GENERAL ASSEMBLY

Principles of Data Visualization

Patrick D. Smith

Lead Instructor, General Assembly DSI

LEARNING OBJECTIVES

- Learn how to visualize your data with Tableau
- Lean how to utilize Python’s plotting libraries to visualize your data

Opening

Principles of Data Visualization

Visualizing your data is extremely important to be proficient at as a data scientist. Why?

1. **You need to be able to explore your data visually. This is essential!**

- You need to get an intuition for your data.
- Visualization should always be done before you start modeling your data.
- If you model data without visualizing it first, you are asking to run into problems down the line!

Principles of Data Visualization

Visualizing your data is extremely important to be proficient at as a data scientist. Why?

2. You will be always be required to report on your findings working as a data scientist.

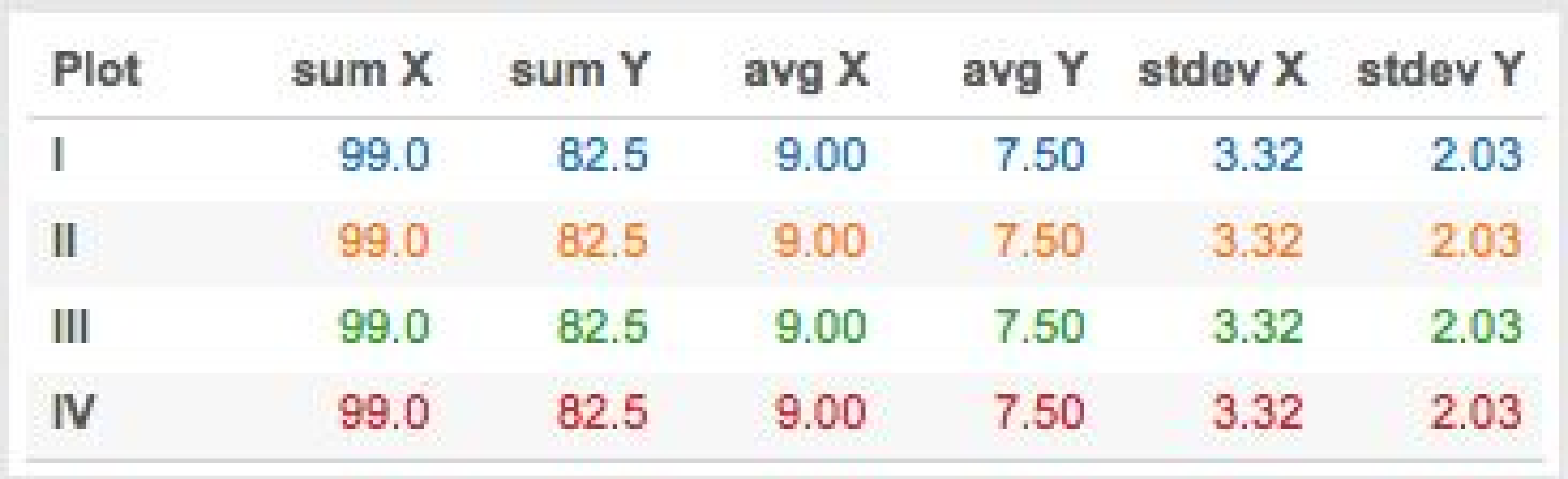
- Technical coworkers such as other data scientists or analysts will want to get an intuition for the data too
- Visualization will make your findings compelling and intuitive to non-technical coworkers.

Principles of Data Visualization

Basics of Visualization

Principles of Data Visualization

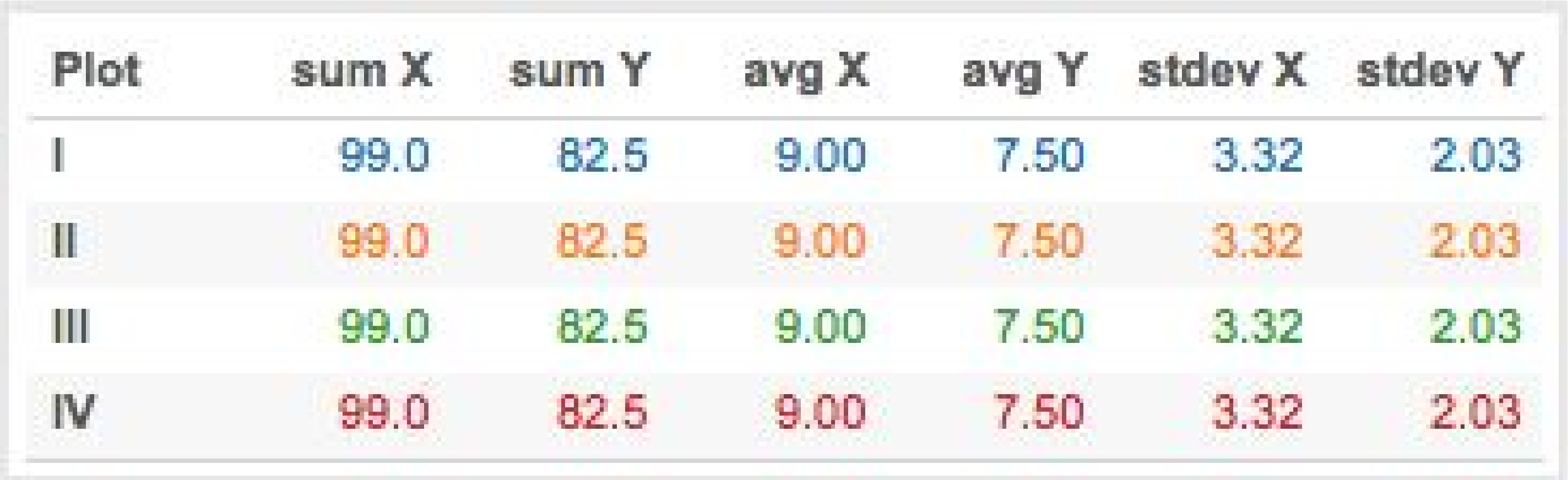
Here is the summary statistics for four plots. What do you think the visualization for each of these plots would look like?



Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03

Principles of Data Visualization

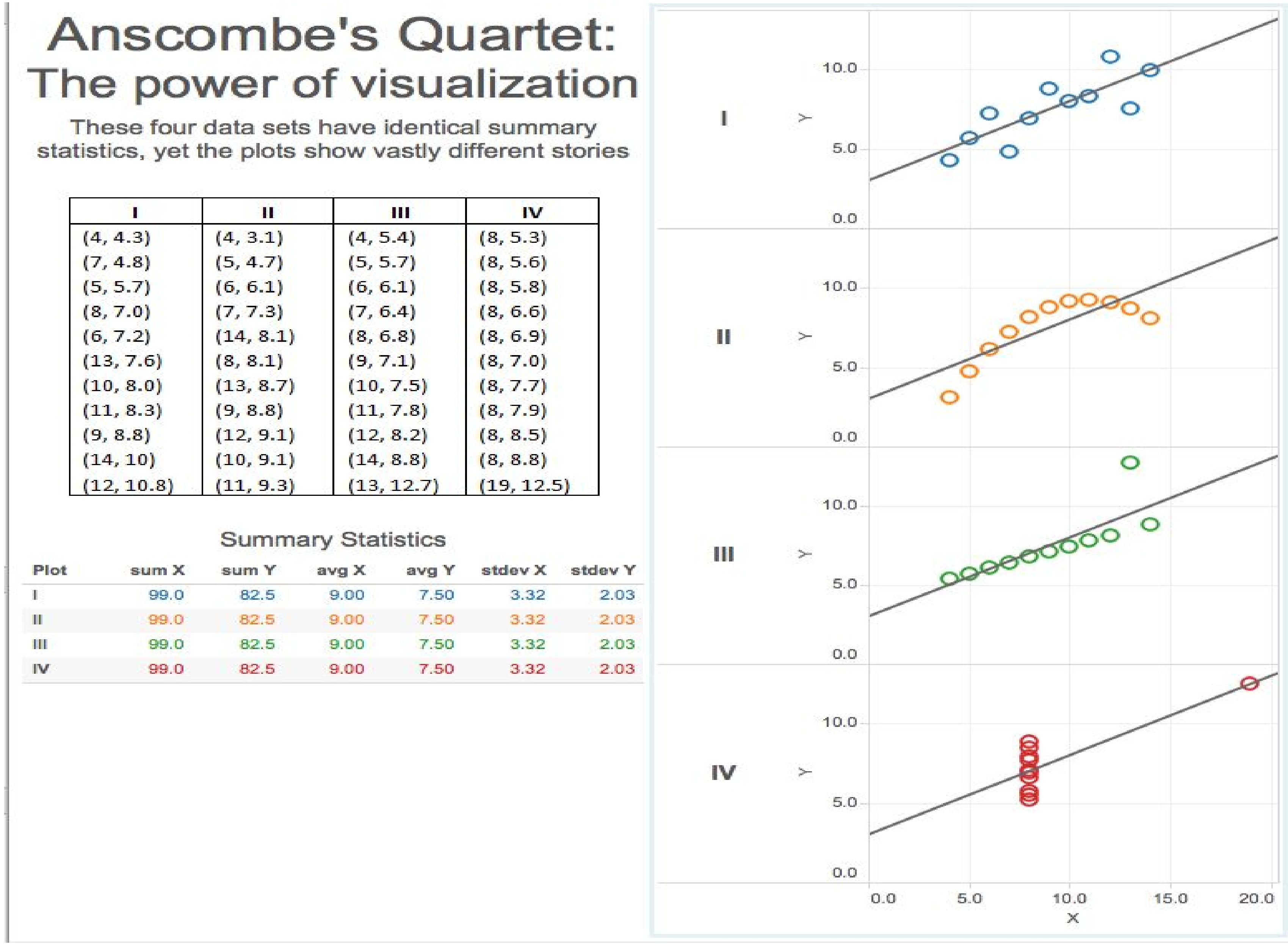
Here is the summary statistics for four plots. What do you think the visualization for each of these plots would look like?



Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03

You can probably already guess that the answer is, although the four plots have the same summary statistics, they actually are completely different and this can be seen when we visualize them together.

Principles of Data Visualization



Anscombe's Quartet: The power of visualization

These four data sets have identical summary statistics, yet the plots show vastly different stories

I	II	III	IV
(4, 4.3)	(4, 3.1)	(4, 5.4)	(8, 5.3)
(7, 4.8)	(5, 4.7)	(5, 5.7)	(8, 5.6)
(5, 5.7)	(6, 6.1)	(6, 6.1)	(8, 5.8)
(8, 7.0)	(7, 7.3)	(7, 6.4)	(8, 6.6)
(6, 7.2)	(14, 8.1)	(8, 6.8)	(8, 6.9)
(13, 7.6)	(8, 8.1)	(9, 7.1)	(8, 7.0)
(10, 8.0)	(13, 8.7)	(10, 7.5)	(8, 7.7)
(11, 8.3)	(9, 8.8)	(11, 7.8)	(8, 7.9)
(9, 8.8)	(12, 9.1)	(12, 8.2)	(8, 8.5)
(14, 10)	(10, 9.1)	(14, 8.8)	(8, 8.8)
(12, 10.8)	(11, 9.3)	(13, 12.7)	(19, 12.5)

Summary Statistics

Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03

Principles of Data Visualization

Anscomb's quartet reminds us not to rely completely on just the summary stats of our data. And that, especially during exploratory data analysis (EDA), which we will get to in Week 2, making some exploratory visual plots could keep us from making some incorrect assumptions about our data..

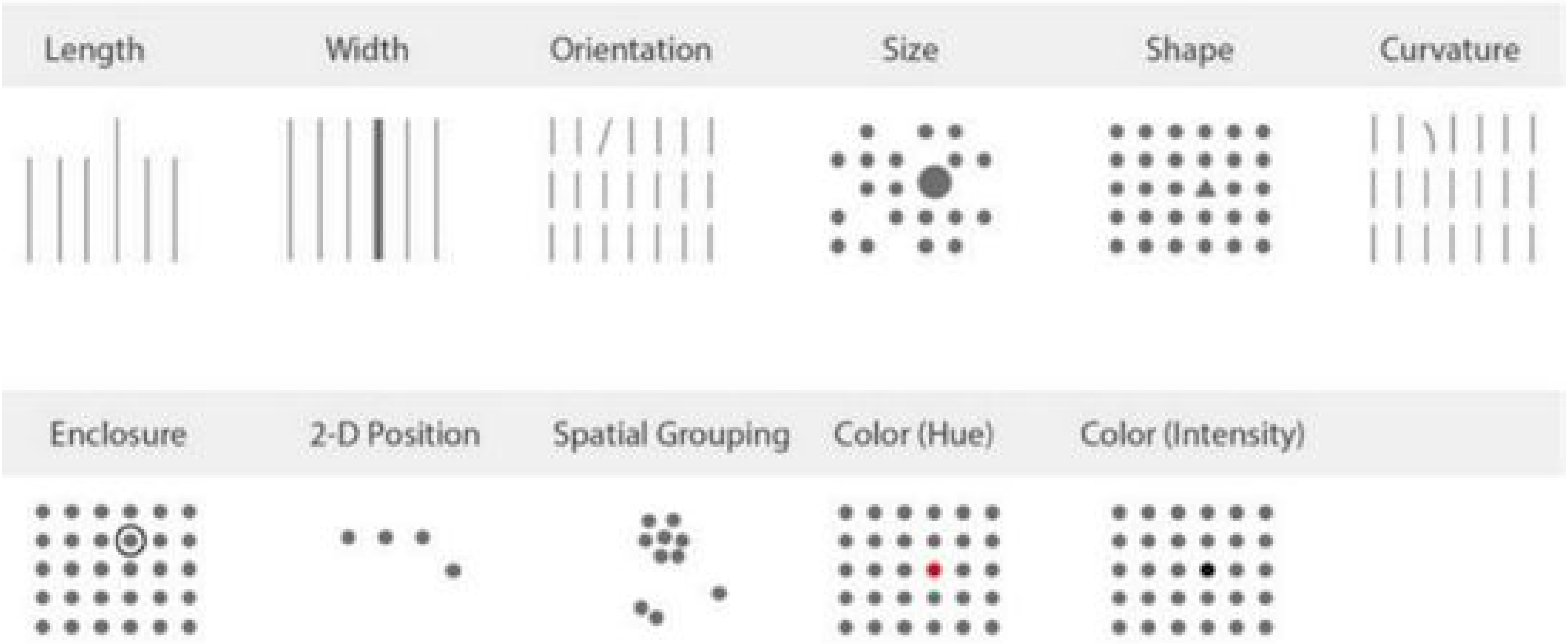
Principles of Data Visualization

What do you think might be impotant attributes that a data visualization should have?

Let's take a look at what Jeffrey Shaffer, who teaches data visualization at the University of Cincinnati, thinks is important.

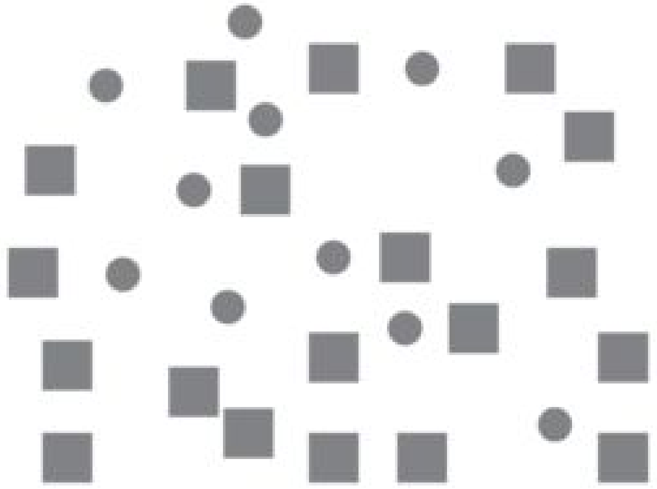
Principles of Data Visualization

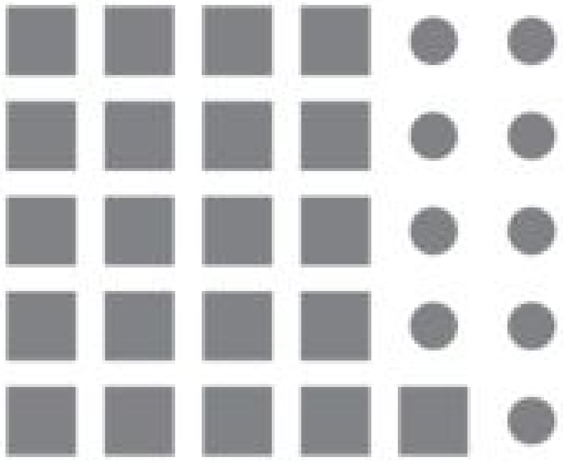
Something really interesting, is that some attributes have more of an effect on our brains. The one we tend to focus on most is position, then color and size.

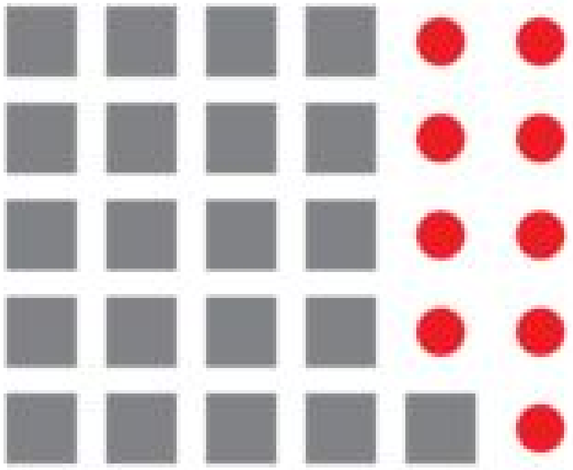


Principles of Data Visualization

Let's take a look at three visualizations. Which one of the three catches your attention the most and why?

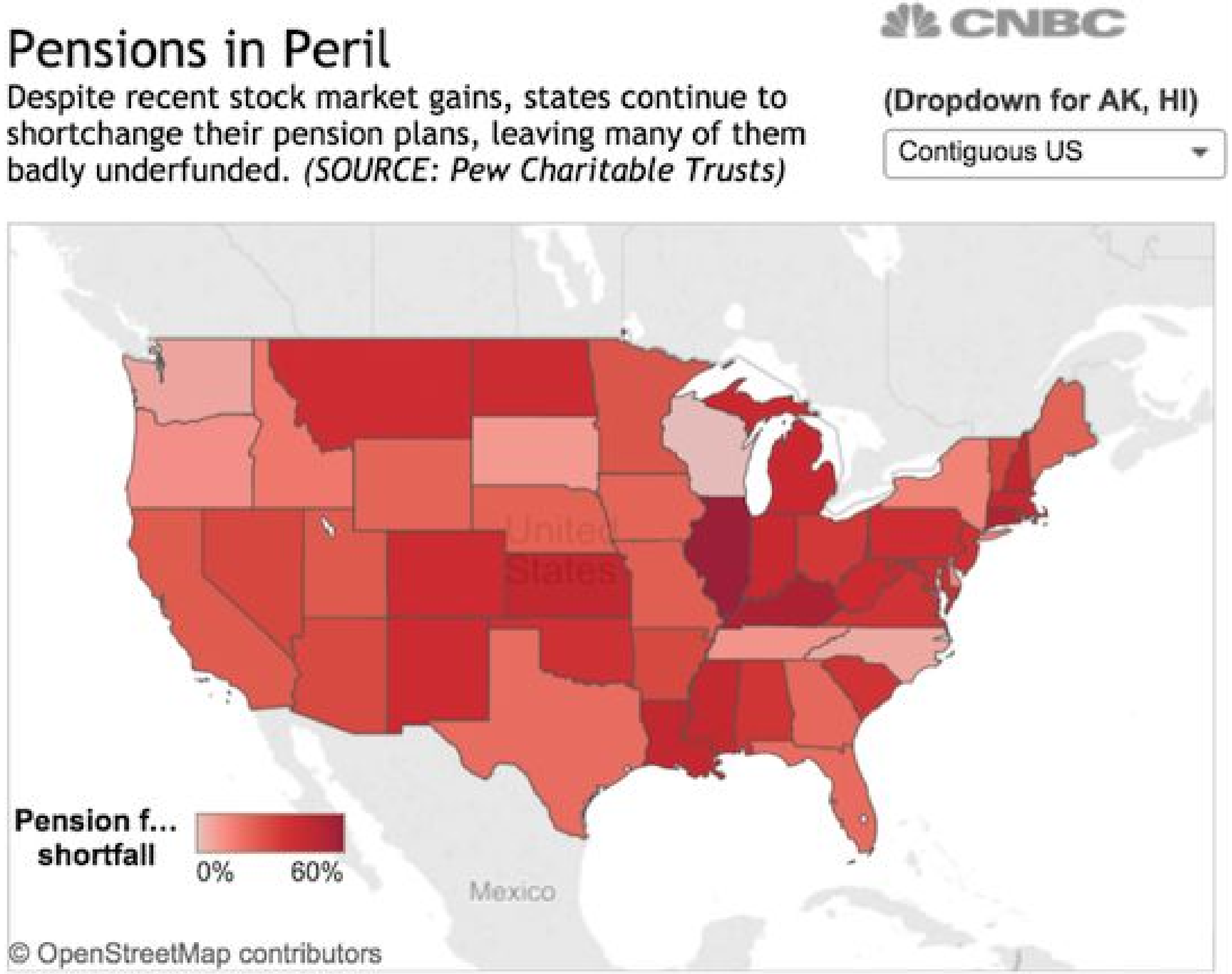






Principles of Data Visualization

Sequential colors are used to show values ordered from low to high.



Pensions in Peril

Despite recent stock market gains, states continue to shortchange their pension plans, leaving many of them badly underfunded. *(SOURCE: Pew Charitable Trusts)*

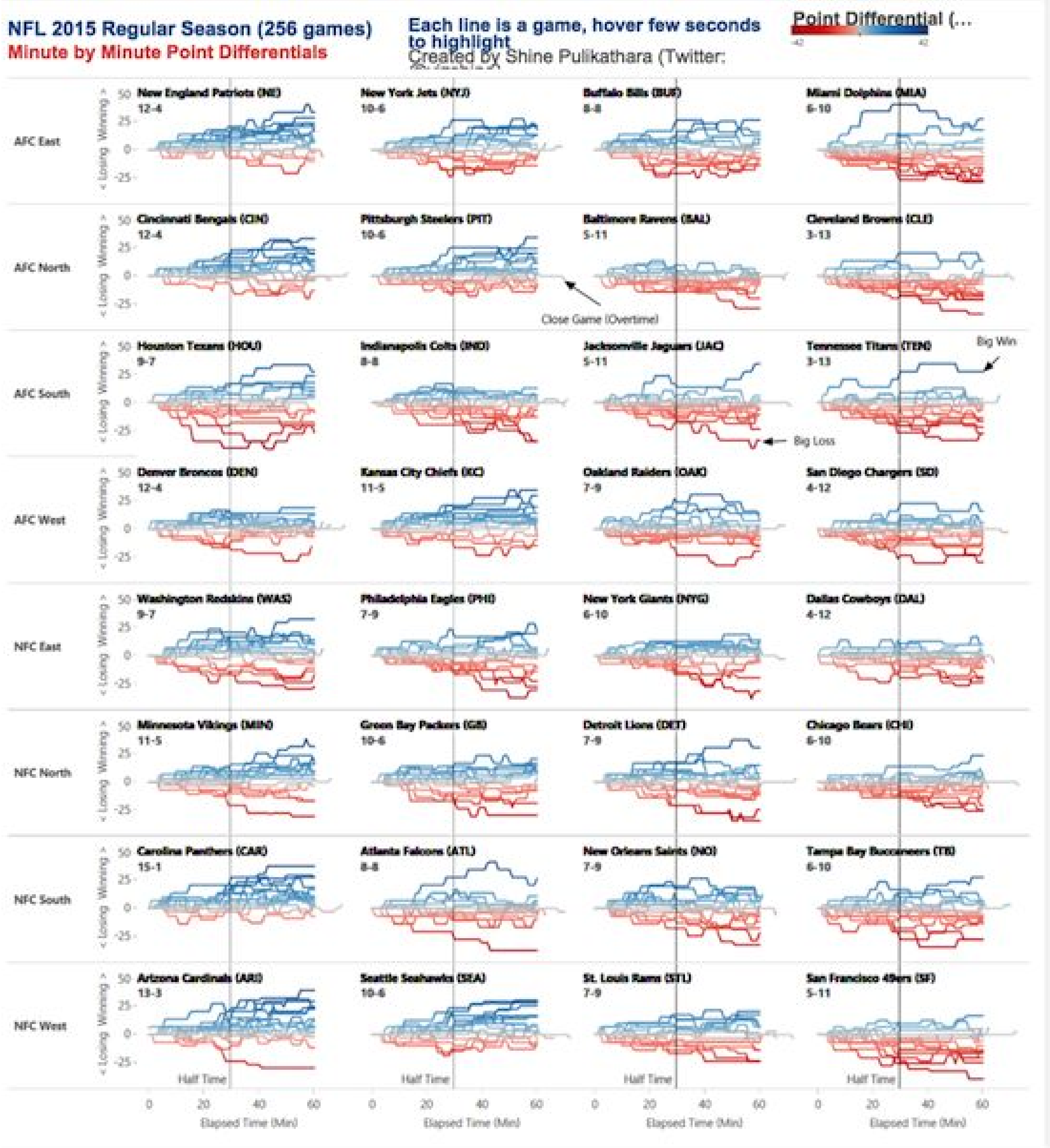
(Dropdown for AK, HI)

Contiguous US

© OpenStreetMap contributors

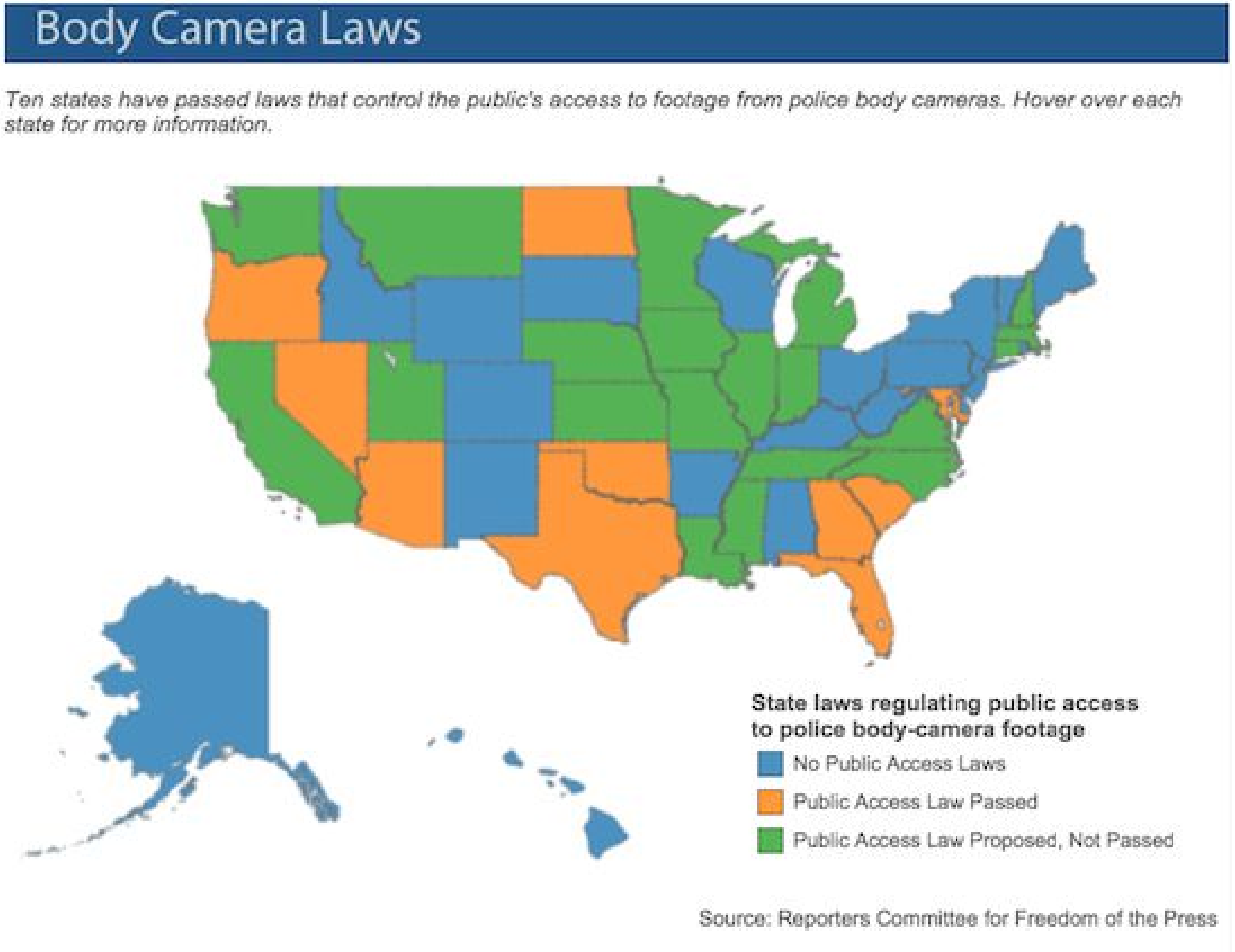
Principles of Data Visualization

Divergent colors are used to show ordered values that have a critical midpoint, like an average or zero.



Principles of Data Visualization

Categorical colors are used to distinguish data that falls into distinct groups.



Body Camera Laws

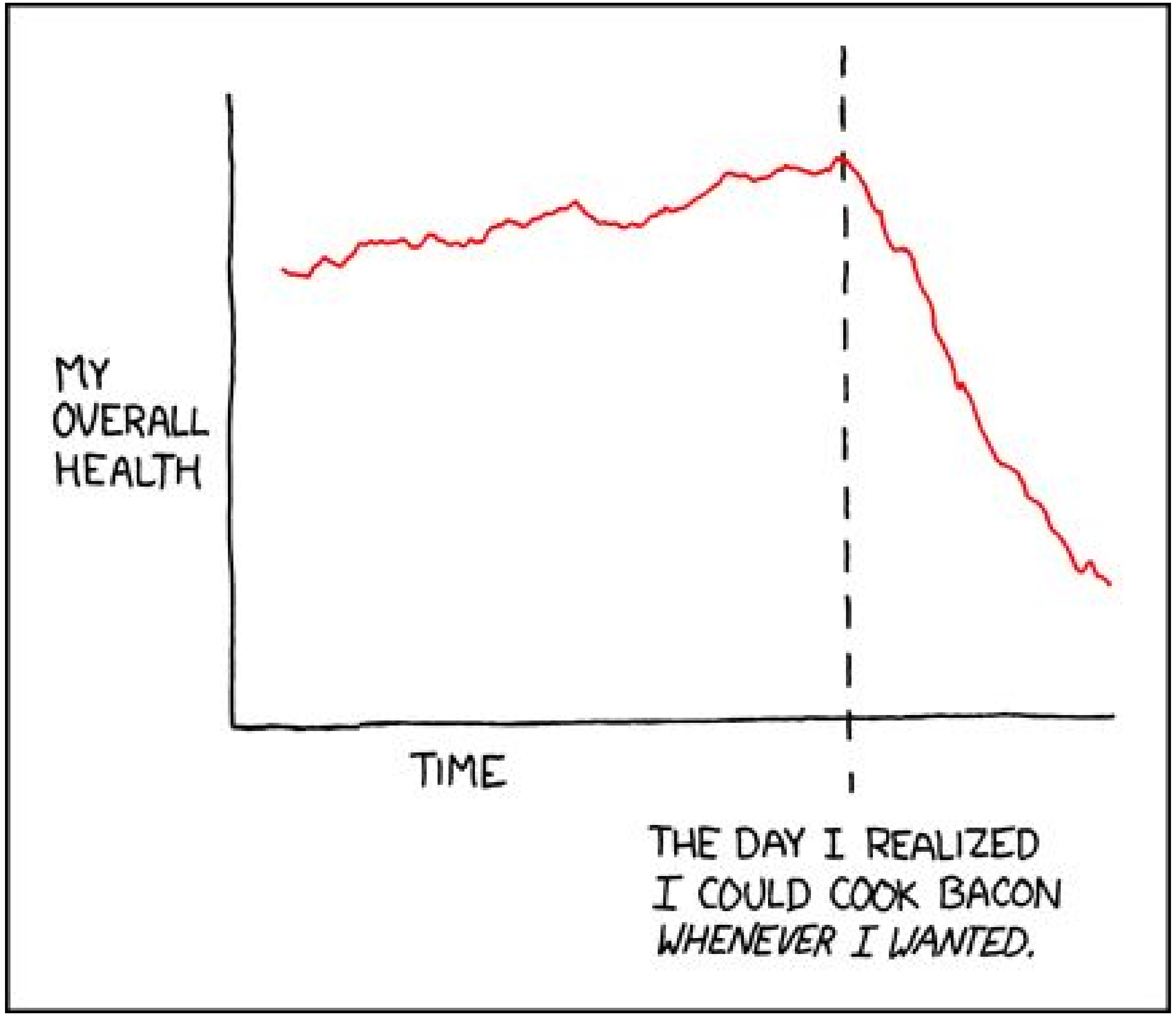
Ten states have passed laws that control the public's access to footage from police body cameras. Hover over each state for more information.

State laws regulating public access to police body-camera footage
No Public Access Laws
Public Access Law Passed
Public Access Law Proposed, Not Passed

Source: Reporters Committee for Freedom of the Press

Principles of Data Visualization

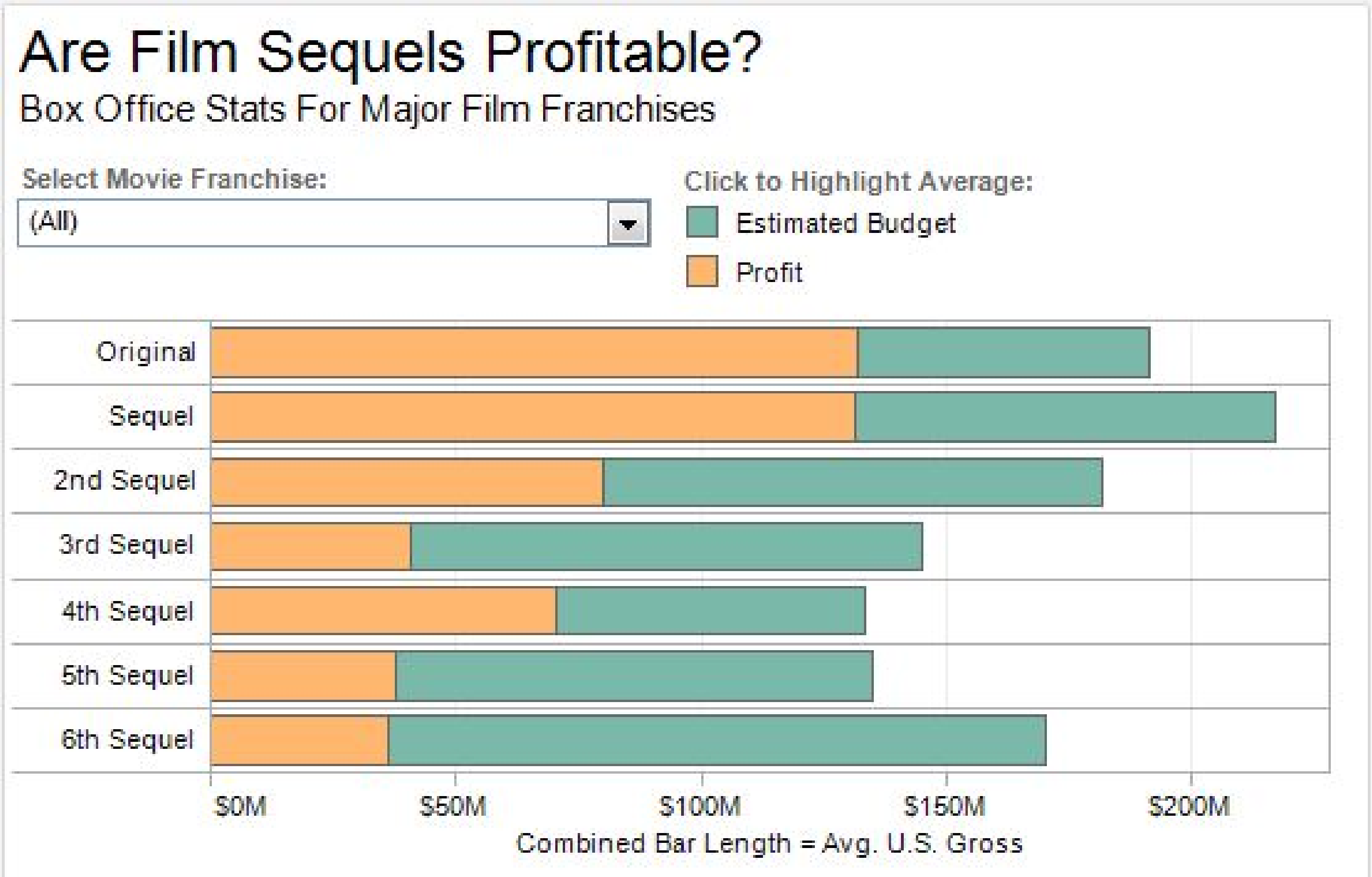
In addition to taking visualization attributes into consideration, you should also consider carefully, which kind of chart or graph to use. Let's look at a few commonly used charts and graphs.



Principles of Data Visualization

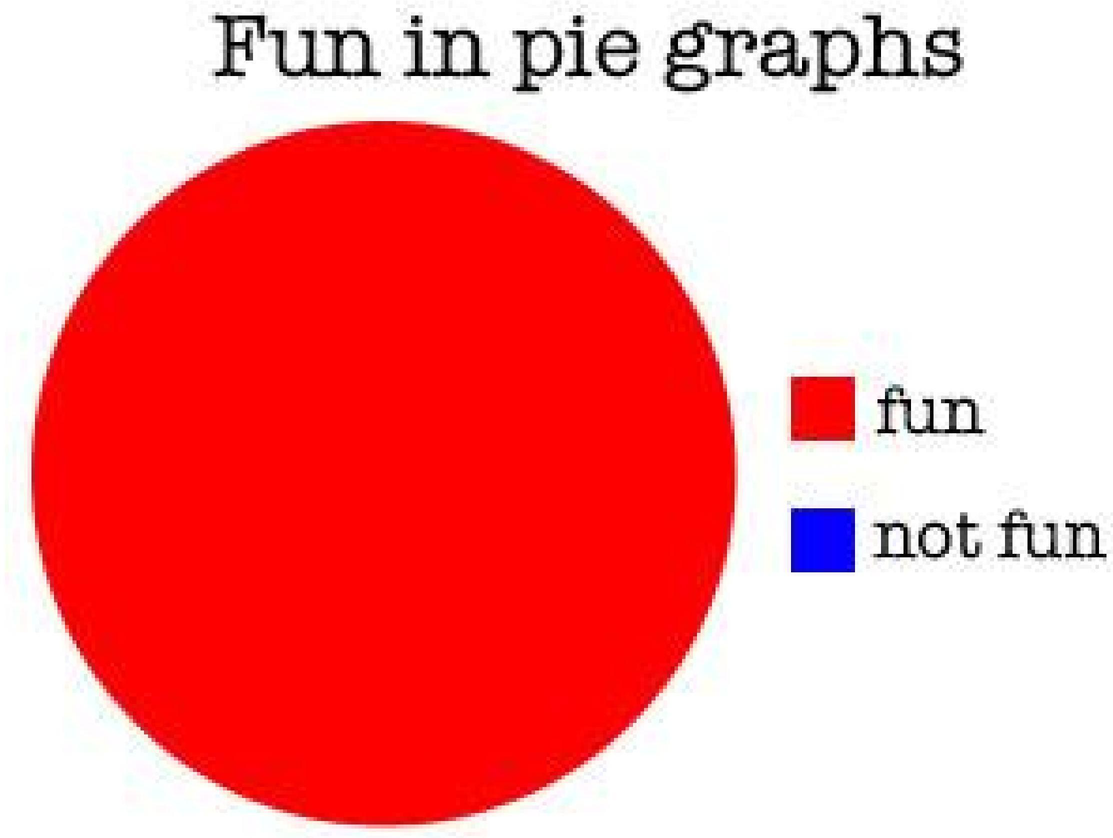
Bar Charts

Bar charts are one of the most common ways to visualize data. Why? Because it's easy to compare information, revealing highs and lows quickly. Bar charts are most effective when you have numerical data that splits neatly into different categories.

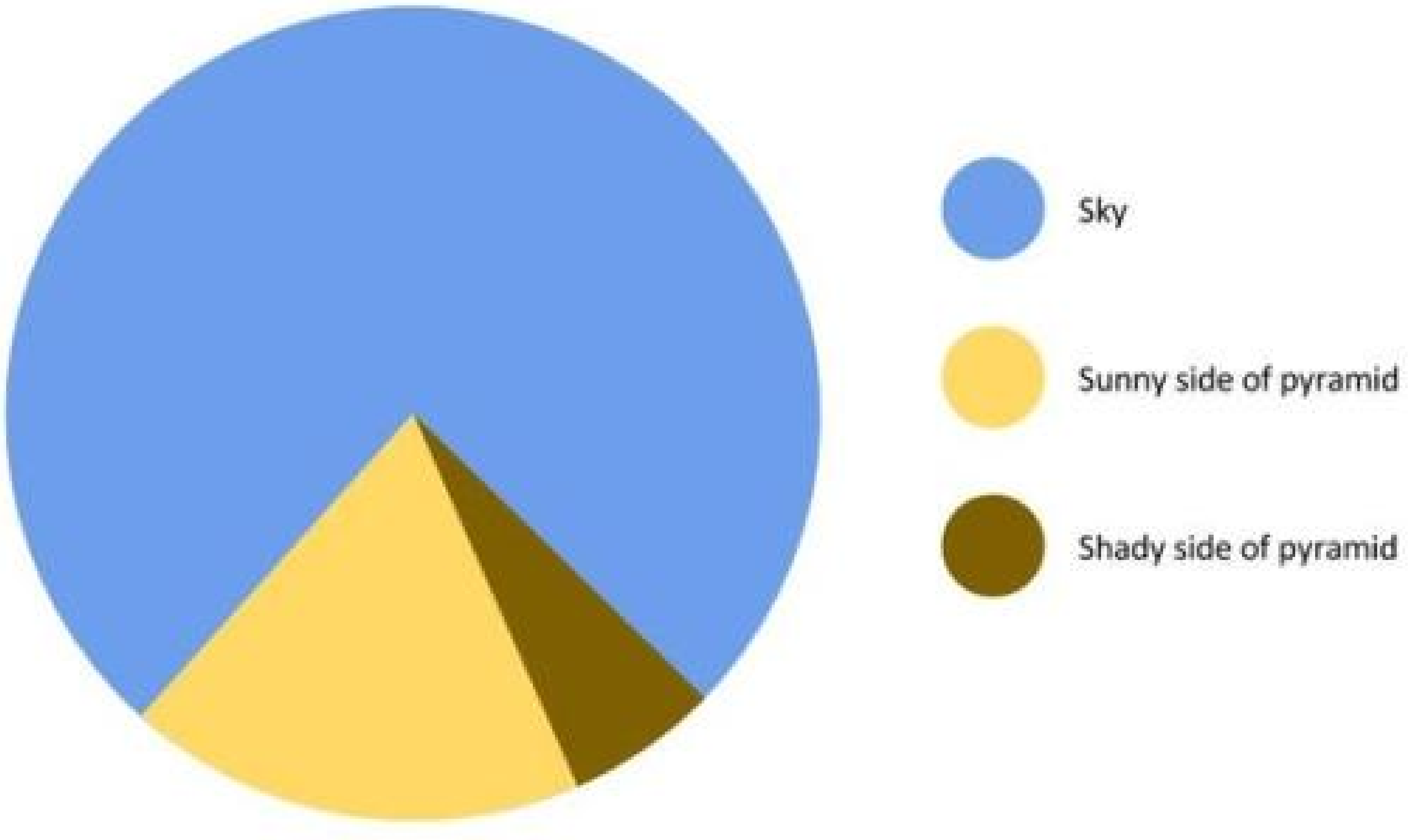


Pie Charts

The only time that pie charts should be used is to show relative proportions or percentages of information. Pie charts are the most commonly mis-used chart type. If you want to compare data, leave it to bars or stacked bars.

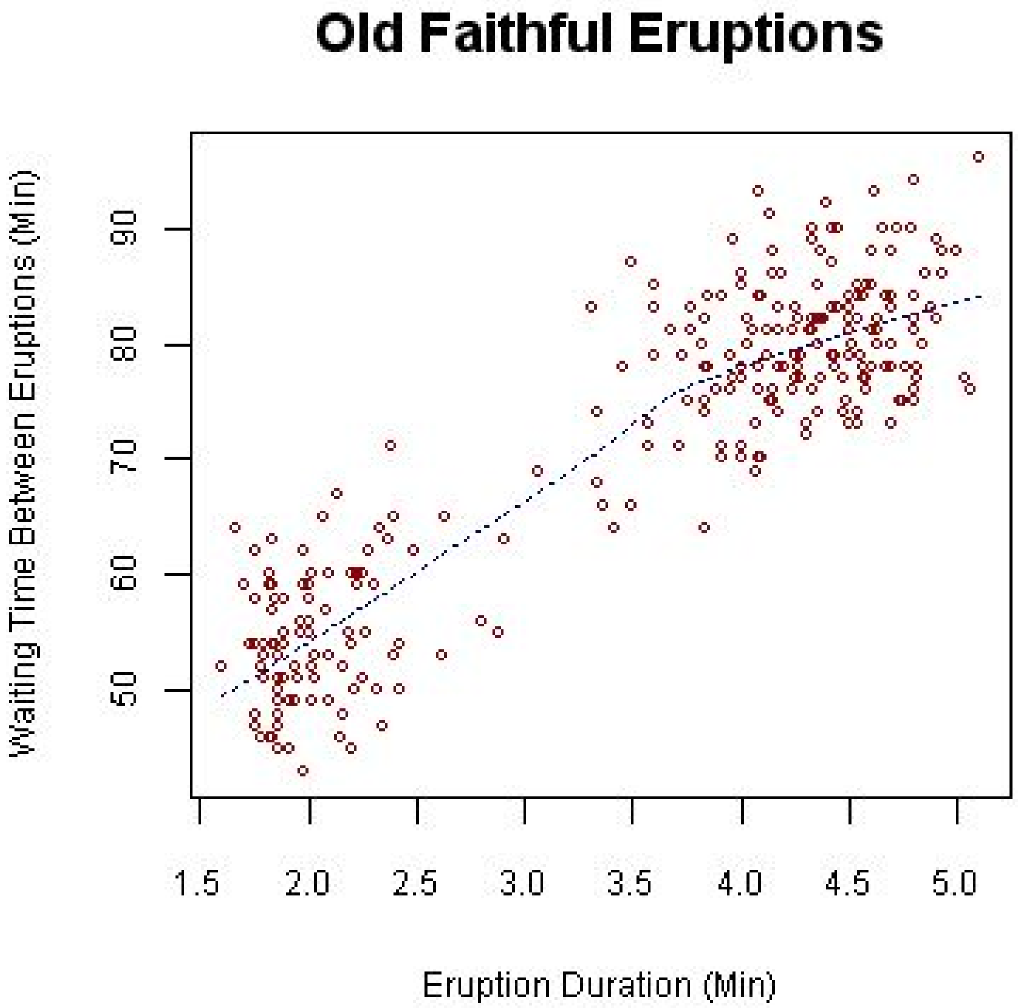


A Potentially Useful Pie Chart?



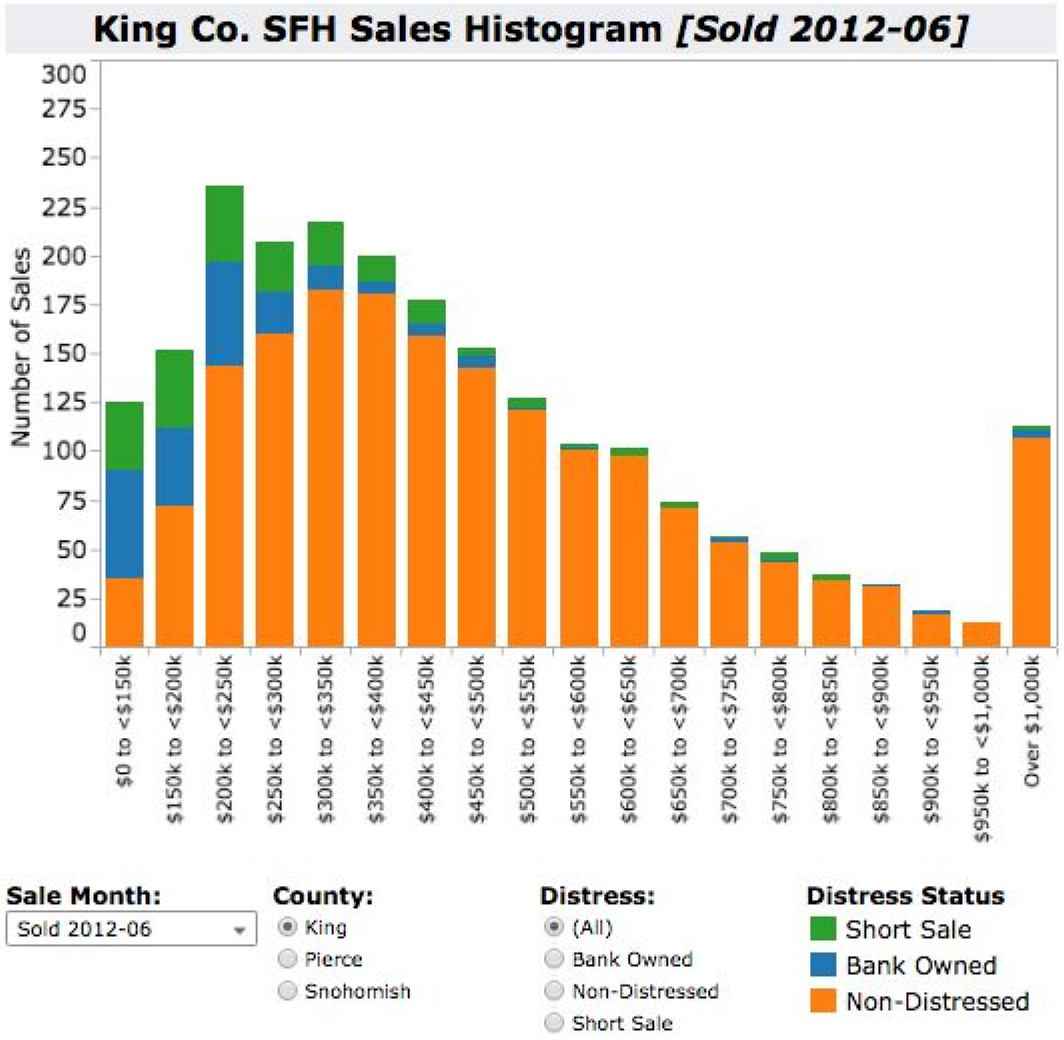
Scatter Plot

Scatter plots are a great way to give you a sense of trends, concentrations and outliers. This will give you a good idea of where you may want to investigate further.



Histrograms

Histograms are useful when you want to see how your data are distributed across groups.



Sale Month: Sold 2012-06

County:
● King
○ Pierce
○ Snohomish

Distress:
● (All)
○ Bank Owned
○ Non-Distressed
○ Short Sale

Distress Status
🟩 Short Sale
🟦 Bank Owned
🟧 Non-Distressed

Principles of Data Visualization

Python Visualization

Principles of Data Visualization

We are going to cover the basics of visualizing data in python using some popular python packages.

- **matplotlib** is the low-level but powerful standard plotting package for python.
- **seaborn** builds on top of matplotlib. It is much easier to use and looks better, but is more restricted in functionality.
- **plotly** is a fancy plotting library with very nice visuals but very different syntax than matplotlib.

Principles of Data Visualization

Prerequisites:

- Plotly
- Matplotlib
- Seaborn

Don’t have one or all of these? Let’s take a moment and install them.

Principles of Data Visualization

Prerequisites:

- Plotly
- **Matplotlib**
- Seaborn

Principles of Data Visualization

Prerequisites:

- Plotly
- Matplotlib
- **Seaborn**

Prerequisites:

- **Plotly**
- Matplotlib
- Seaborn

Principles of Data Visualization

Plotting with Tableau

Conclusion