitative

Can show the overall distribution of the data on a continuous interval allowing for smoother distributions by smoothing the noise

Source

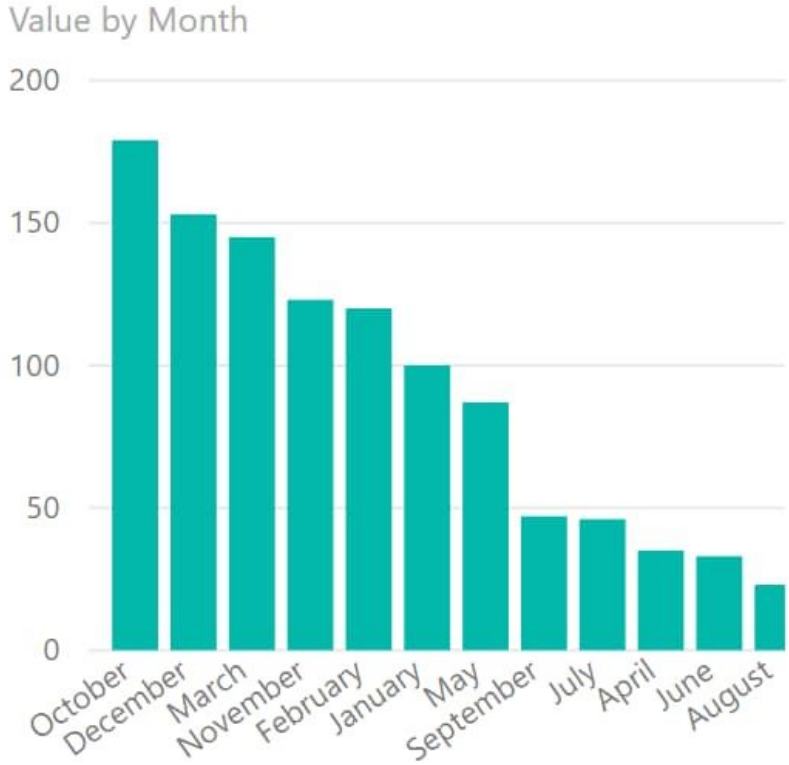
Bar Chart

Data Types: all the data types

Used to represent the same variable over a number of domains

Can show frequency distributions for discrete variables

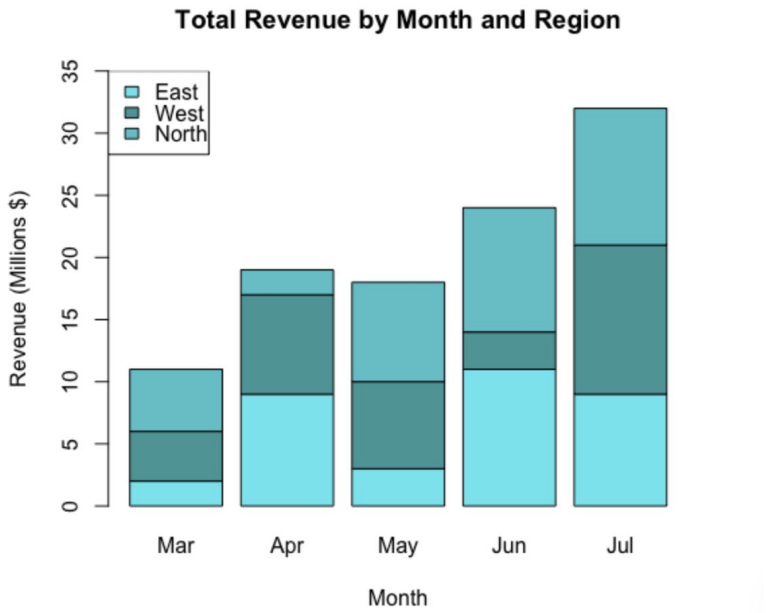
Should be sorted in order if x-labels are not ordinal

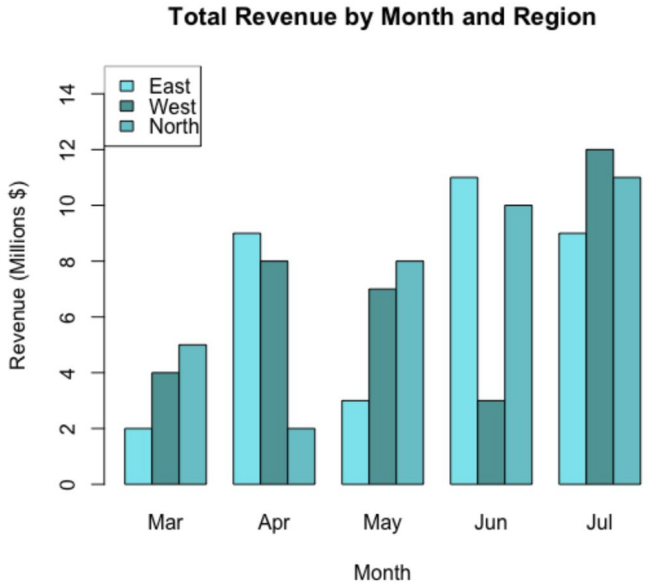


Source

Bar Chart

Can be used to display a subset of data multiple ways





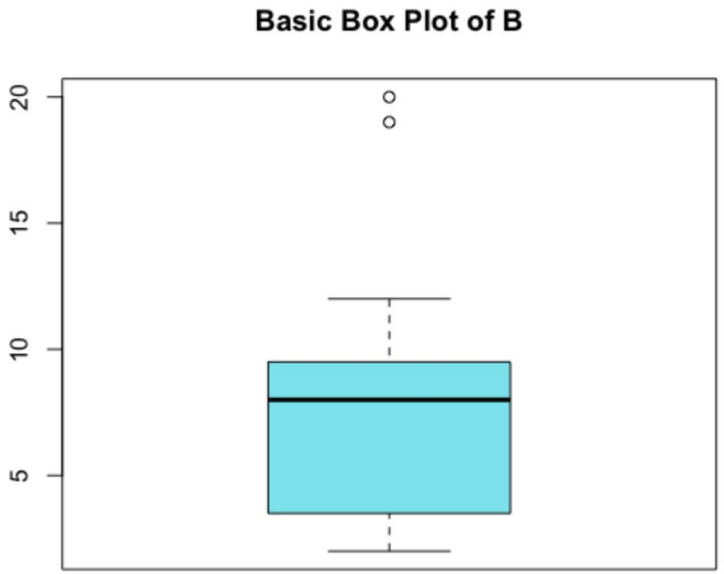
Box Plot

Data Types: Continuous, Quantitative

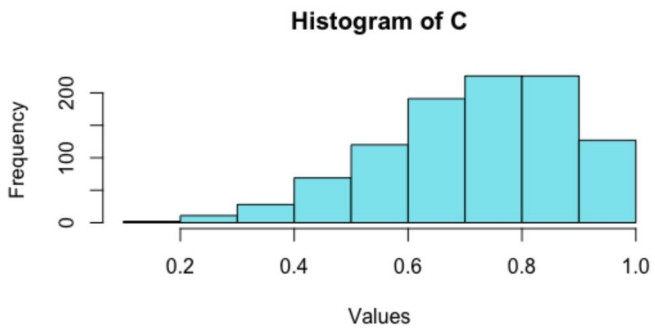
Used to visually represent the five number summary

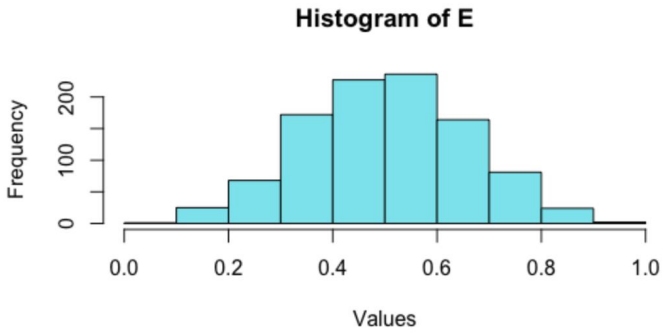
Can show the distribution skew

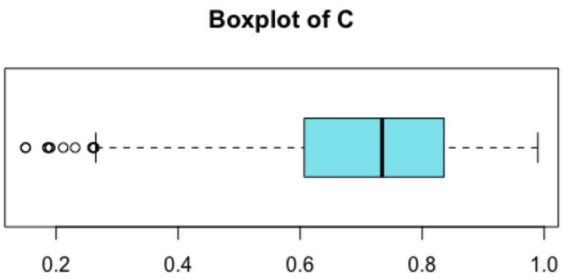
Can visually represent the outliers

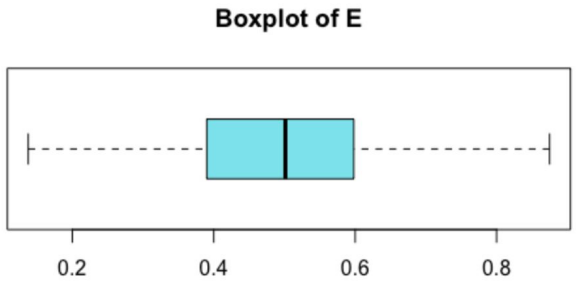


Box Plot


Histogram of C
Frequency
0
100
200
0.2
0.4
0.6
0.8
1.0
Values


Histogram of E
Frequency
0
100
200
0.0
0.2
0.4
0.6
0.8
1.0
Values


Boxplot of C
0.2
0.4
0.6
0.8
1.0

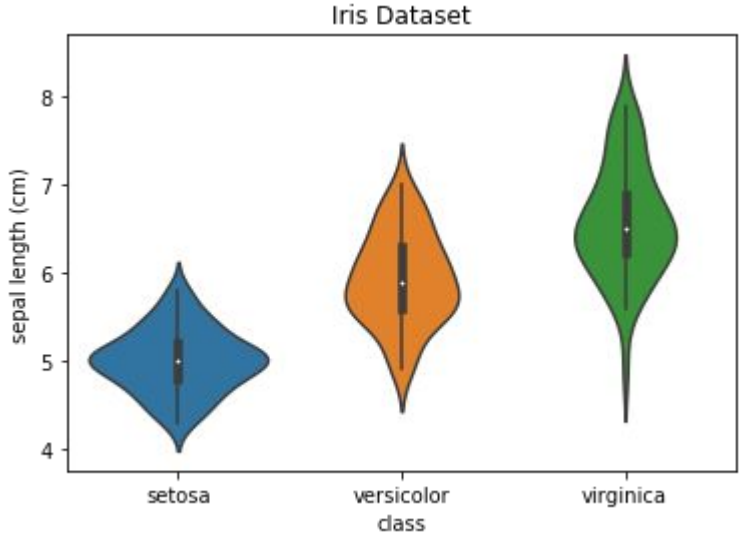

Boxplot of E
0.2
0.4
0.6
0.8

Box Plots to Violin Plots

Data Types: Continuous, Quantitative

Similar to a box plot, but adds in the kernel density plot in each side (distribution shape)

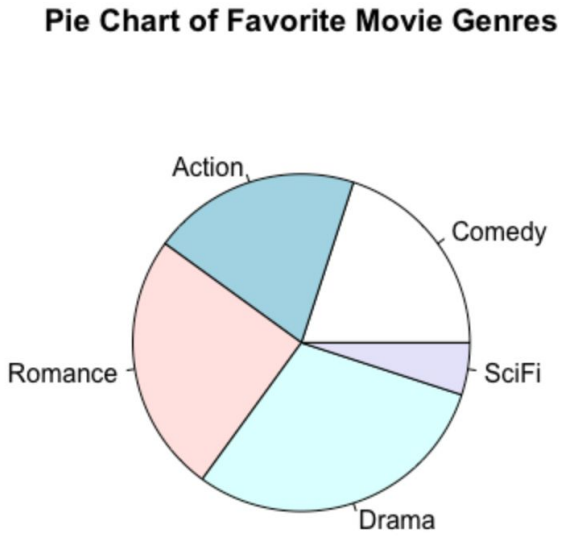
It shows the summary statistics in the plot as well



Pie Charts

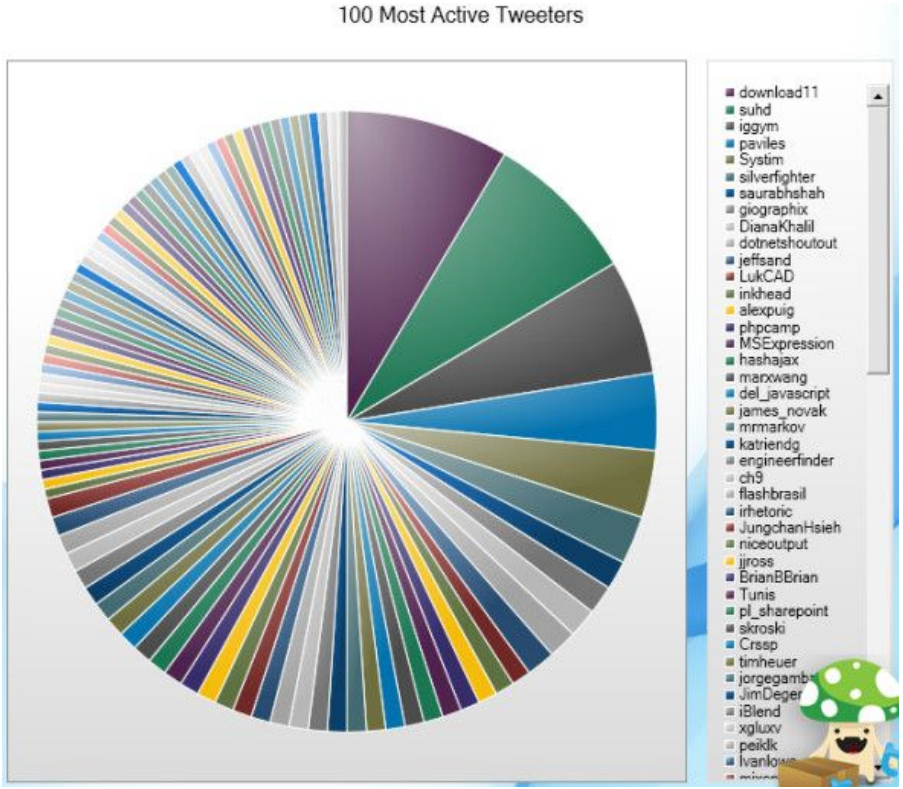
Data Types: All Types of data

Used to show proportions in data, like bar charts



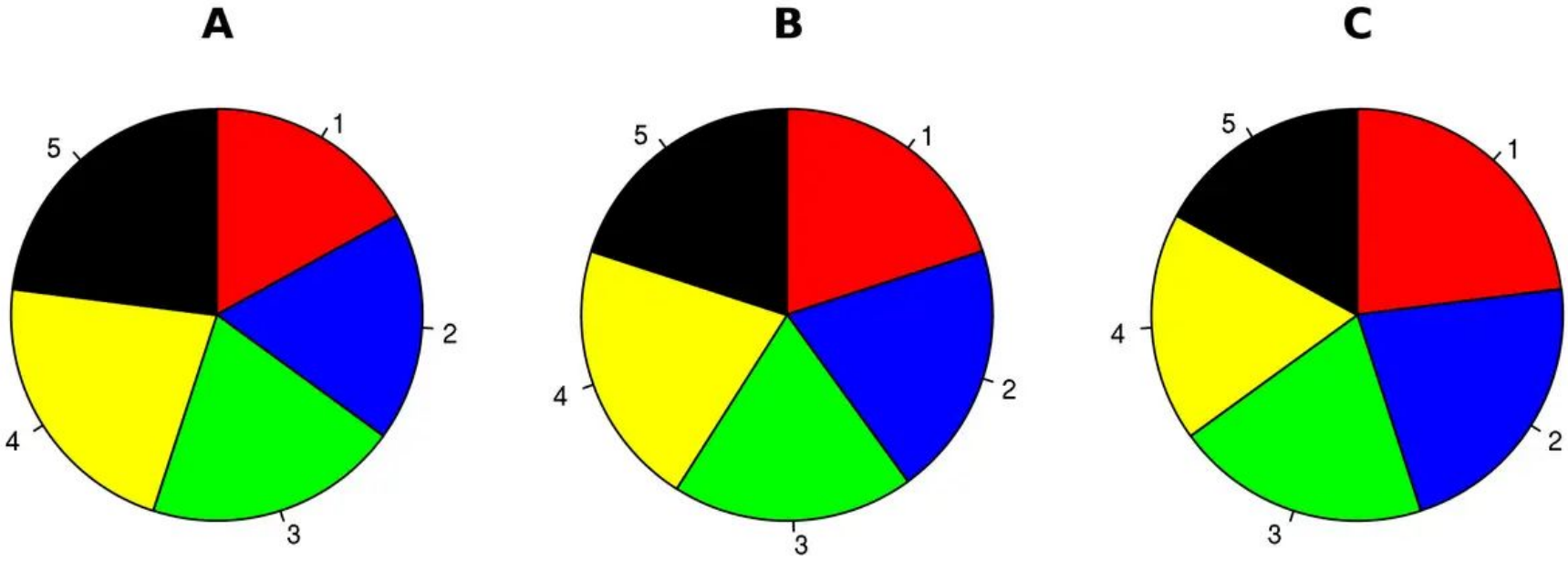
Pie Charts

- Reasons why Pie Charts can be ineffective:
  - Difficult to compare one subset to another
  - Difficult to compare the subsets of two Pie Charts
  - They show proportions well, but that doesn't necessarily translate to creating understanding about the total value represented, or the specific value of any subset
- Reasons why Pie Charts can be effective:
  - Approachable to non-technical audience
  - Visually appealing (Donut Charts rising in popularity)
- Limitations/Alternatives
  - Limit the number of subsets to 6 or fewer
  - Consider a Bar Chart as an alternative

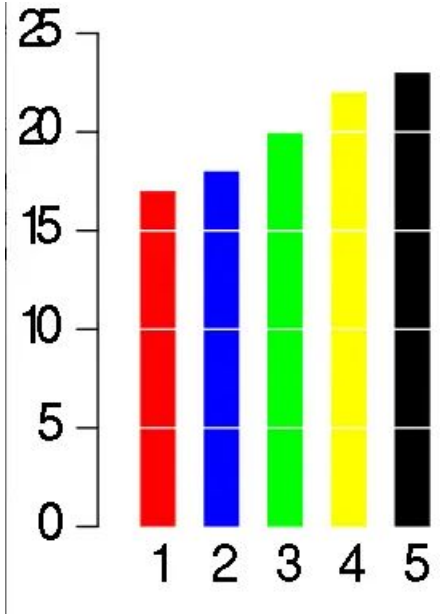


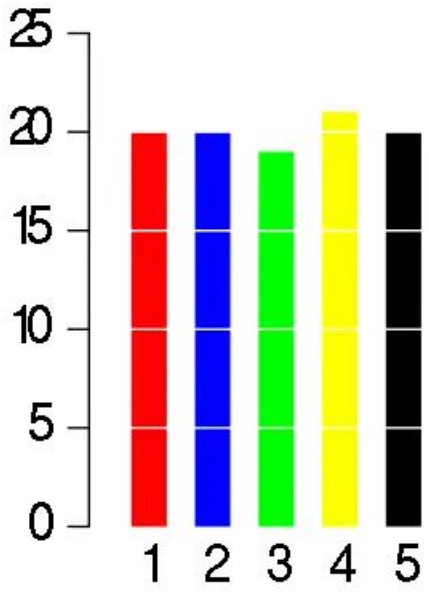
Source

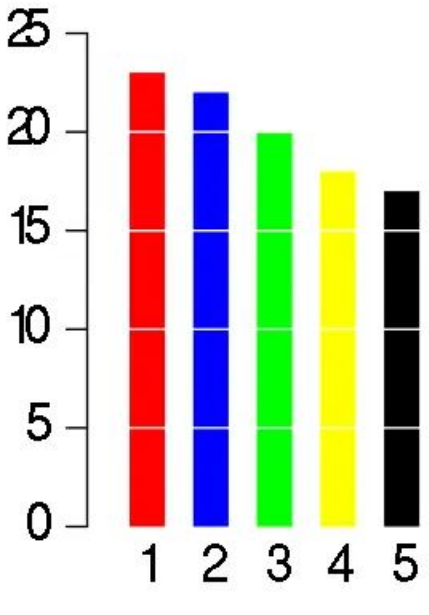
Pie Charts



Pie Charts





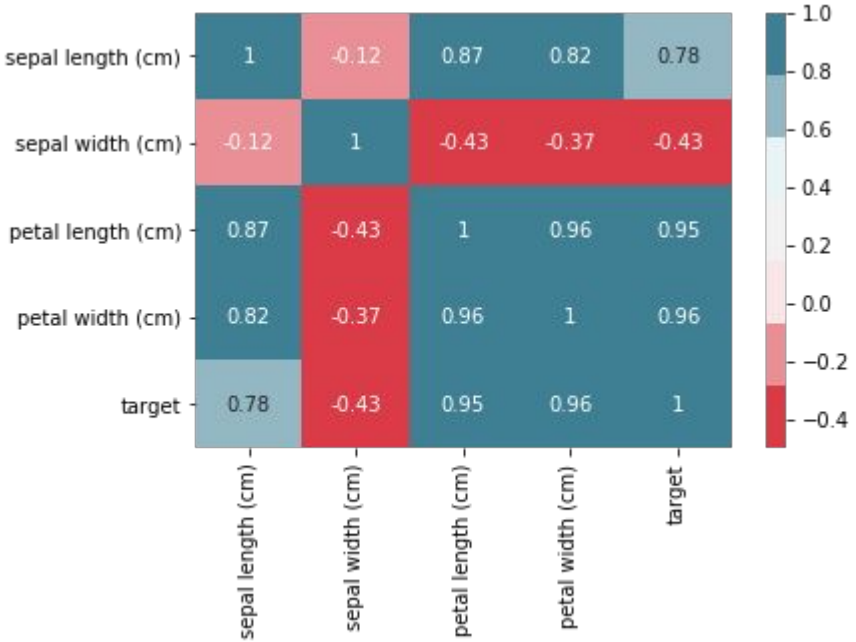


Heatmaps

Visually represent a matrix

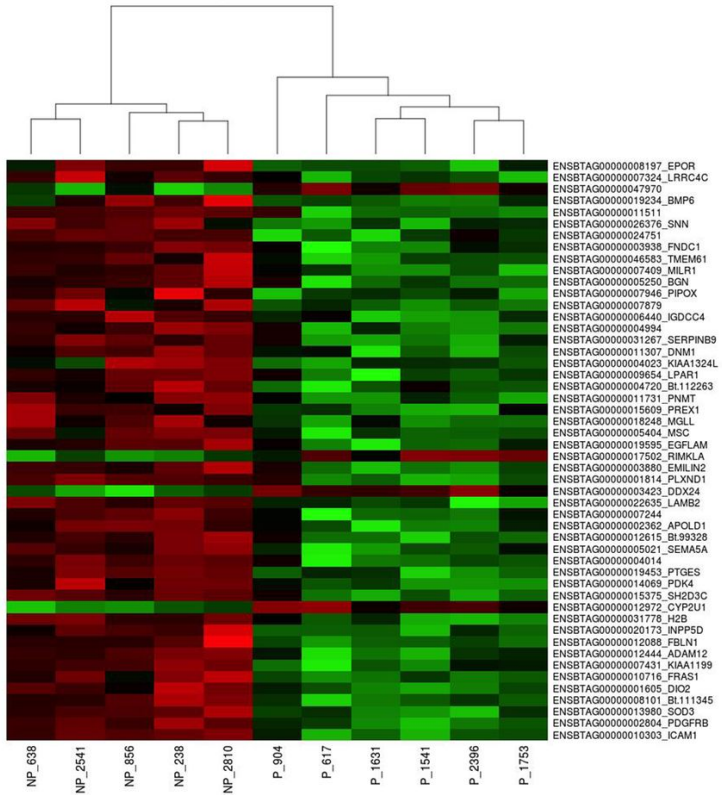
Colors represent the numbers of elements of the matrix

Typically used to show covariance and correlation



Heatmaps

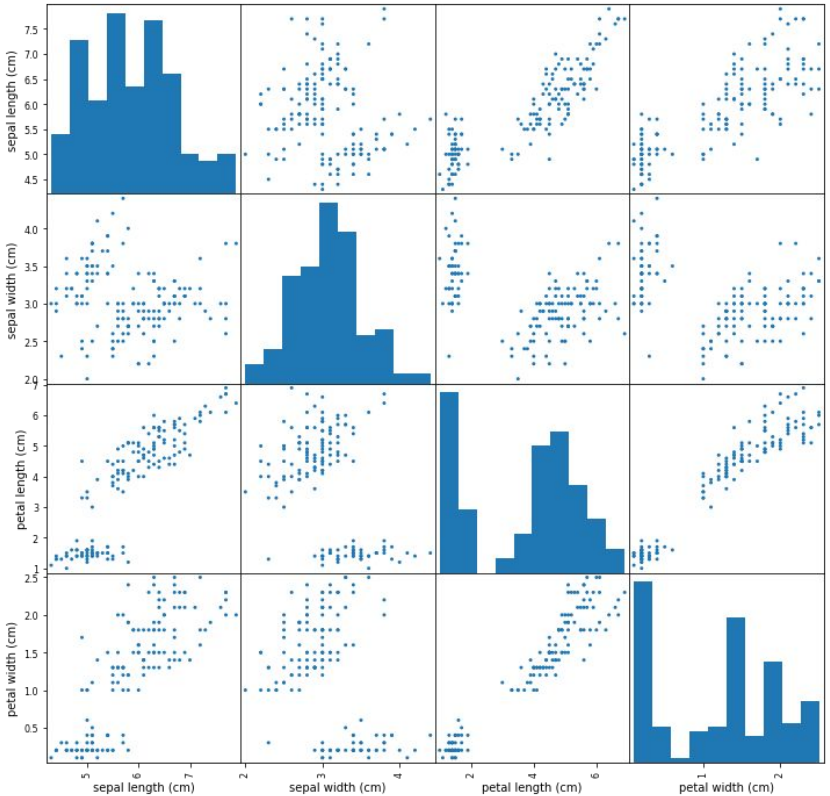
Commonly used in gene expression arrays



Pairplots

Shows pairwise data comparisons

Be wary of using pairplots with large datasets



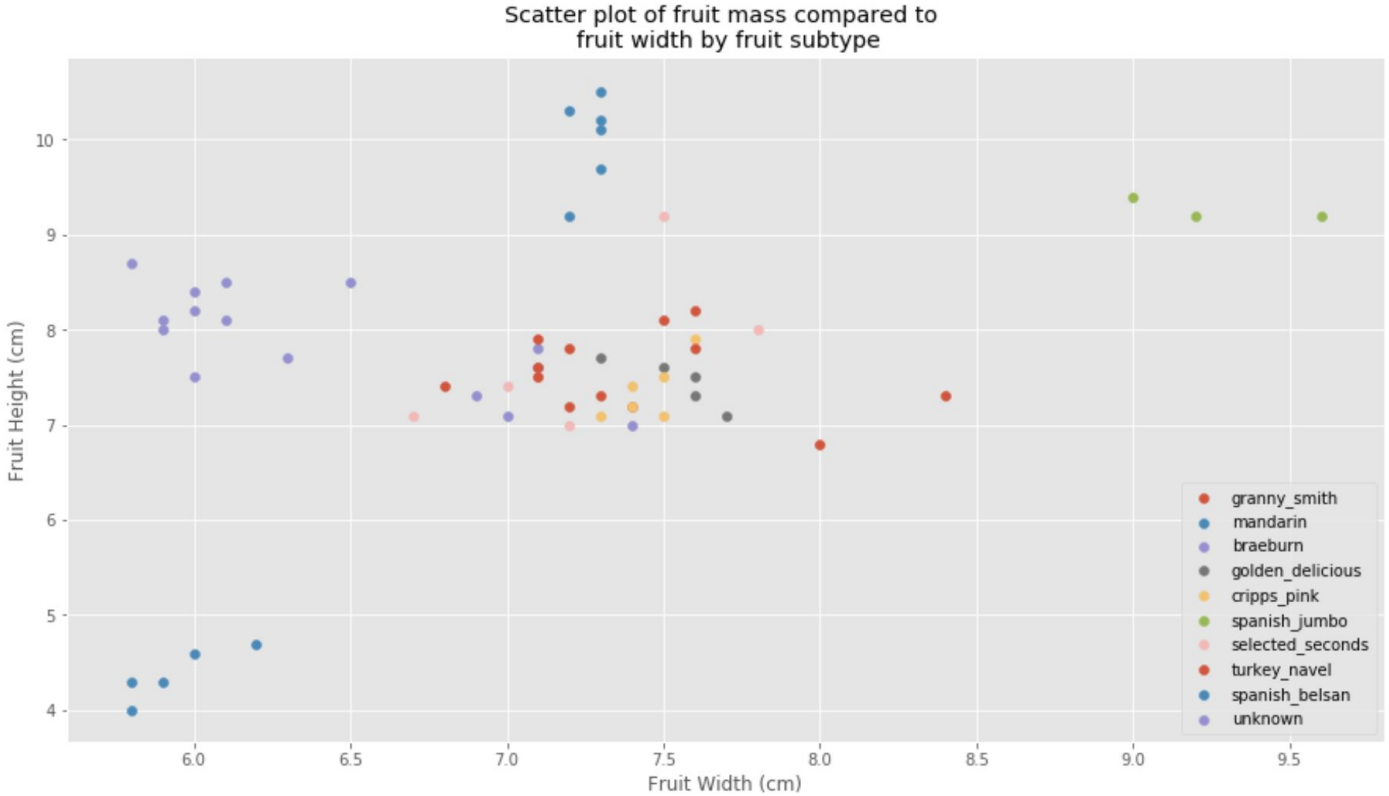
Helpful Tips for making graphs

- Check your data
  - What are the takeaways from the visual you want to create?
  - Do you have the appropriate data to send the message you want?
- Explain your encodings
  - Legends and labels!
- Label Axes and Units
- Keep your Geometry in Check
  - If you are using shapes to represent proportions, make sure they are correct
- Include your sources if they aren’t your own data
- Consider your audience

What makes a good graph

- Visual Structure
  - Make sure to represent the data with the correct graph
  - Consider number of axes and your background
  - Scale and tick marks -
    - All scales should begin with the zero value
    - Tick marks only use for quantitative scales
  - Grid Lines
  - Text

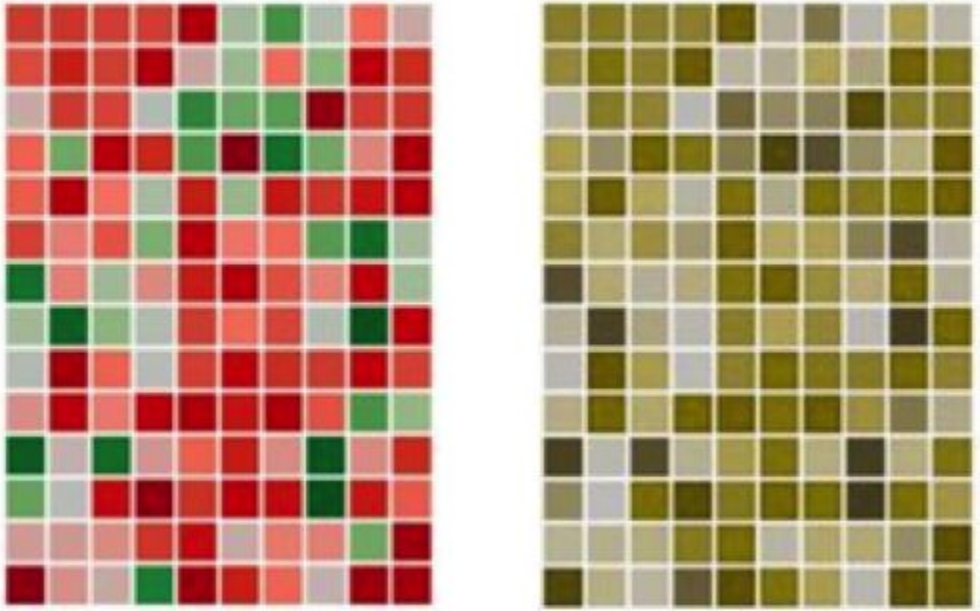
What makes a good graph



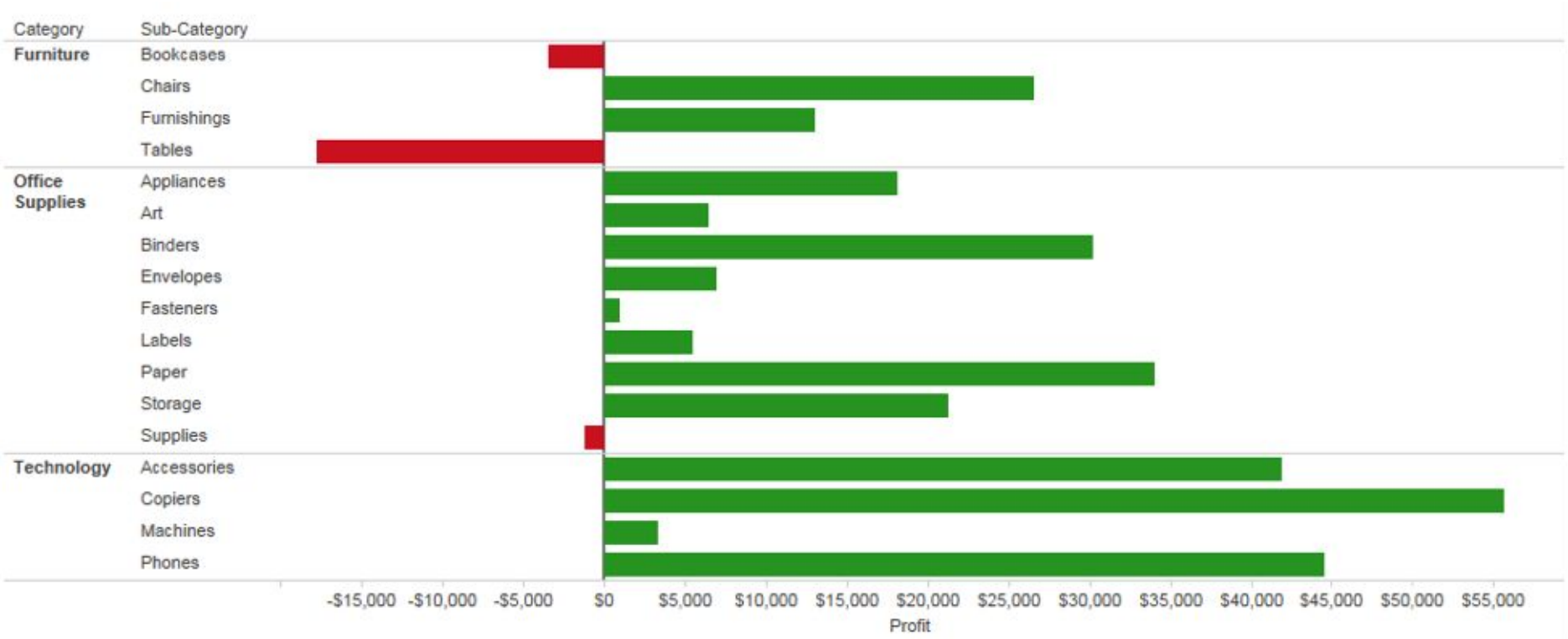
What makes a good graph

- Visual Structure
  - Make sure to represent the data with the correct graph
  - Consider number of axes and your background
  - Scale and tick marks -
    - All scales should begin with the zero value
    - Tick marks only use for quantitative scales
  - Grid Lines
  - Text
- Color Considerations
  - Color Blindness is something to consider
    - Red and Green together can be problematic in some circumstances
    - Try to use a colorblind friendly palette

What makes a good graph



What makes a good graph



What makes a good graph

- Overall:
  - Keep it simple and clean
    - What is the point you are trying to get across?
  - Make it readable
    - Make sure your font sizes on your axes, ticks, titles, and legends are legible
  - Use the correct graph for your data
  - Colors and symbols matter just as much in a good visualization as the data you are using
  - Remember your audience
    - Your graph might make sense to another data scientist, but what about a CFO, your grandmother, or someone who doesn't know the first thing about data?
  - A graph is only a graph until you provide context - we will get to storytelling in just a bit

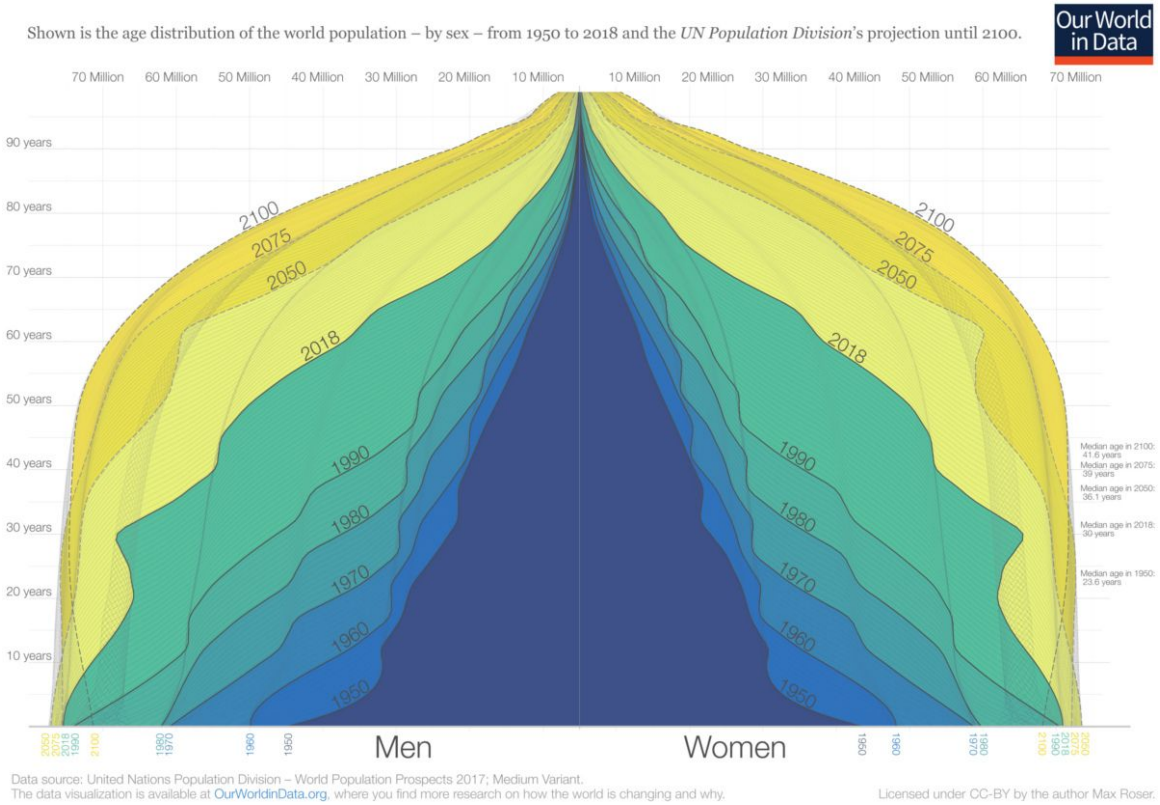
Good, Bad, or Terrible?

- We are now going to have an interactive discussion over the next few graphs. Please identify if it is:
- A good graph
  - What makes it good
- A bad graph
  - What makes it bad?
  - Is there anything you would change to make it a better graph?
- A terrible graph
  - There is no hope for this graph

Good, Bad, or Terrible?

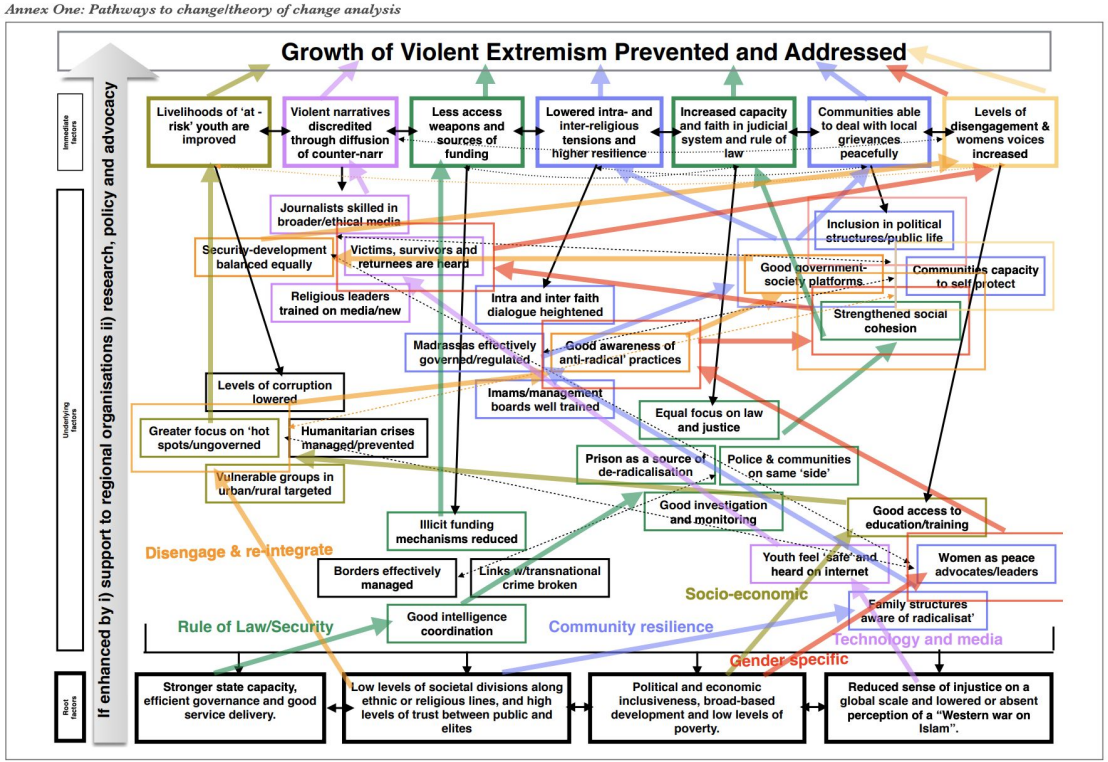
Source

Good, Bad, or Terrible?



Source

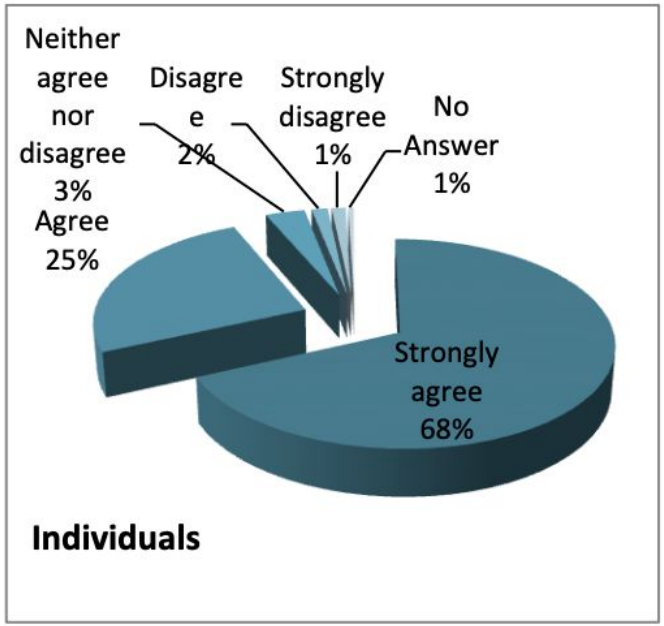
Good, Bad, or Terrible?

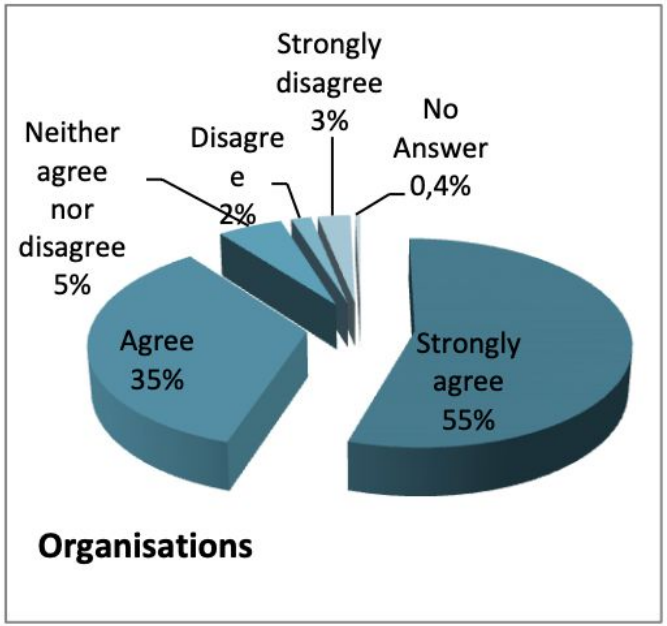


Source

Good, Bad, or Terrible?

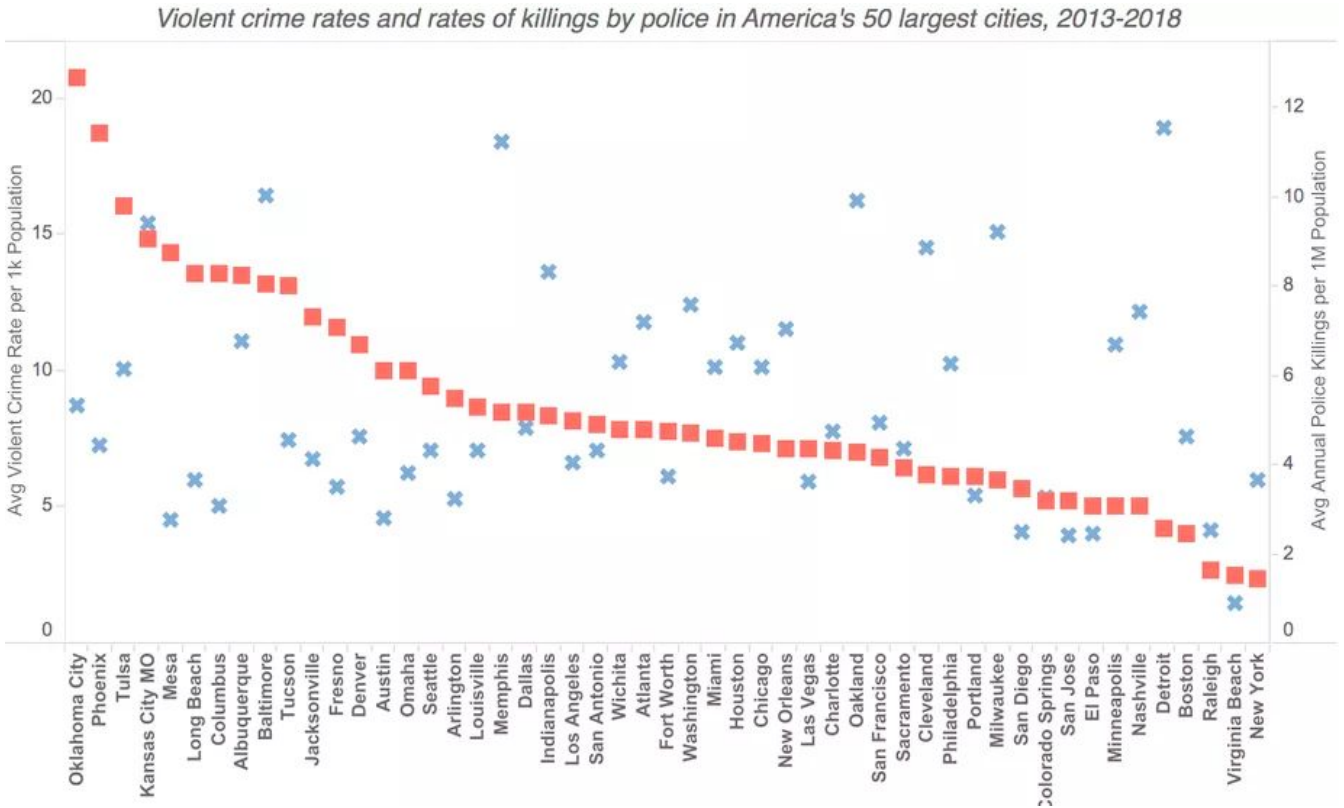
Figure 3 - Question 29. Regarding the statement "Citizens should be able to manage their own health data", do you…





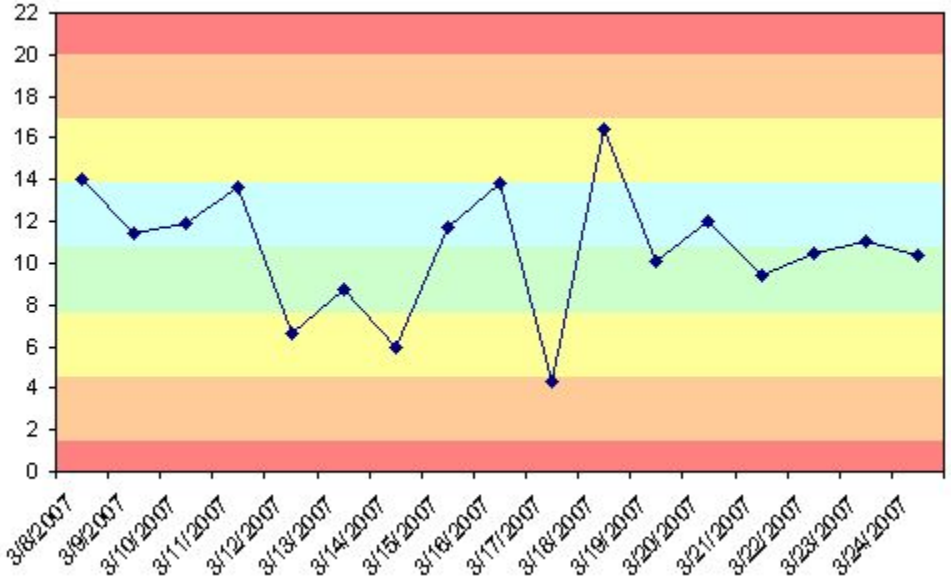
Source

Good, Bad, or Terrible?



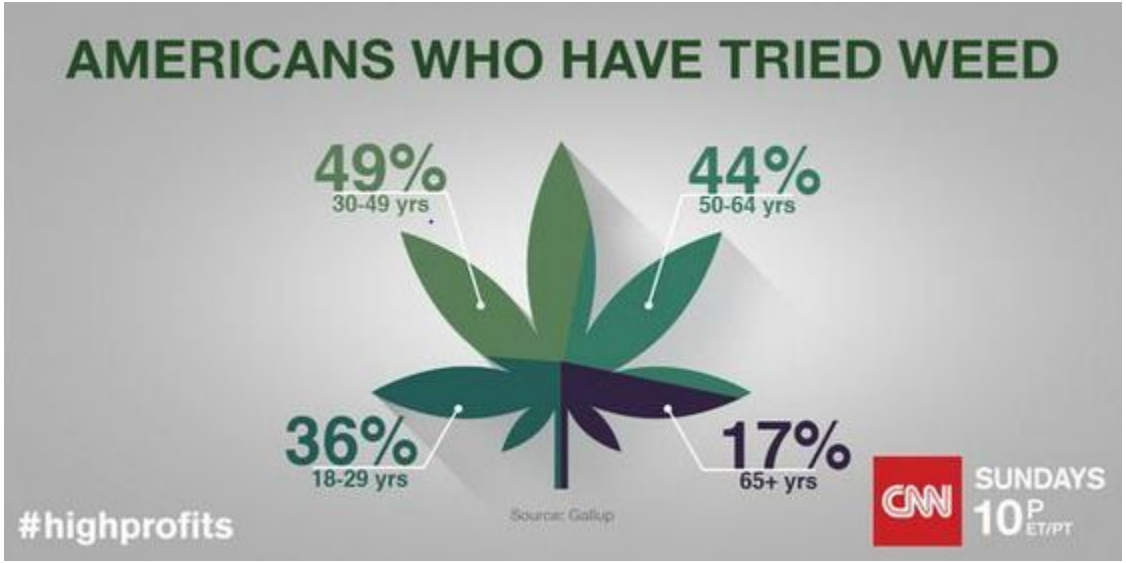
Source

Good, Bad, or Terrible?



Source

Good, Bad, or Terrible?



Source

Data Storytelling

Data storytelling is changing the way businesses make decisions

There are four elements that make a good data story:

1. The Data
2. The Visuals
3. The Narrative
4. The Audience

It is important to understand how these different elements combine and work together in data storytelling. Confusing graphs and visuals can make data much more difficult to understand

Exploratory vs Explanatory

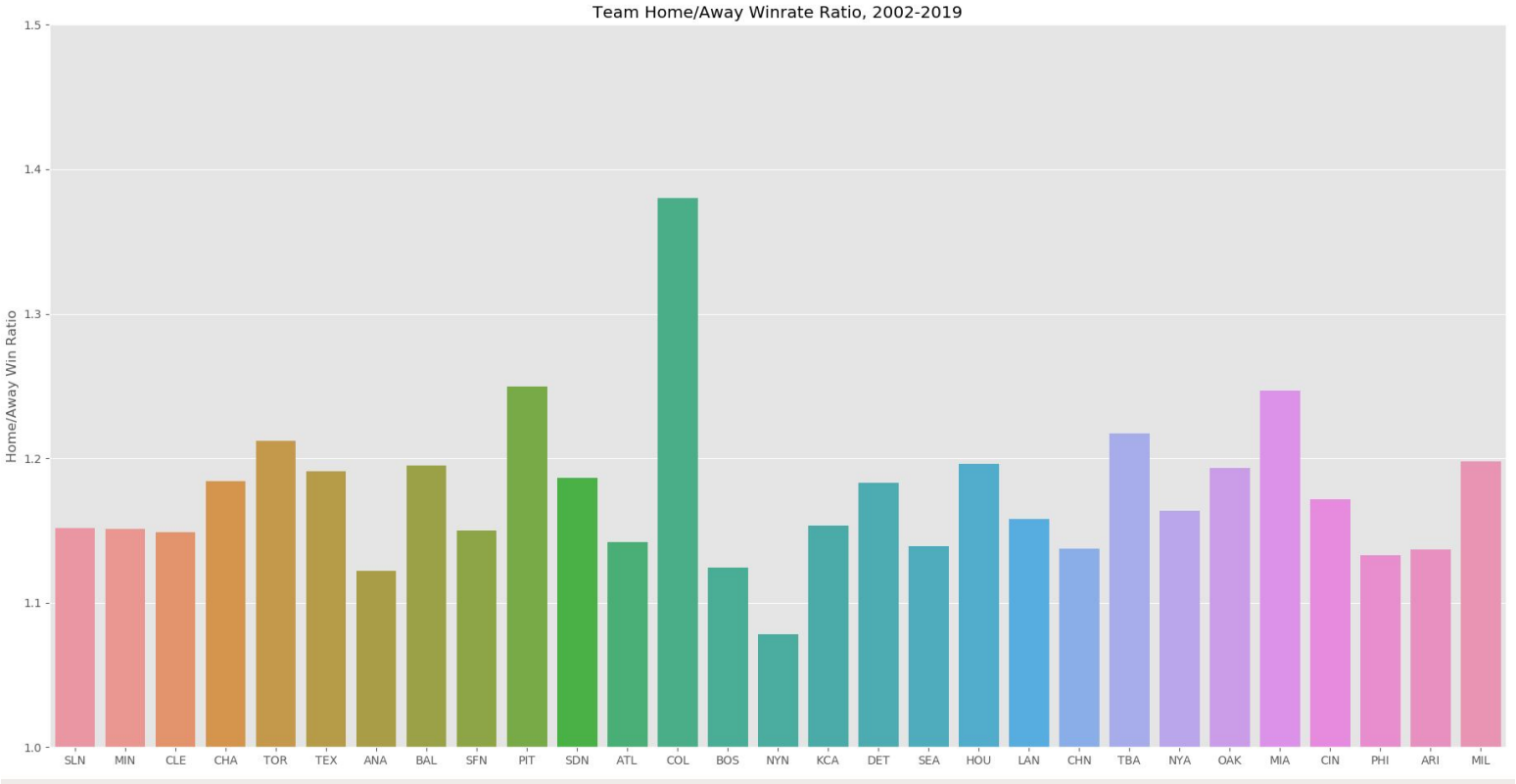
Exploratory:

- Understanding data without any expected agenda or narrative
- Part of data understanding

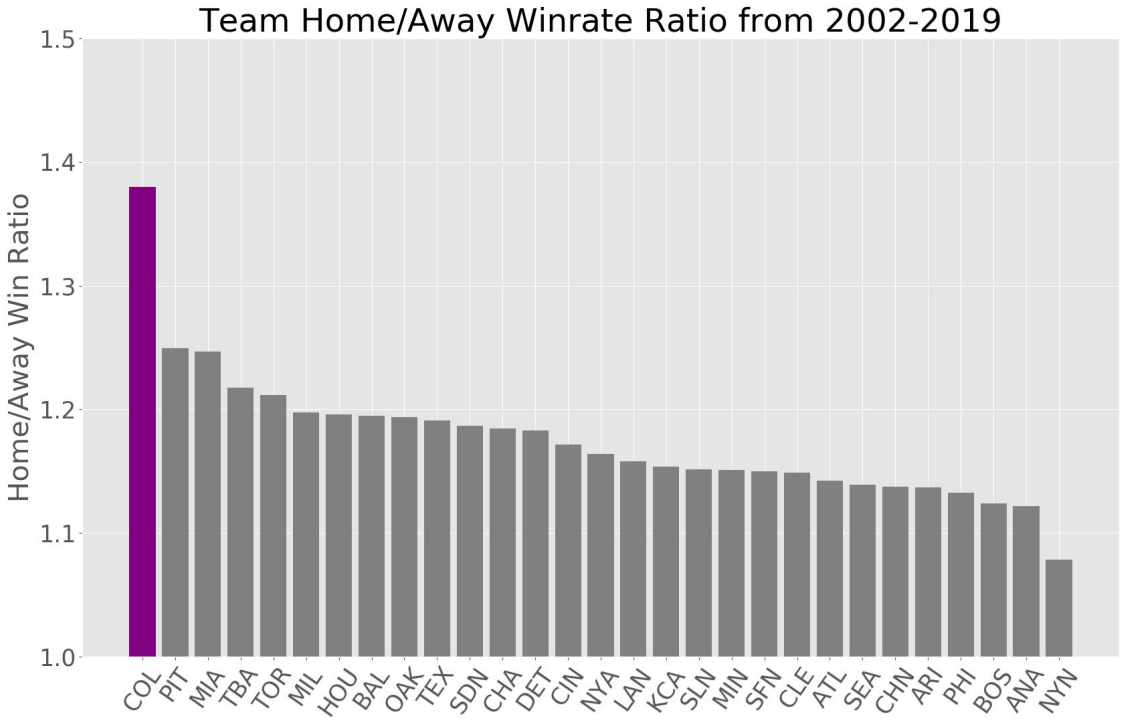
Explanatory:

- Using data viz to editorialize and communicate a narrative
- Part of deployment
- Can be author-driven or viewer-driven

Exploratory Example



Explanatory Example



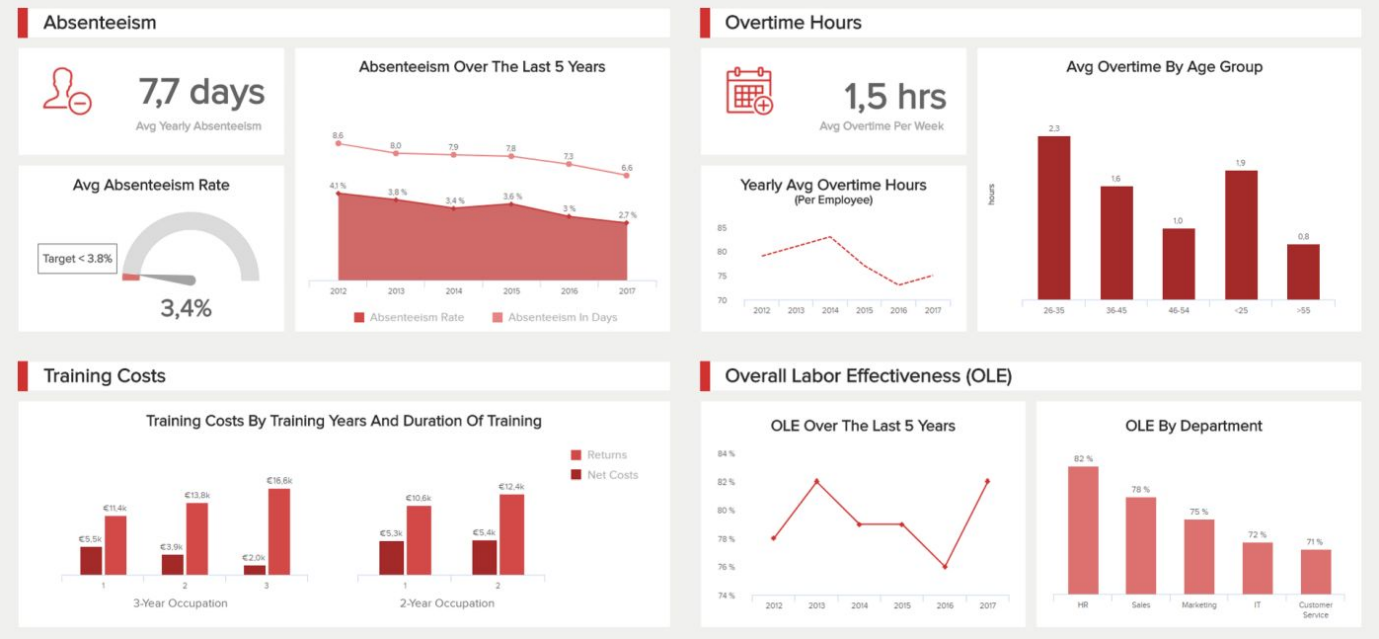
The 8 Commandments of Data Storytelling

1. Begin with a question: Set up your story. What is your audience going to learn?
2. End with an insight: If we can't learn something useful from the data, the story isn't worth telling.
3. Tell a compelling story: People remember stories, not data. Take them on your journey.
4. Explain with visuals, Narrate with words: People understand metrics, trends, and patterns better with visuals. Use words to add your voice to the data.
5. Be honest and credible: The clients we want value honesty. Don't sugarcoat the negatives. And don't mislead with fractioned data.
6. Be clear and concise: Remove everything that is not part of your story. Save the other bits for another time.
7. Know and cater to your audience: What are their interests and goals? Do they want the details, or just the high-level summary?
8. Provide context: Compare metrics over time or to industry benchmarks. Numbers are meaningless without context.

Source

How to tell stories with your data?

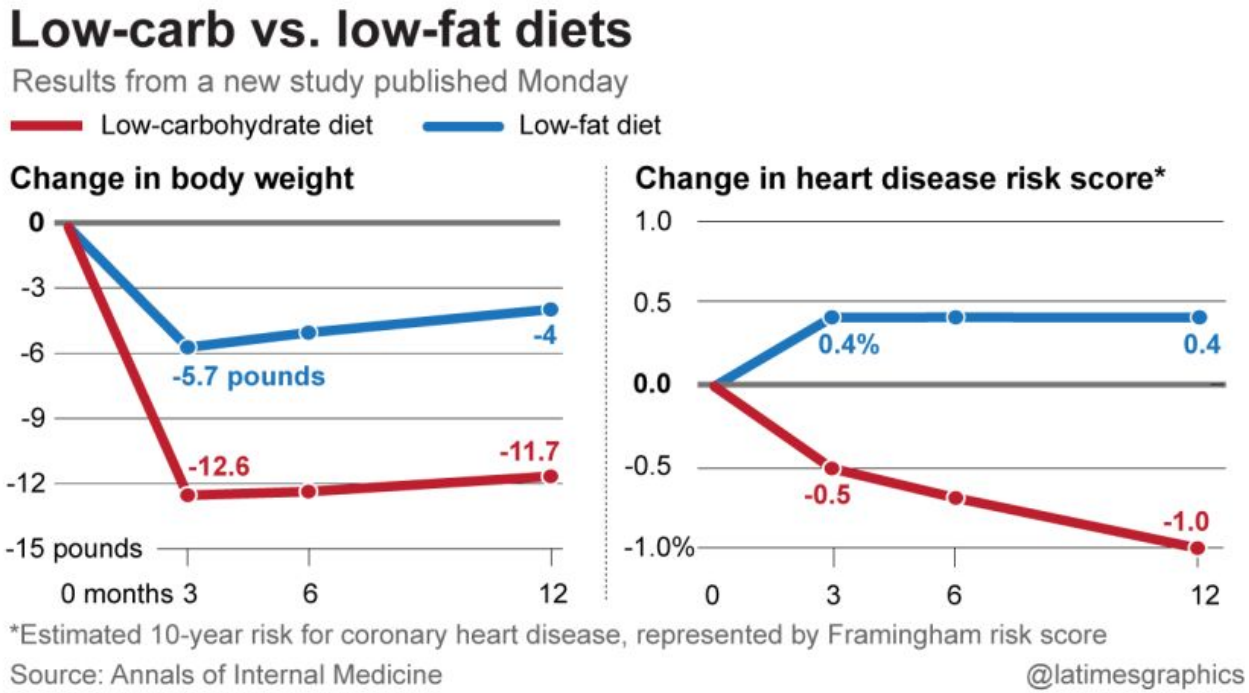
- Turn metrics into actionable concepts



Source

How to tell stories with your data?

- Improve processes and identify interventions with plotting



Source

How to tell stories with your data?

- Simplify & make connections (this is what you will be doing for most of this course)

Thought Exercise:

Let's say your boss comes up to you with all of the NYC MTA (subway) data and wants to make the MTA more efficient.

Your data includes: Turnstyle logs, arrival times, departure times, time delayed from arrival, time delayed from departure, station name, and some other

What would be some visuals that you think would be helpful to drive MTA decisions and make it more efficient?

Learning Objectives

After today's lecture you will:

- Identify the key elements necessary on any graph
- Understand which types of visualizations are good for different data types
- Examine and identify what makes a bad graph
- Use the graphs to support the larger data story