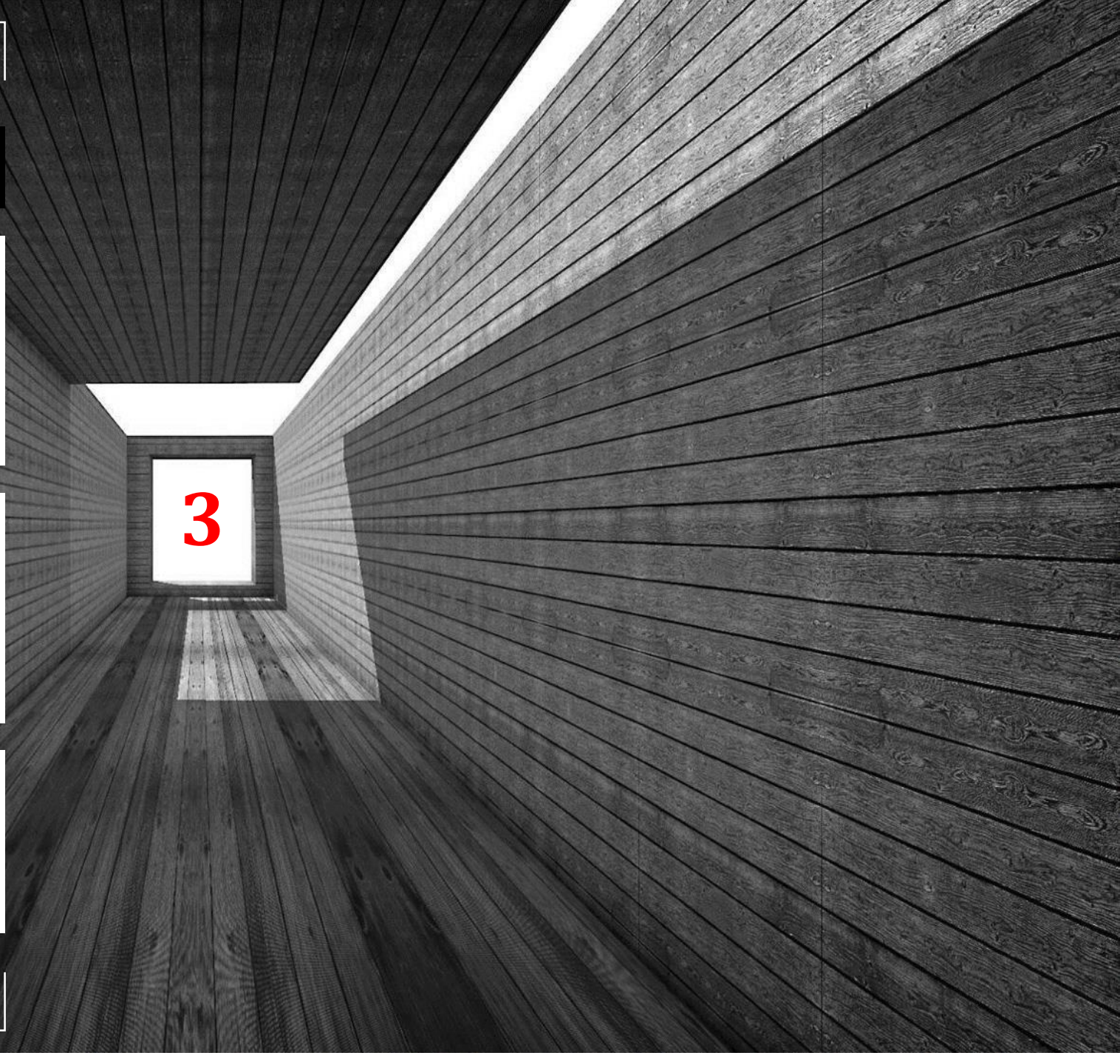
The German University in Cairo

CSEN1095
Data Engineering



Lecture 3
Explore Your Data II

Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg

Visual Representations of Data

In addition to the usual representations we use regularly:

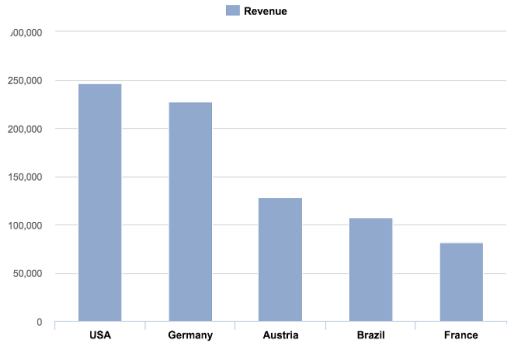
- **Line charts**
- **Bar charts**
- **Pie charts**

We also have:

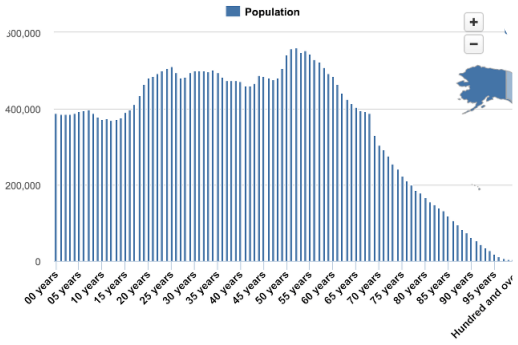
- **Boxplots**
- **Histograms**
- **Scatter Plots**

Additional representations:

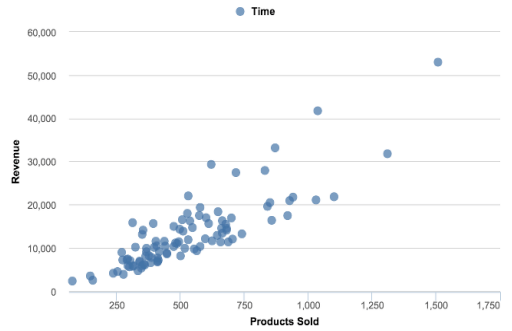
- Quantile and Q-Q Plots
- **Stacked bar and area charts**



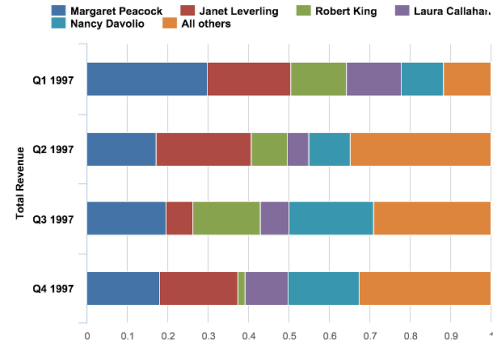
Comparison



Distribution



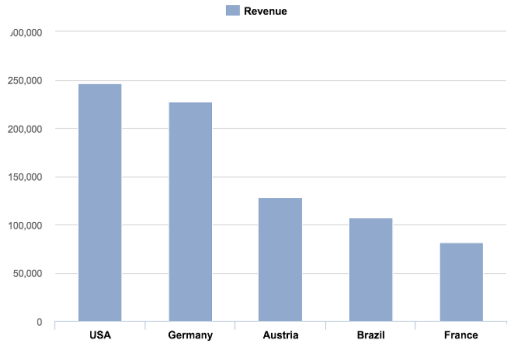
Relationship



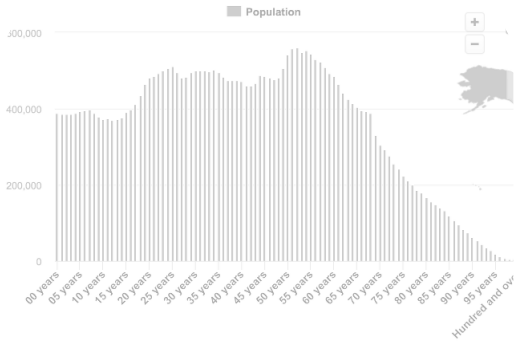
Composition

Methods

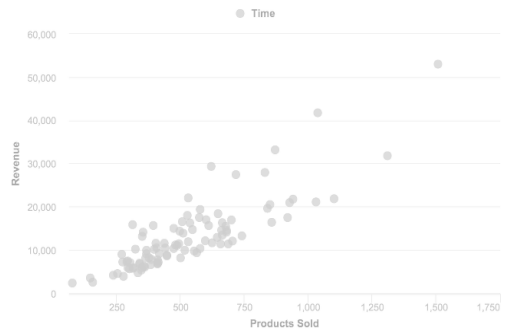
- Evaluate and compare values between two or more data categories
- Column, bar, and line charts
- *Trends – patterns of change*
  - A time axis is involved



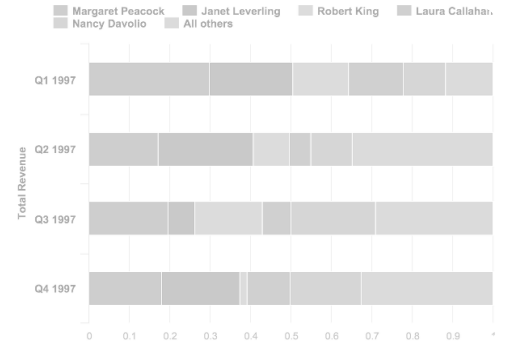
Comparison



Distribution



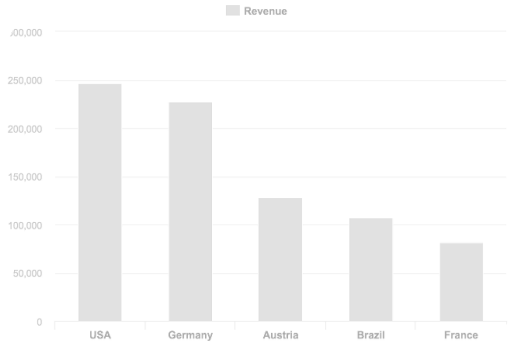
Relationship



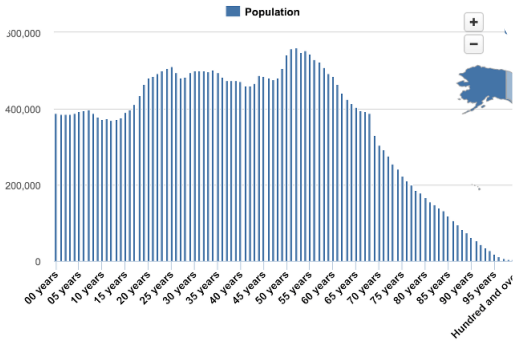
Composition

Methods

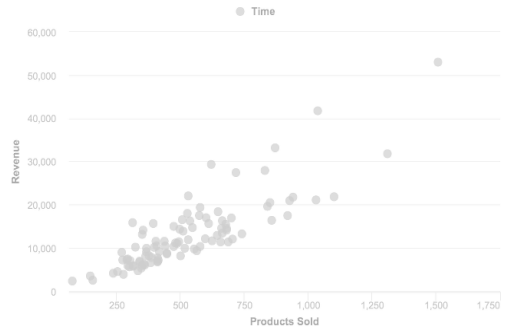
- Visualize full data spectrum
- Often used to visualize deviations or anomalies – Deviation Analysis
- Column and bar histograms, line and area charts, scatter plots, maps



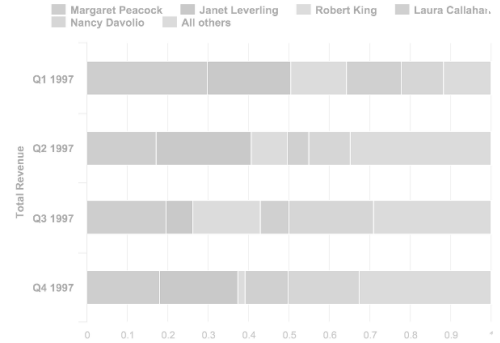
Comparison



Distribution



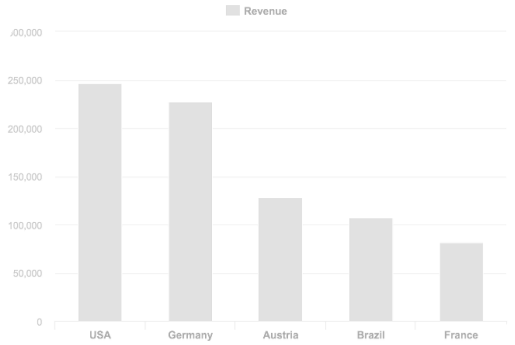
Relationship



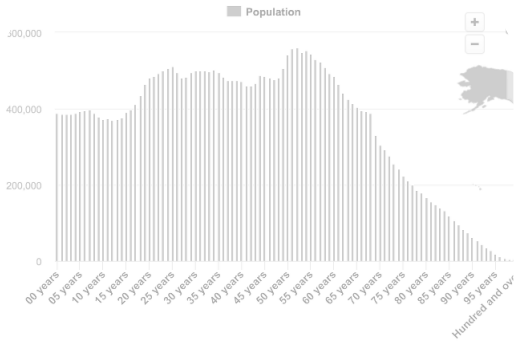
Composition

Methods

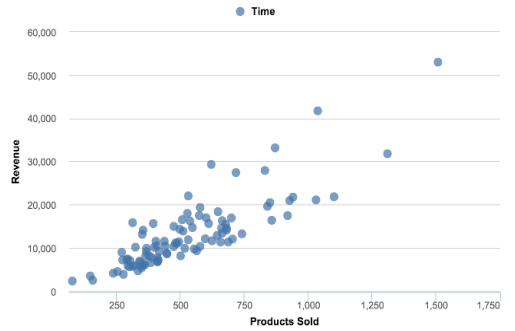
- Show the relationship, correlation, or connection of two or more variables
- Scatter plots, bubble charts (three to four dimensions, using bubble size and color scale), line charts



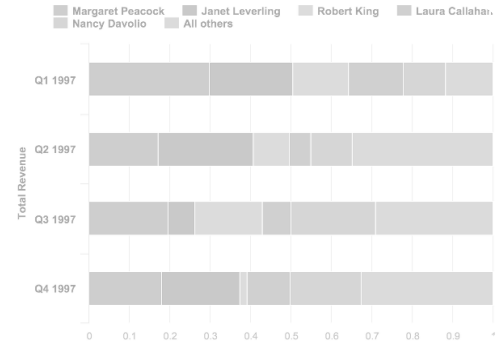
Comparison



Distribution



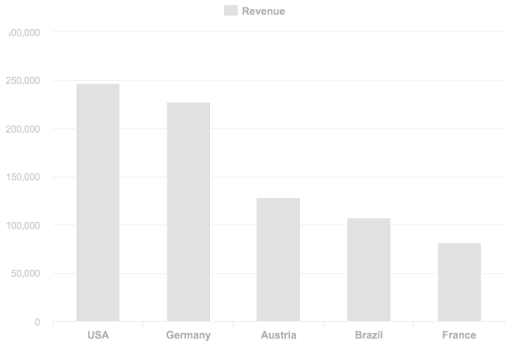
Relationship



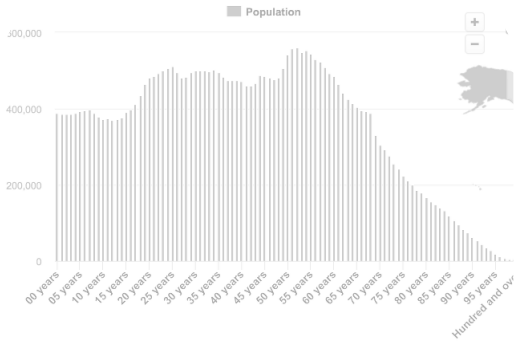
Composition

Methods

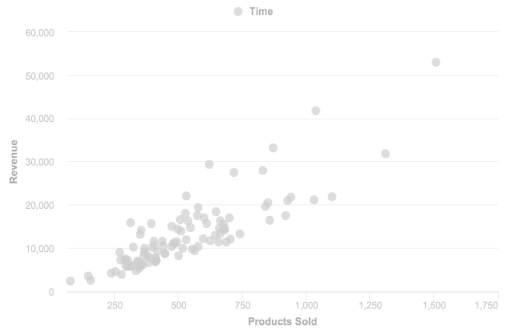
- Show how a total value can be divided into parts or to highlight the significance of each part relative to the total value
- Stacked bars or columns, pie charts, stacked area charts
- Often misunderstood and abused
- You should not use pie or donut charts because people are bad at evaluating angles



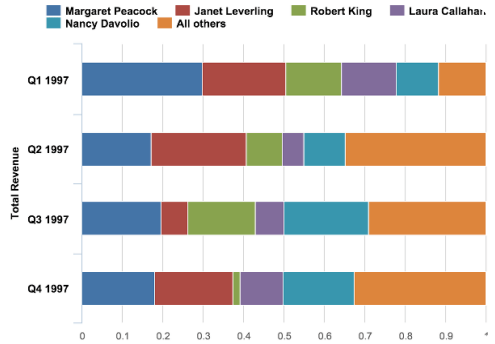
Comparison



Distribution



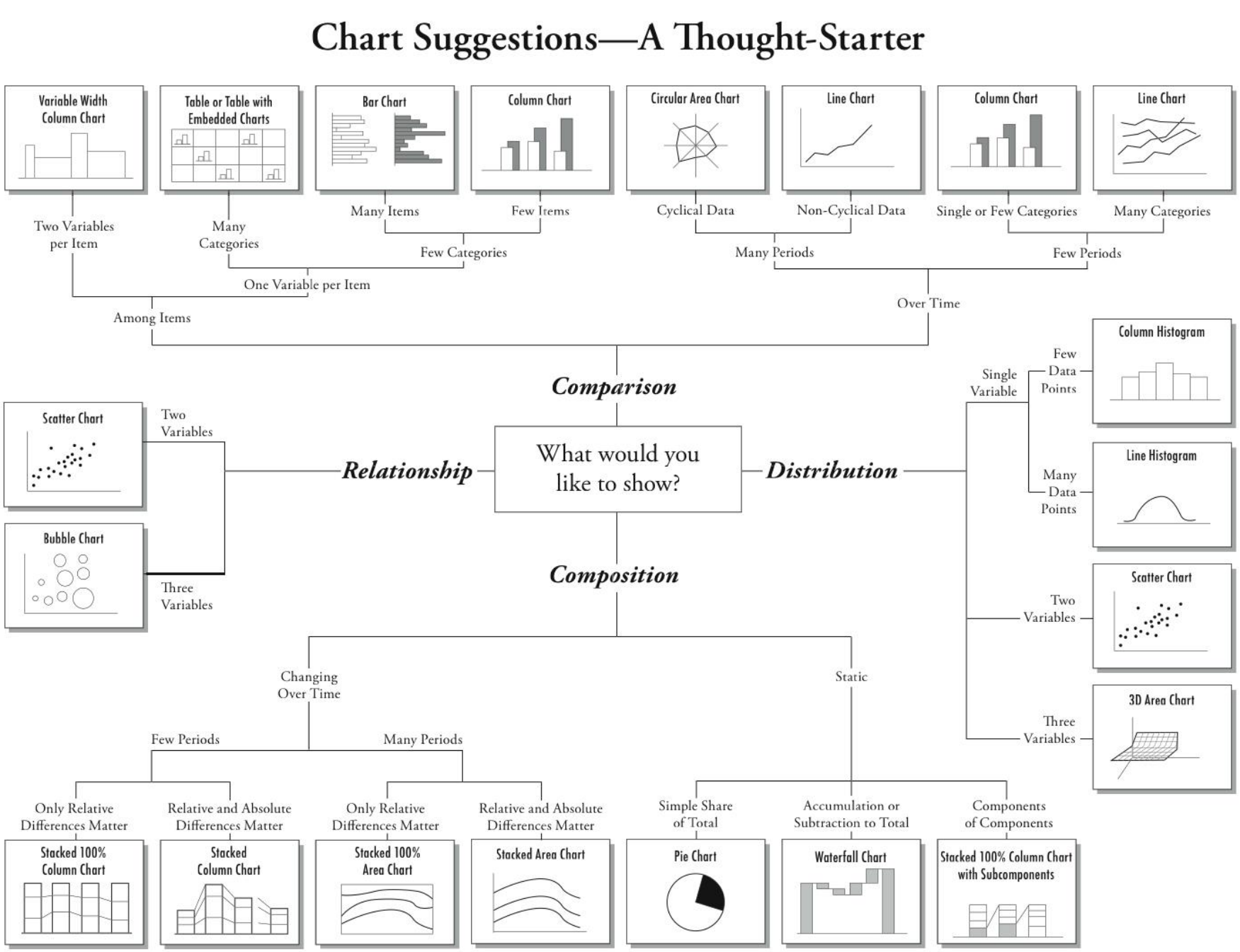
Relationship



Composition

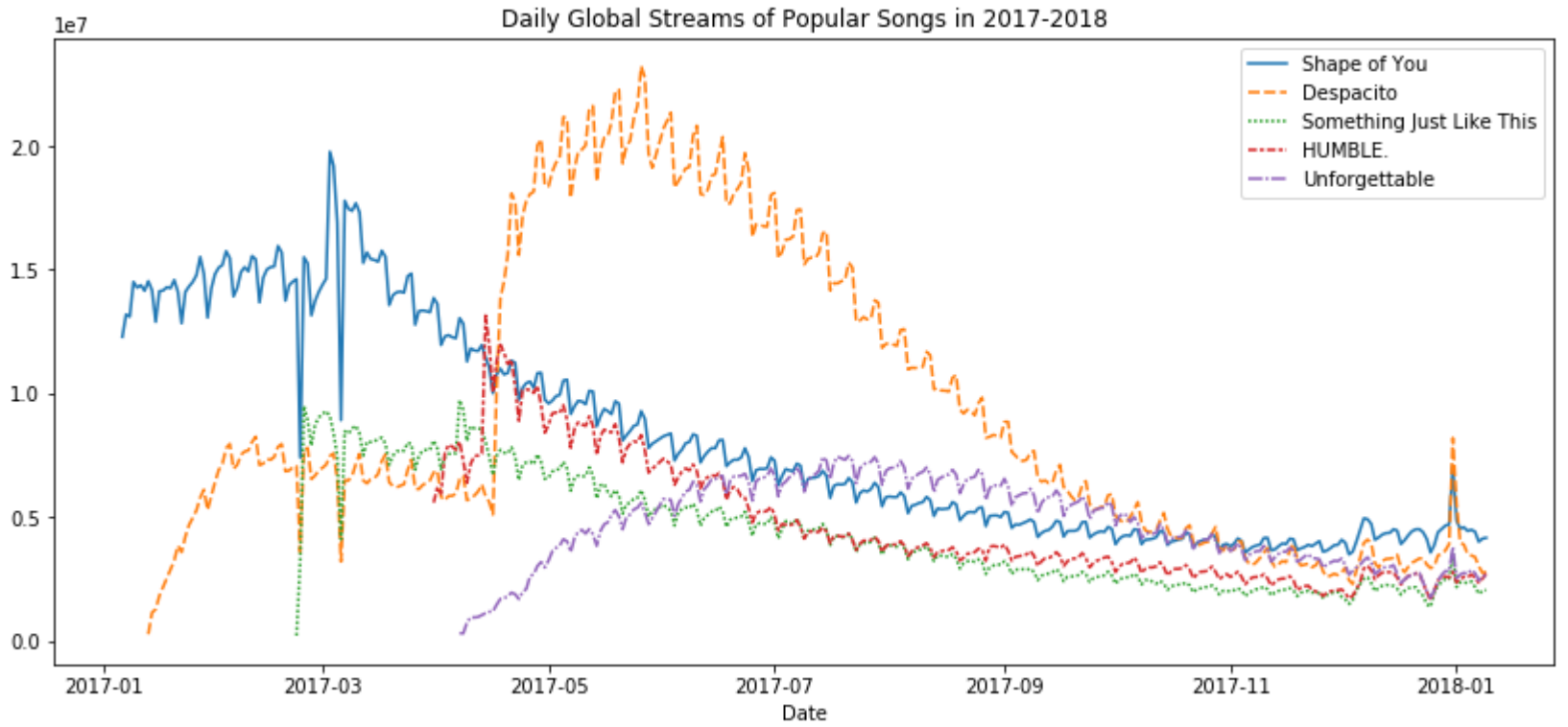
Choice is based on:

- Number of variables
- Number of data points
- Time



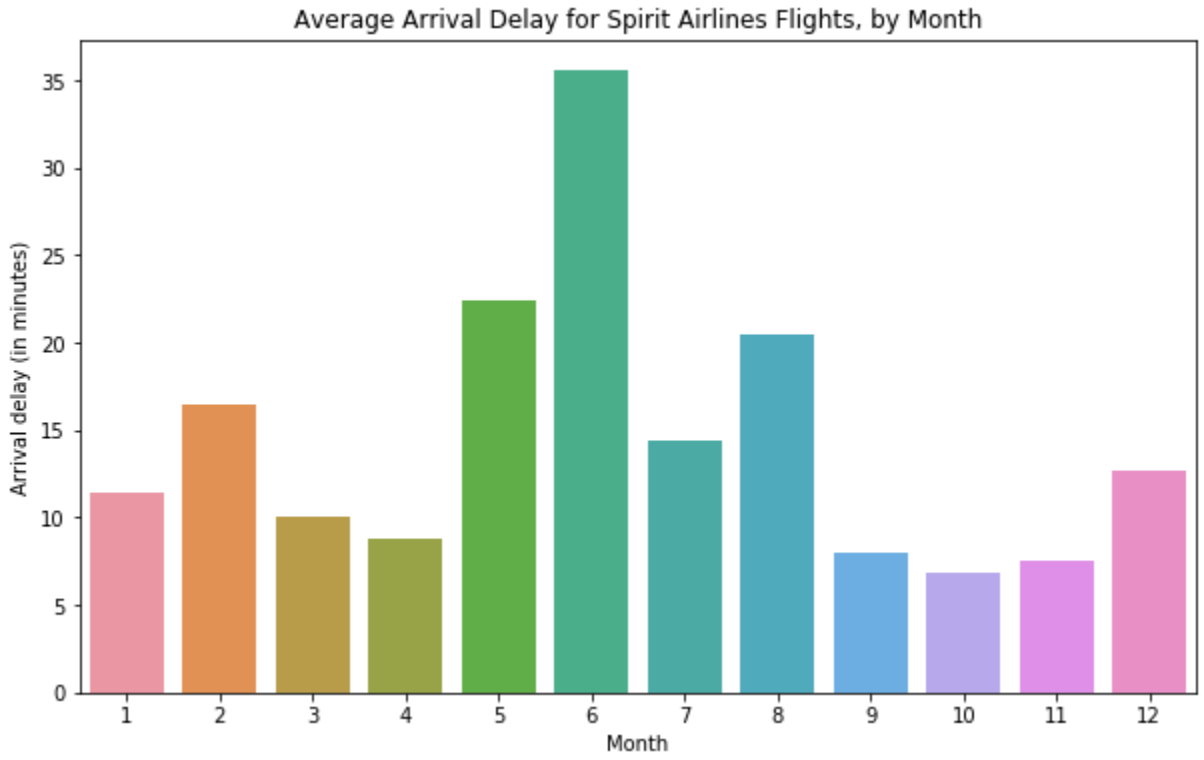
Line Charts

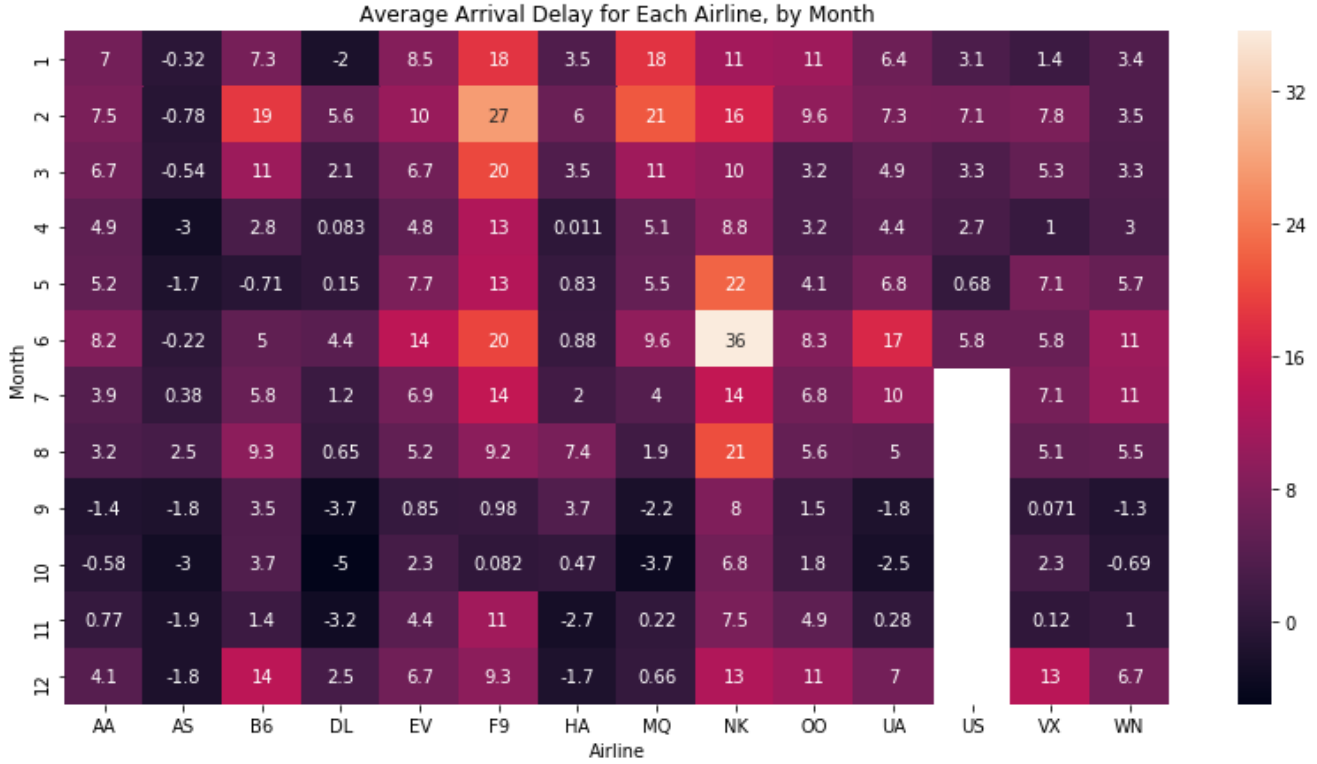
- Trends over time



Source: Kaggle Data Visualization Micro-course

Bar Charts

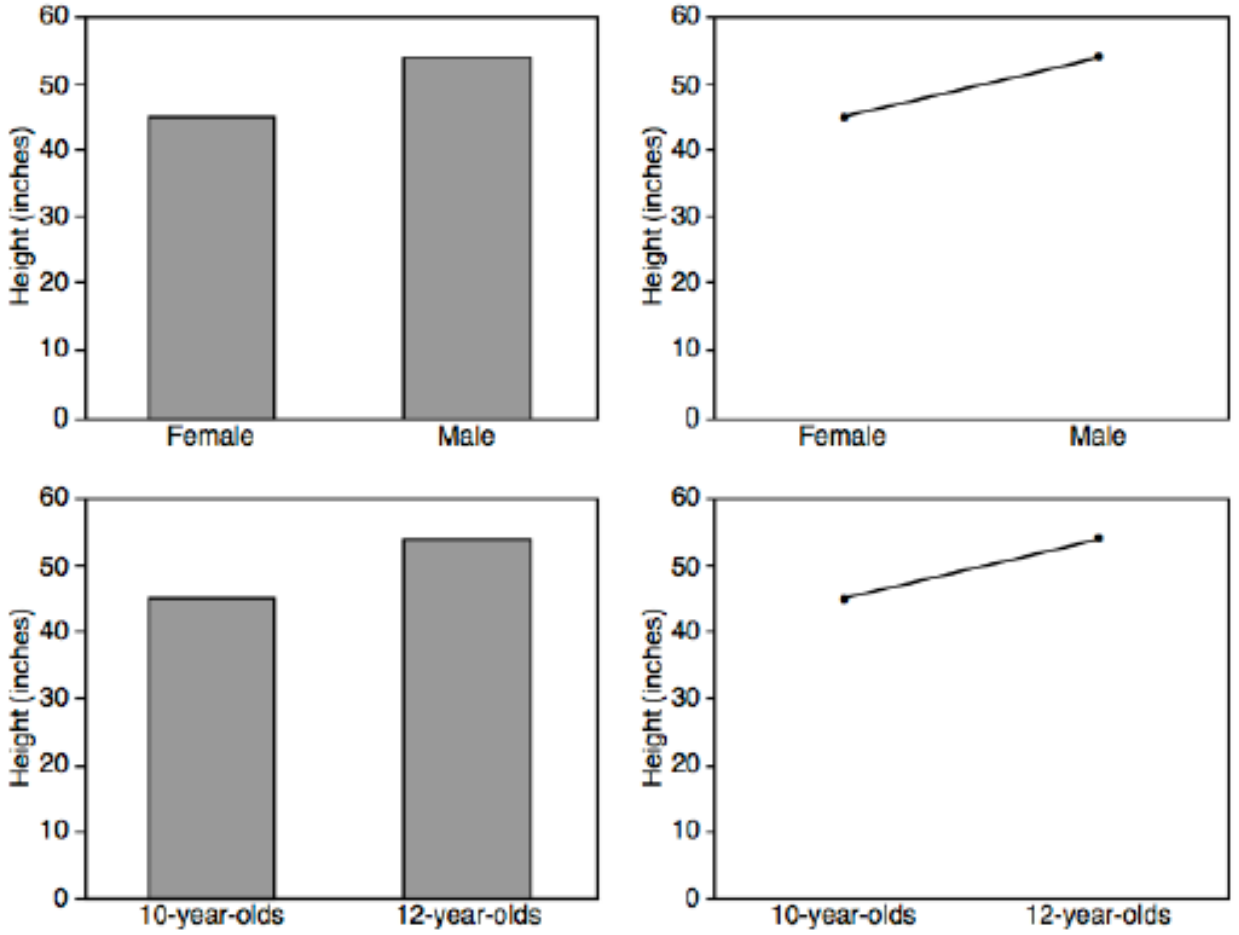




Source: CS109 Stanford's Data Science Course

Bar Charts

Which representation is more accurate?

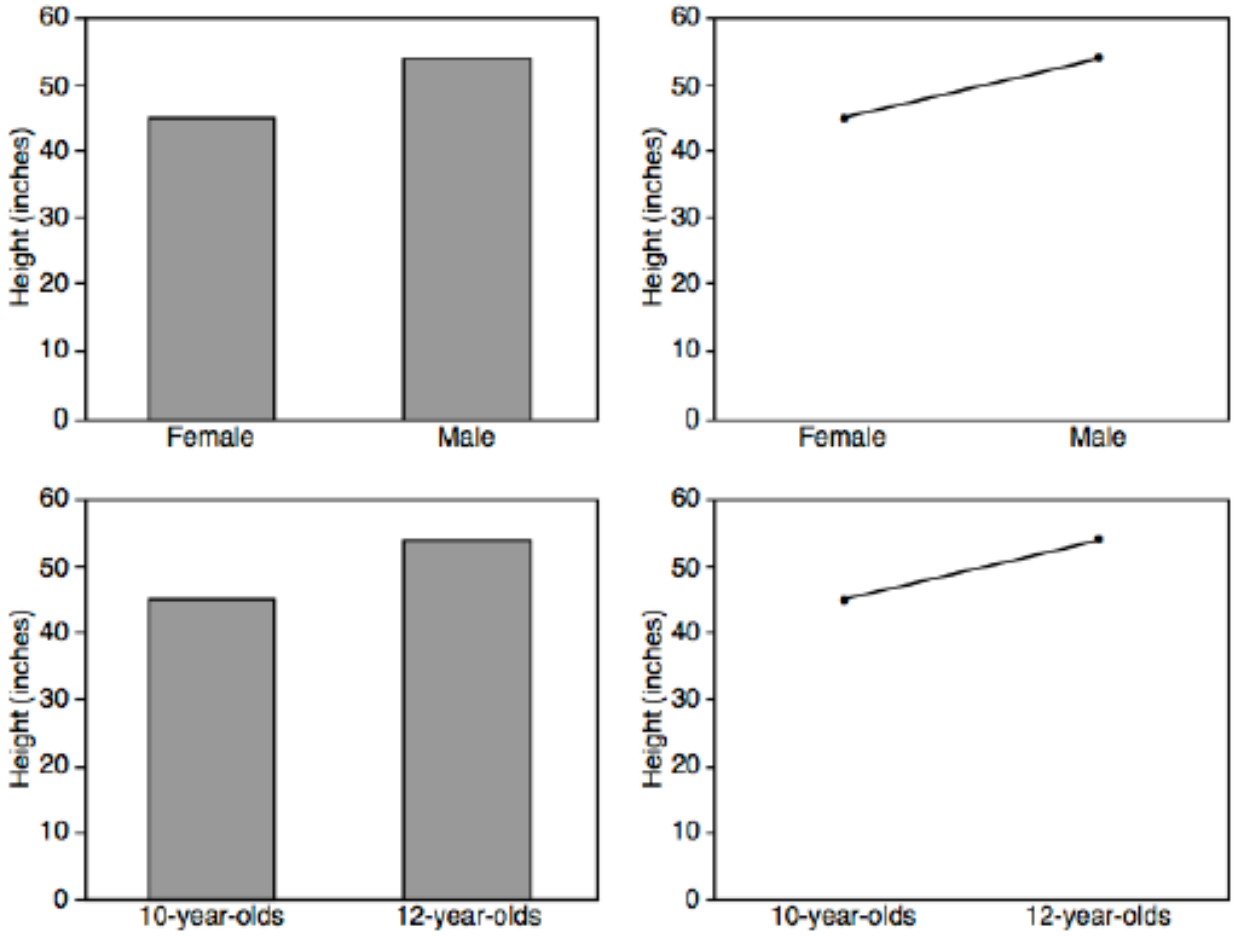


Source: CS109 Stanford's Data Science Course

Bar Charts

Which representation is more accurate?

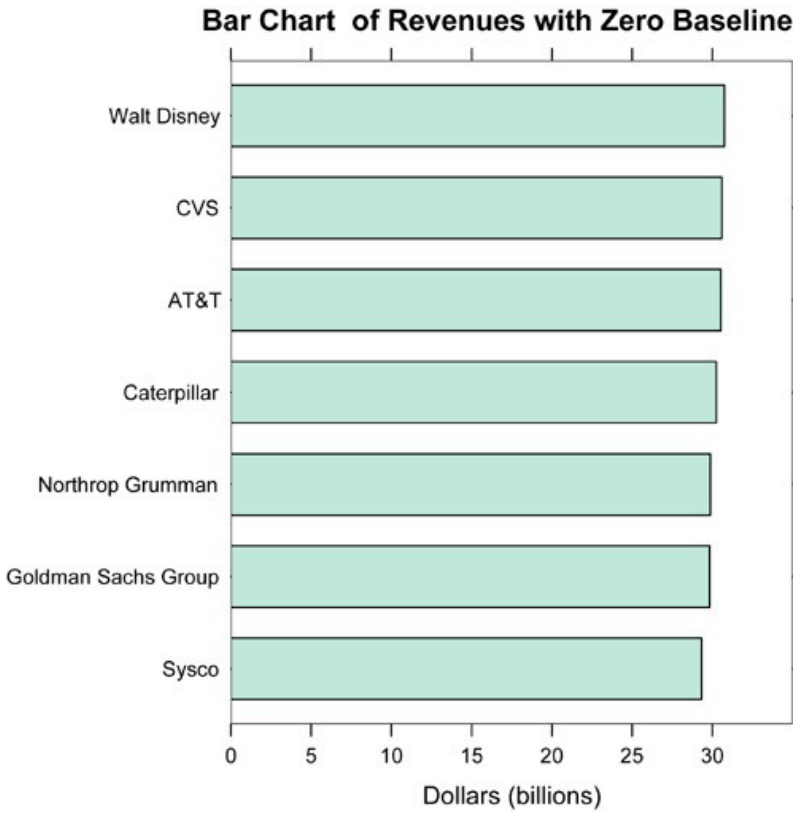
- *What is the variable type?*

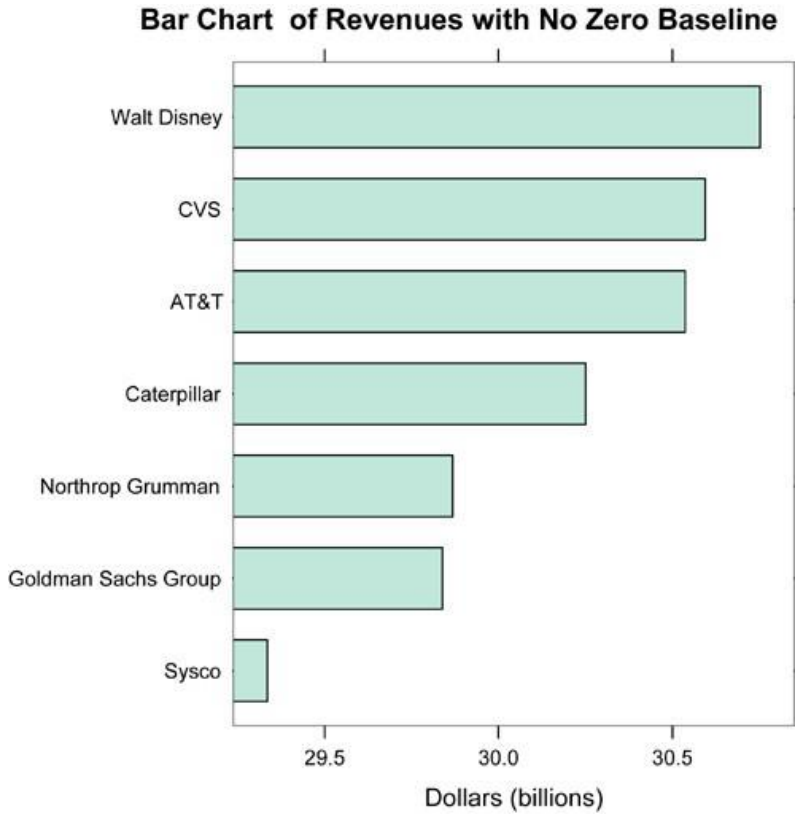


Source: CS109 Stanford's Data Science Course

Bar Charts

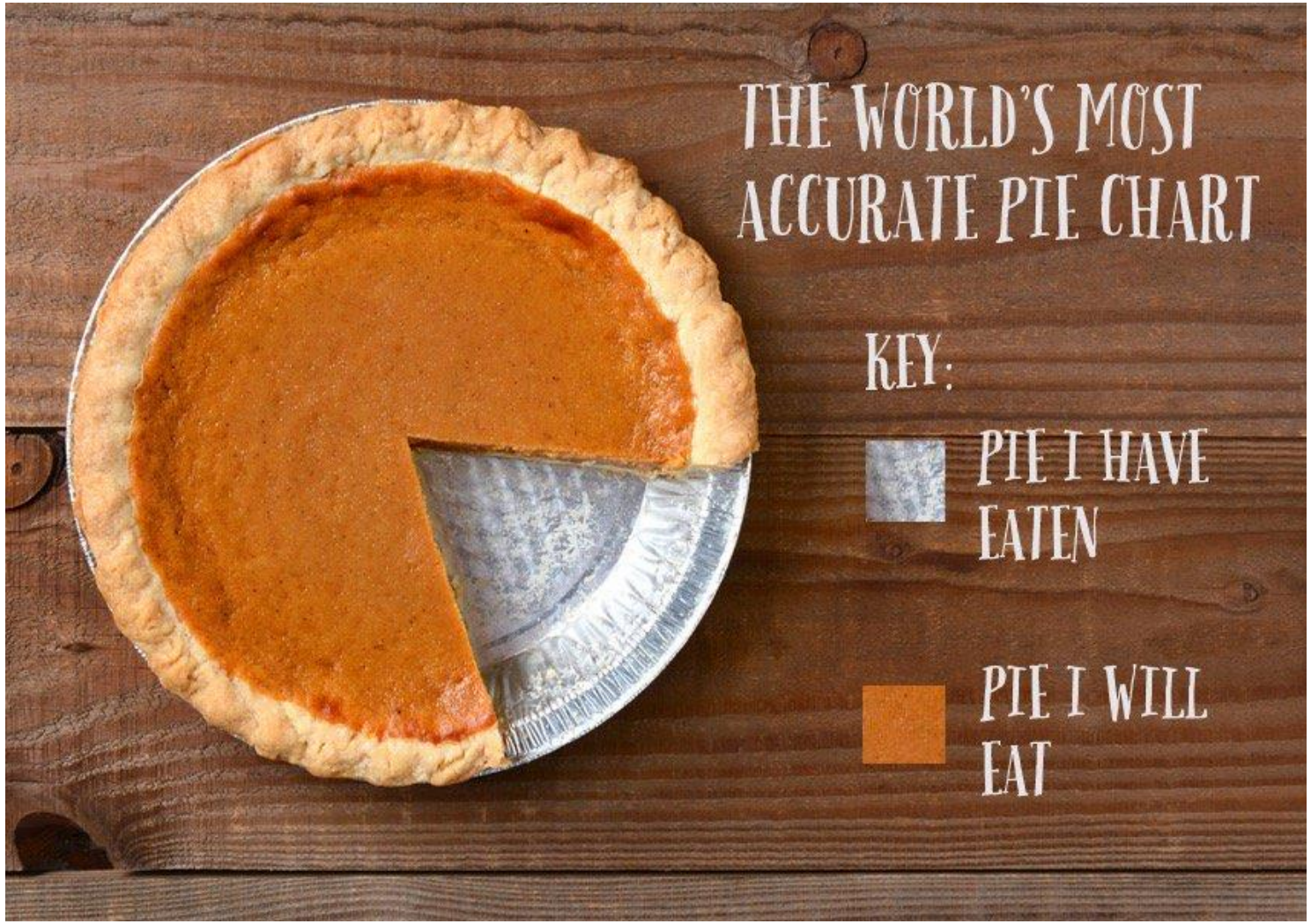
Which representation is more accurate?





Source: CS109 Stanford's Data Science Course

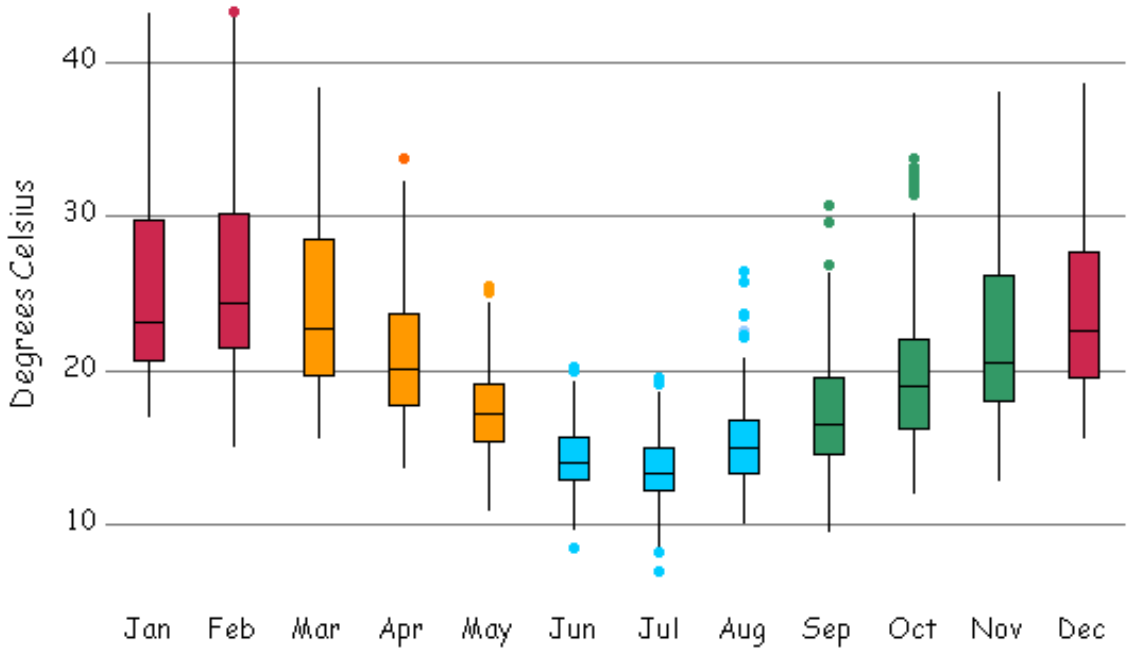
Pie Charts


THE WORLD'S MOST ACCURATE PIE CHART
KEY:
PIE I HAVE EATEN
PIE I WILL EAT

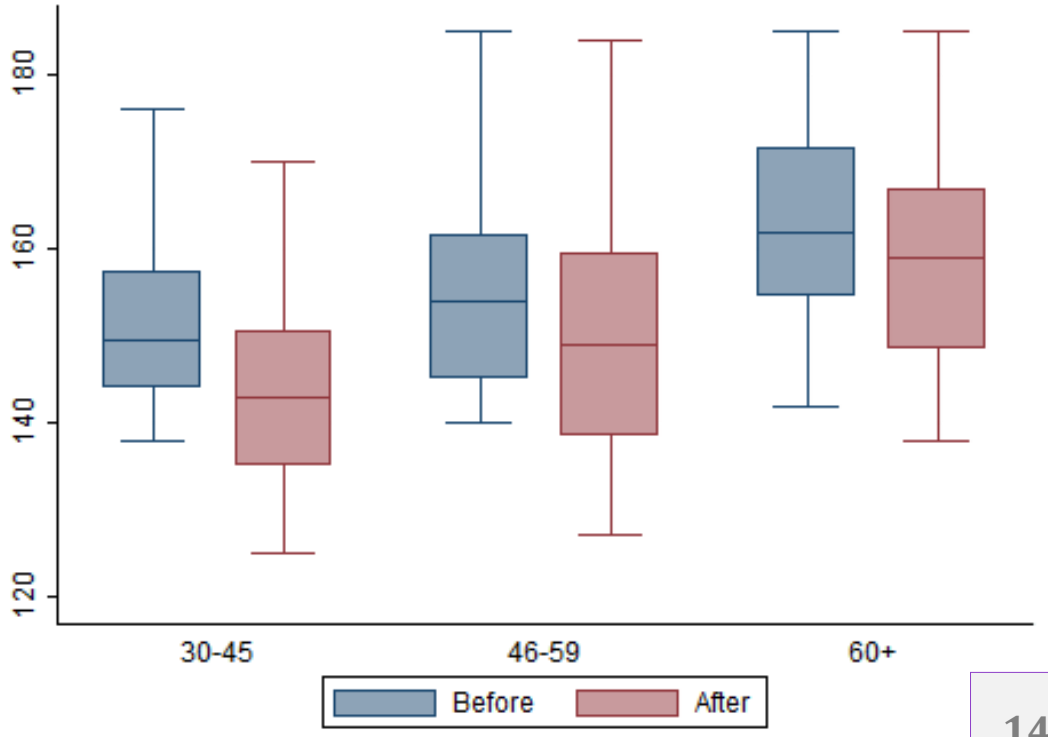
Boxplots

- A boxplot is a simplified visualization to compare a quantitative variable across groups
  - highlights range, quartiles, median and any outliers present in a data set

Maximum daily temperature in Melbourne

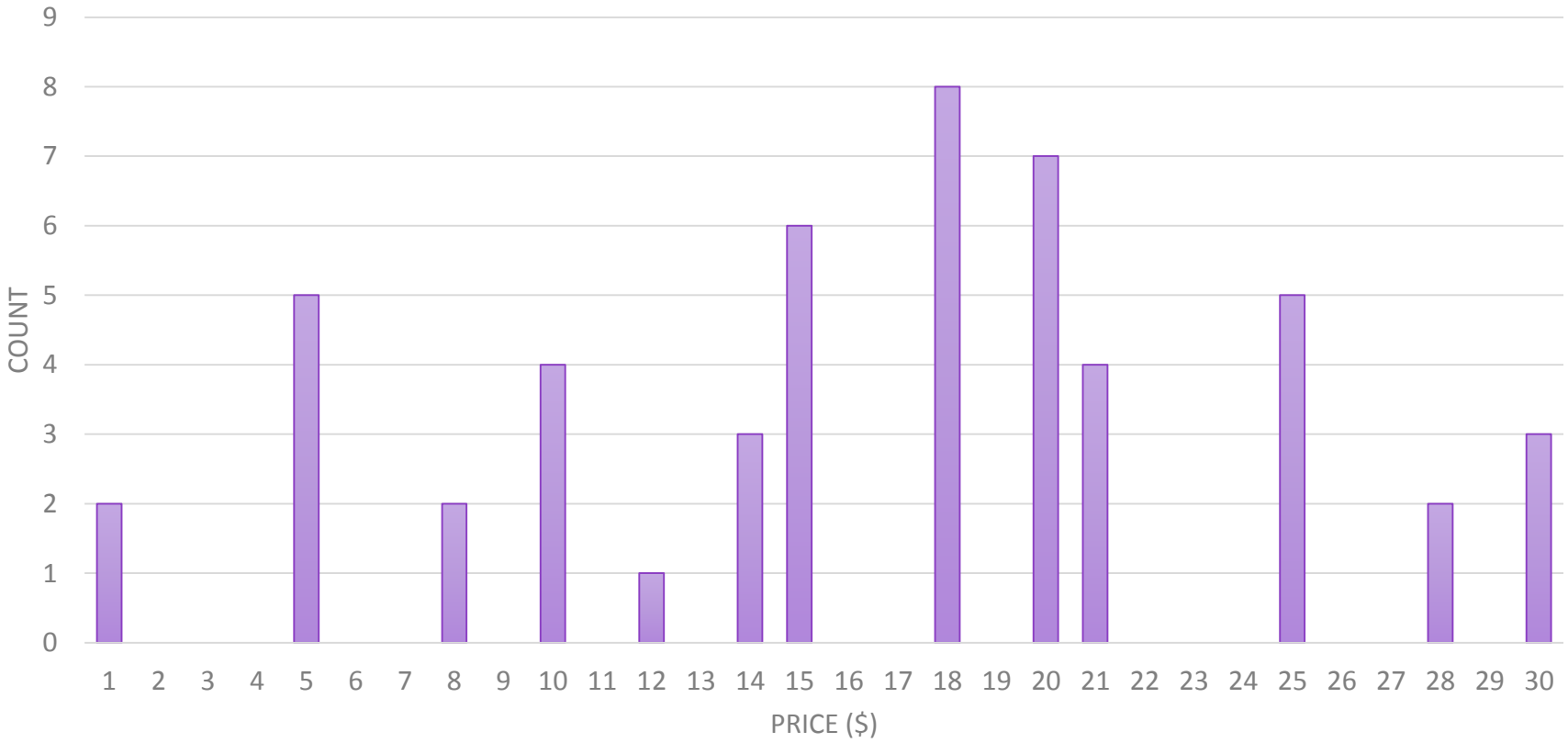


Effects of a new medication to manage high blood pressure



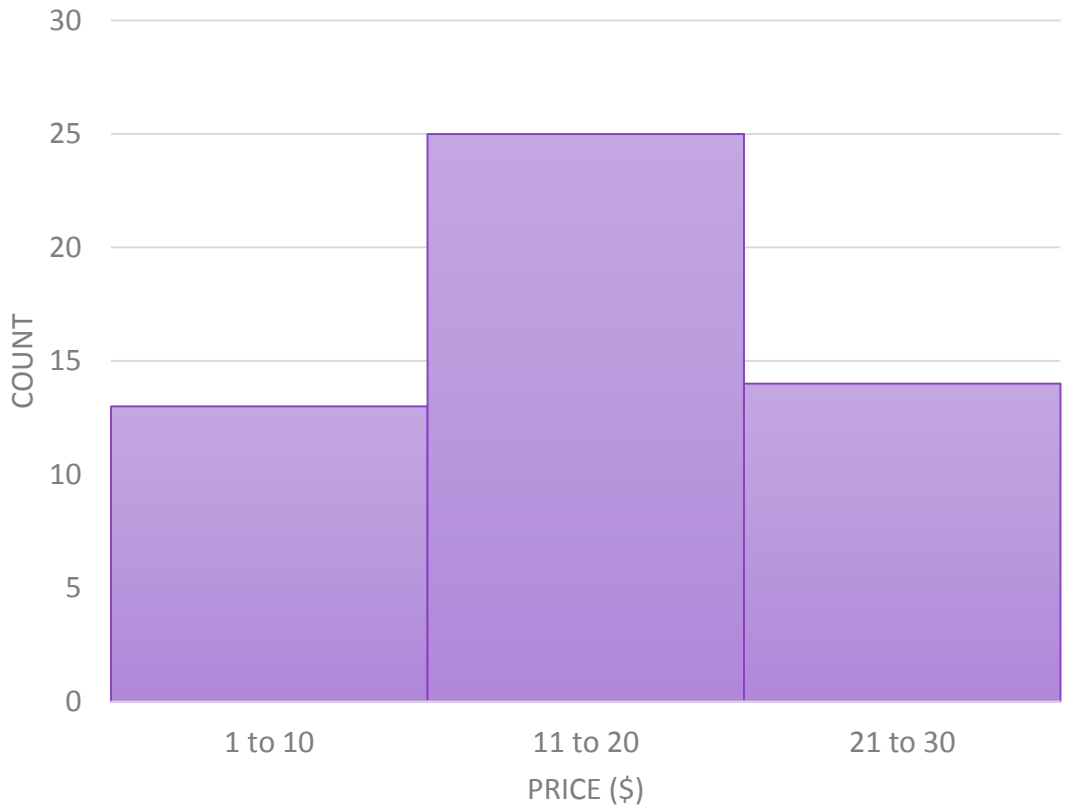
Histograms

- Original observations

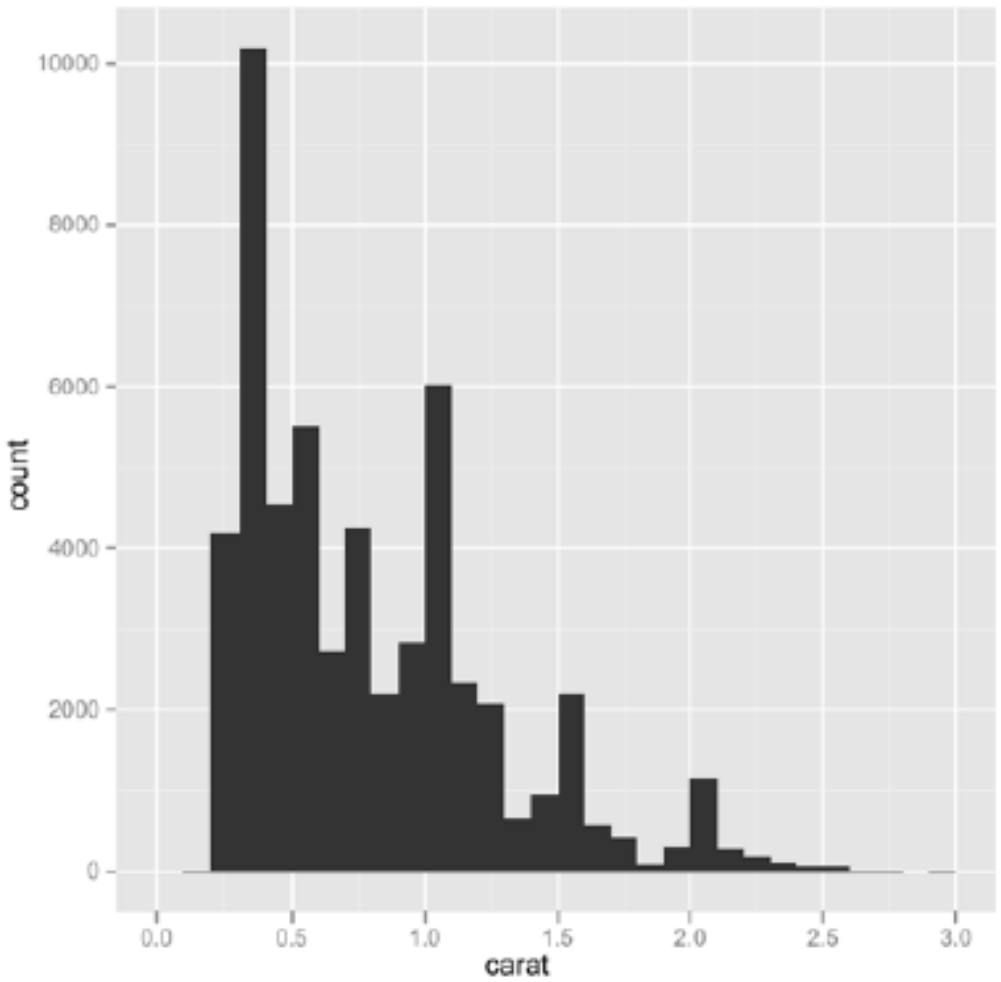


Histograms

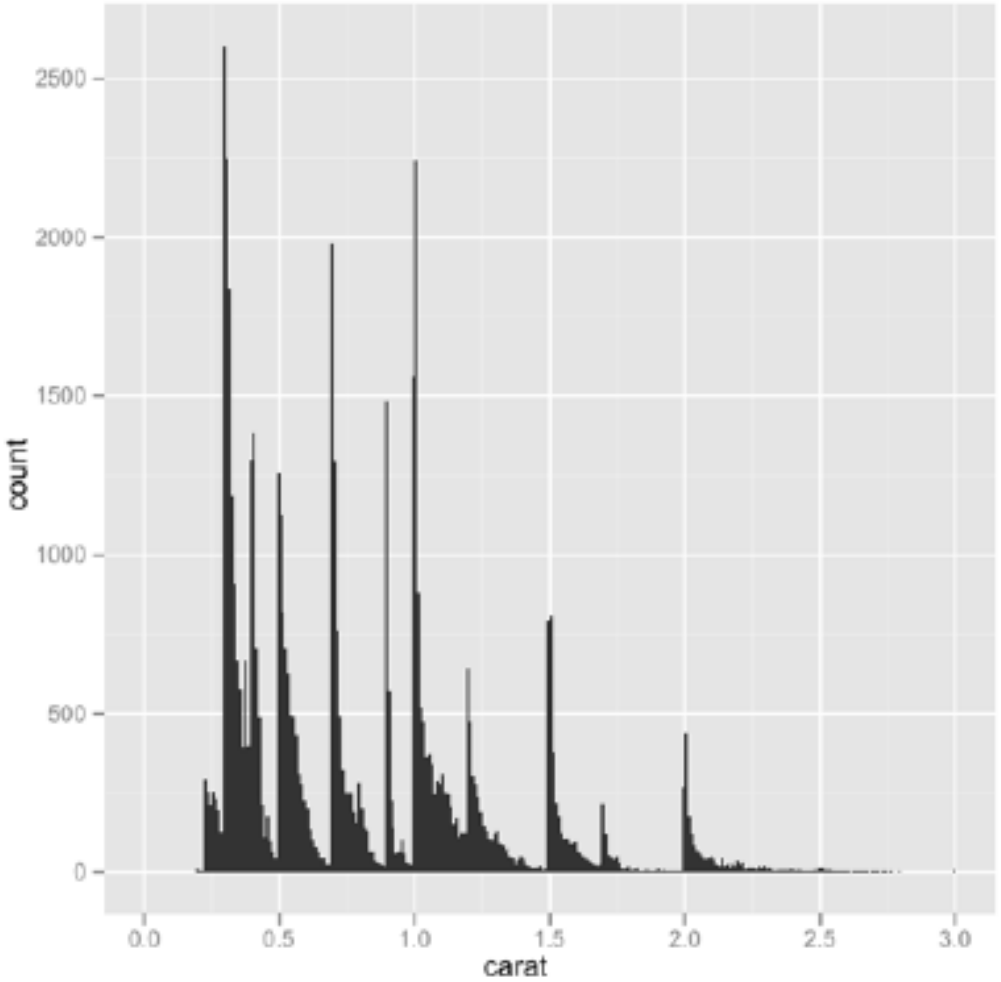
- To reduce further → equal-width buckets ($10 range)



Bin Width



binwidth = 0.1

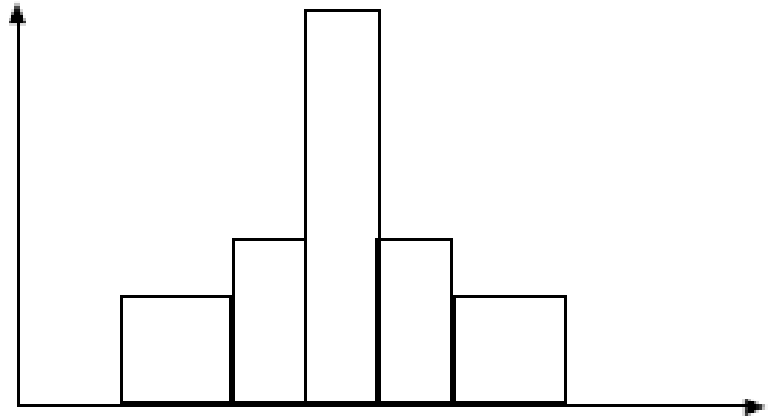


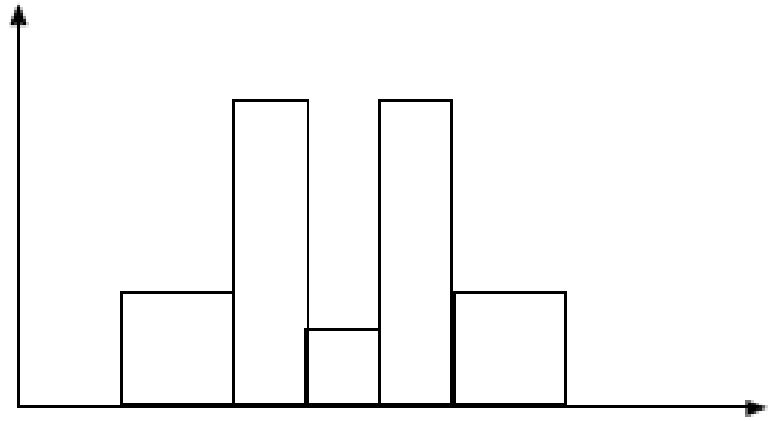
binwidth = 0.01

Source: CS109 Stanford's Data Science Course

Histograms vs. Boxplots

- The two histograms on the right have the same boxplot representation:
  - Same min, same Q1, Q2, Q3, and same max
- But they have different histogram distributions!





Anscombe's Quartet

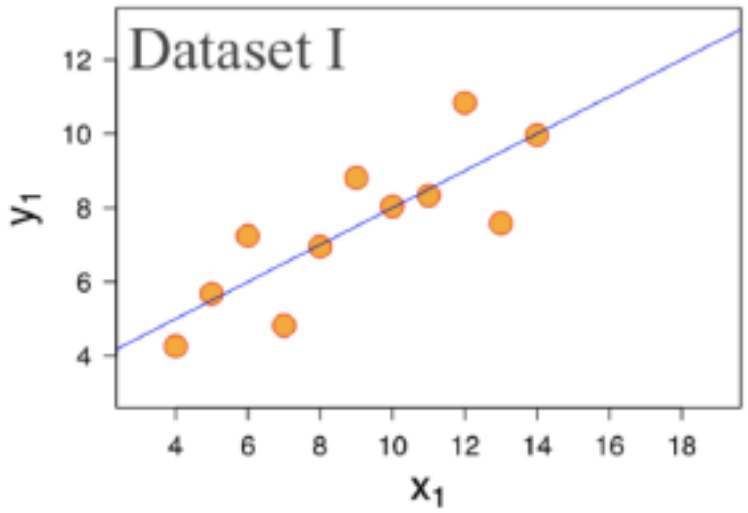
- Four data sets that have identical summary statistics

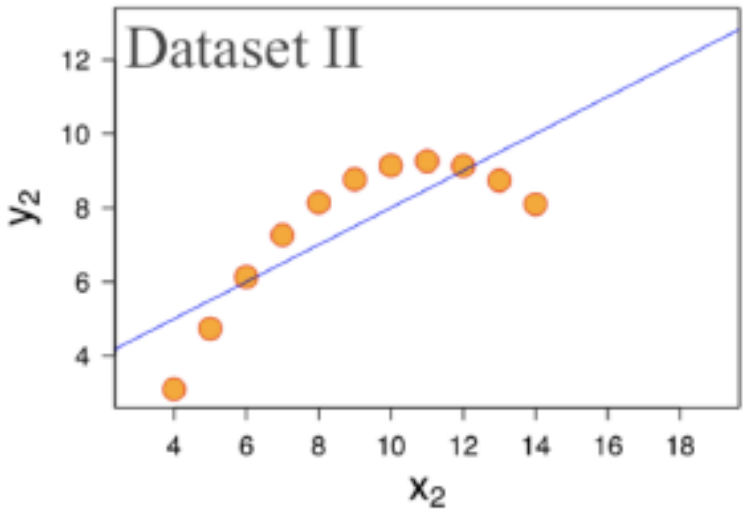
	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

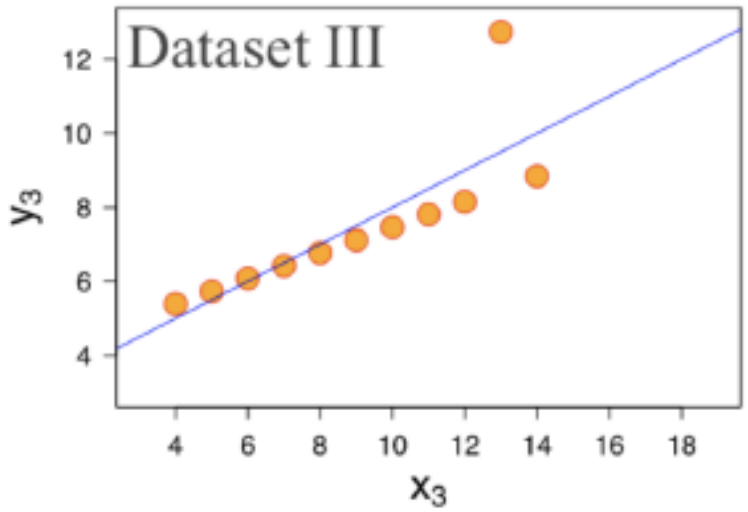
Source: CS109 Stanford's Data Science Course

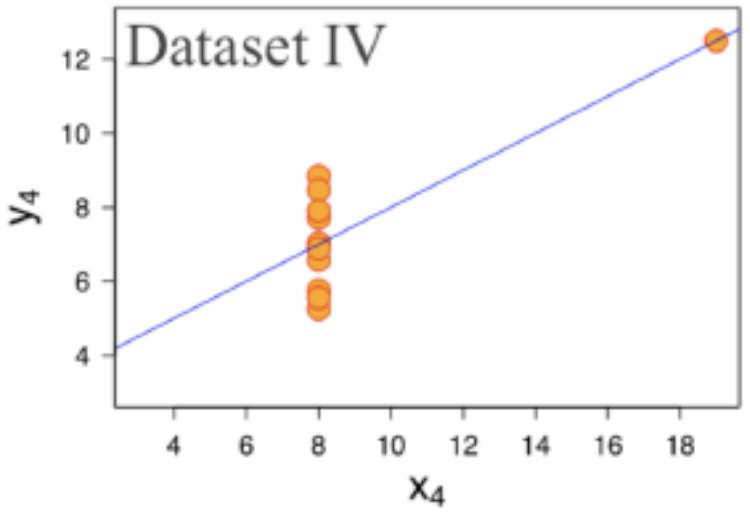
Anscombe's Quartet

- Four data sets that have identical summary statistics
- **Summary statistics don't tell us how the datasets differ**



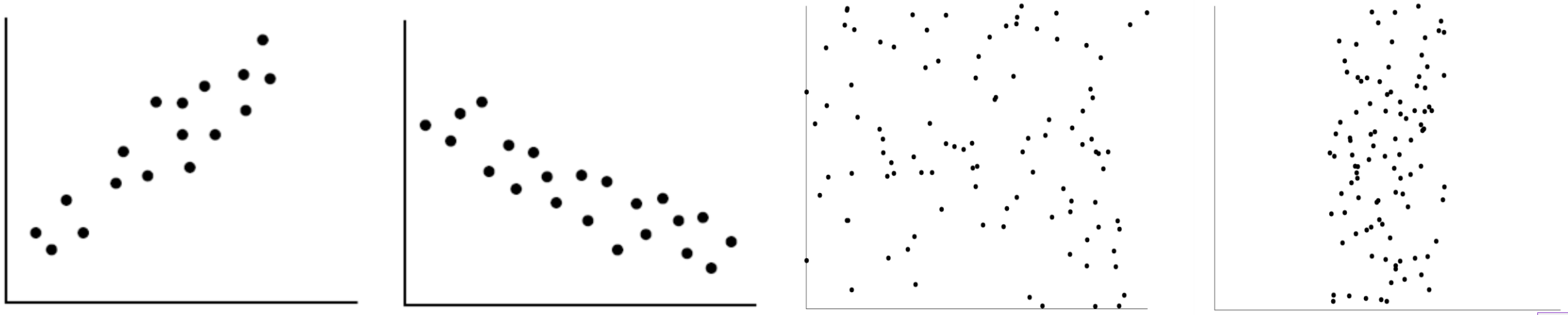






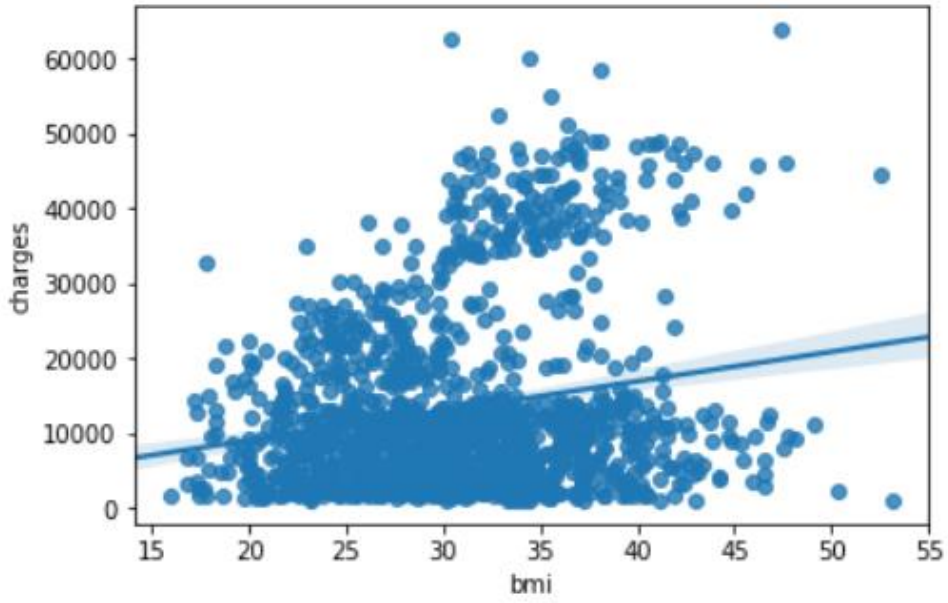
Scatter Plots

- A scatter plot is a way to visualize how multi-dimensional data are distributed across certain values
- A scatter plot is also a way to visualize the relationship between two different attributes of multi-dimensional data

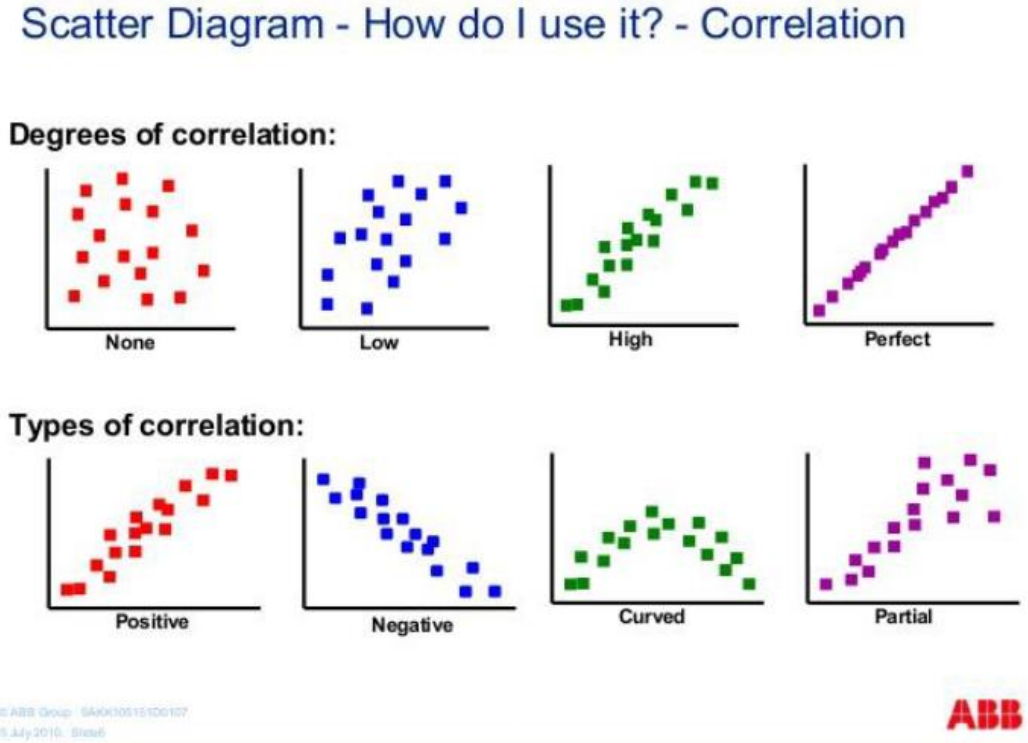


Scatter Plots and Correlations

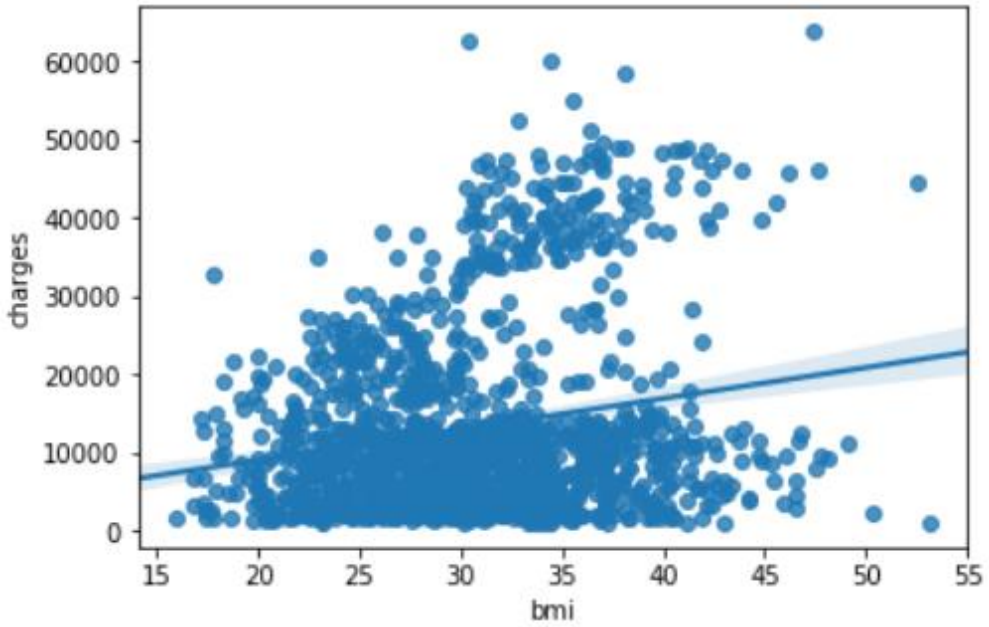
- Each pair of attribute values is treated as a pair of coordinates and plotted as points in plane
  - Attributes are *correlated* if one attribute implies the other
  - *positive*, *negative*, or *null* (uncorrelated)



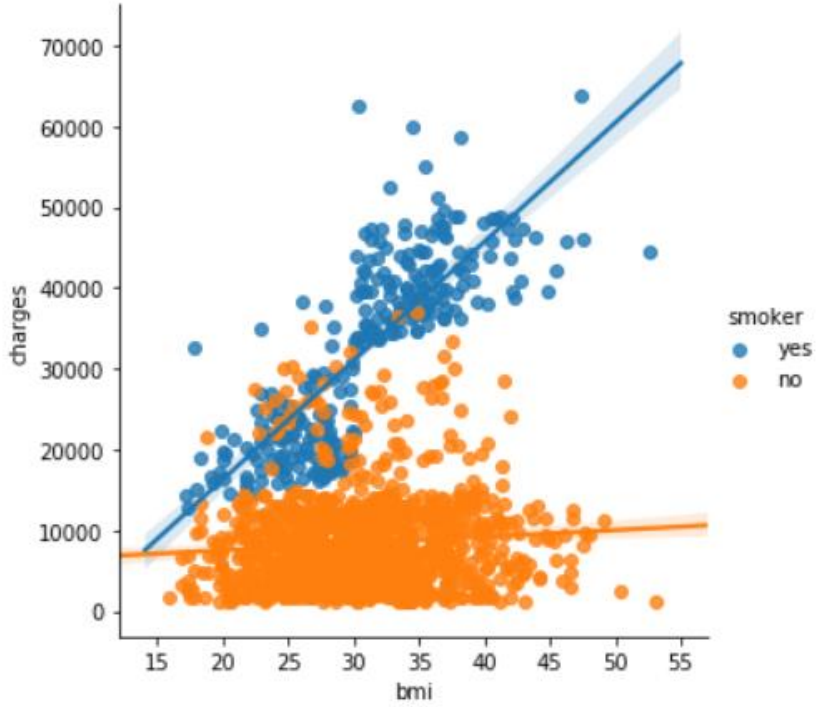
Scatter plot of BMI and insurance charges. Source: Kaggle Data Visualization Micro Course



Scatter Plots and Correlations



Scatter plot of BMI and insurance charges



Scatter plot of BMI, insurance charges, and smoking behavior

Source: Kaggle Data Visualization Micro Course

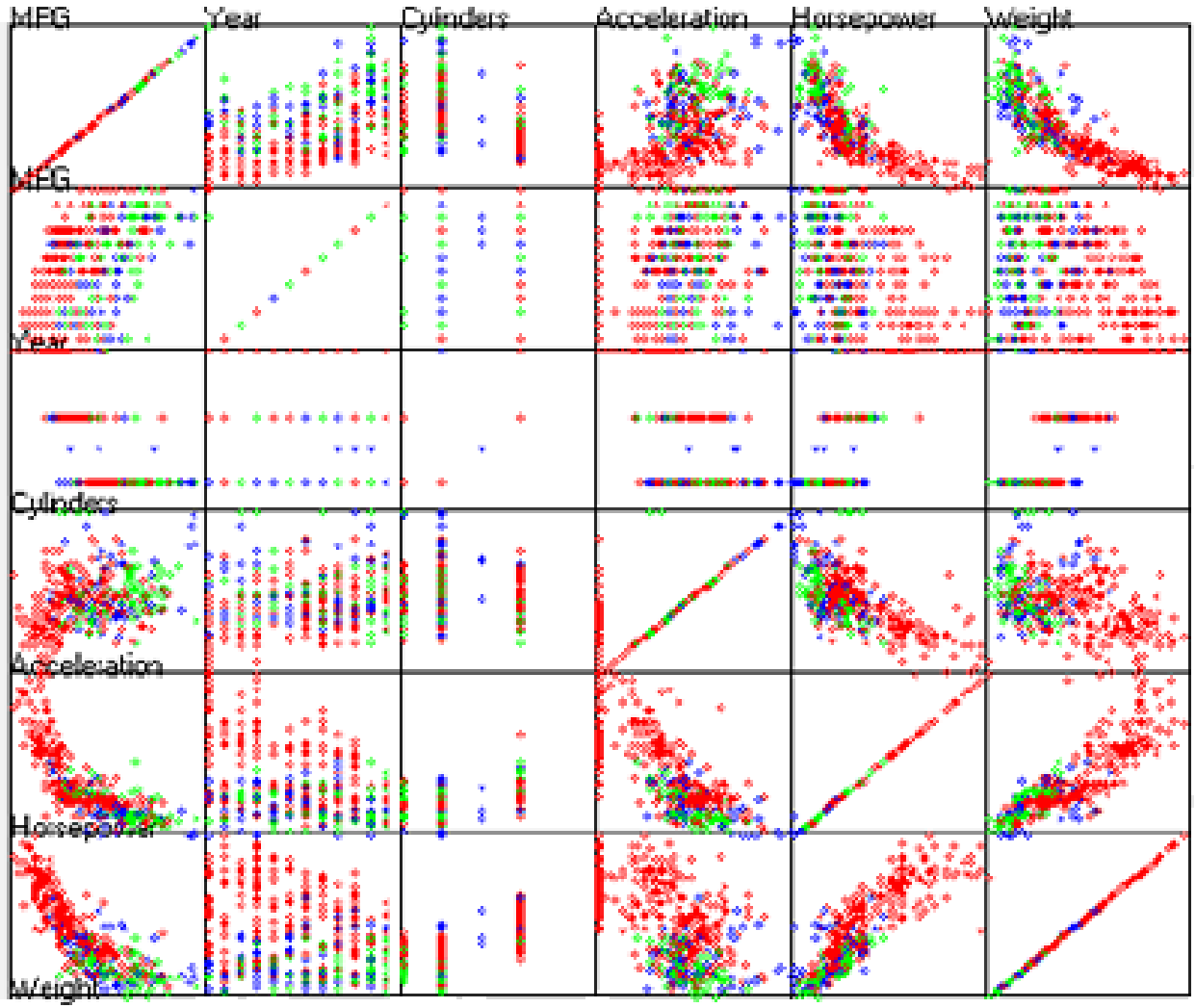
Visualizing High-dimensional Data

Often your dataset seem too complex to visualize:

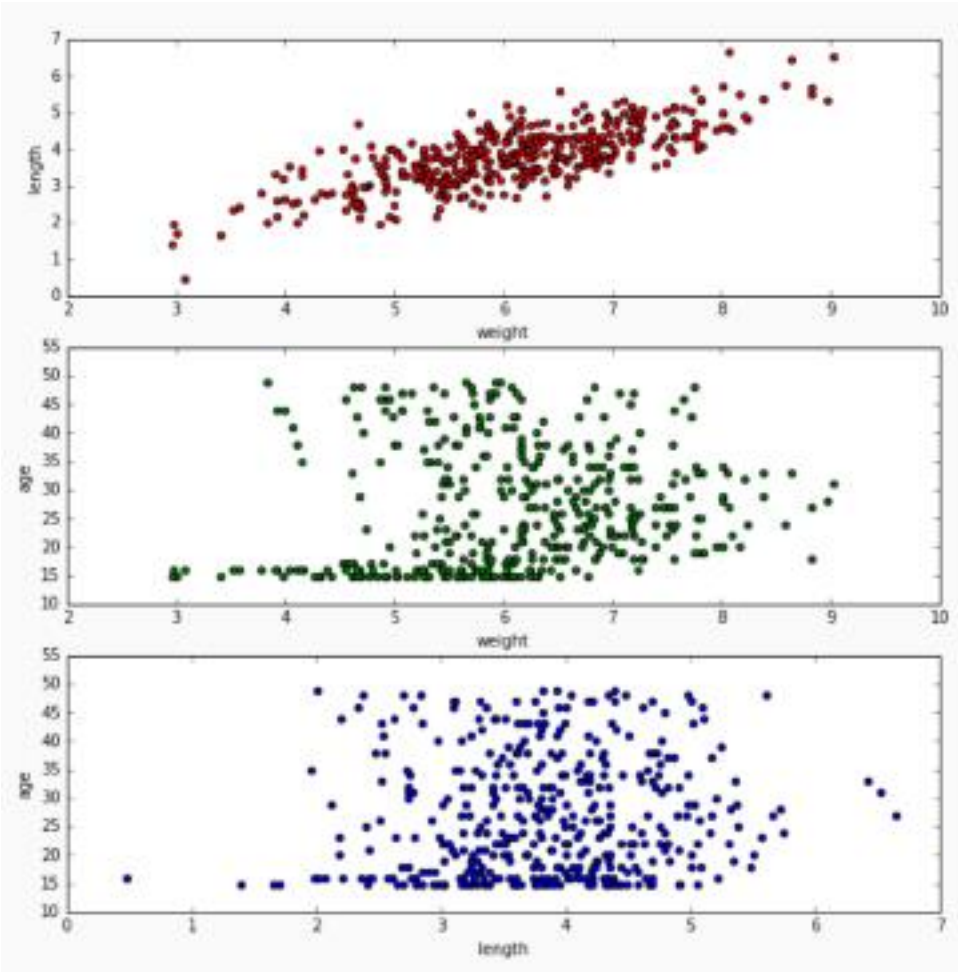
- Data is too high dimensional (how do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like “Cat or No”?)
- When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful!

Scatter Plot Matrix

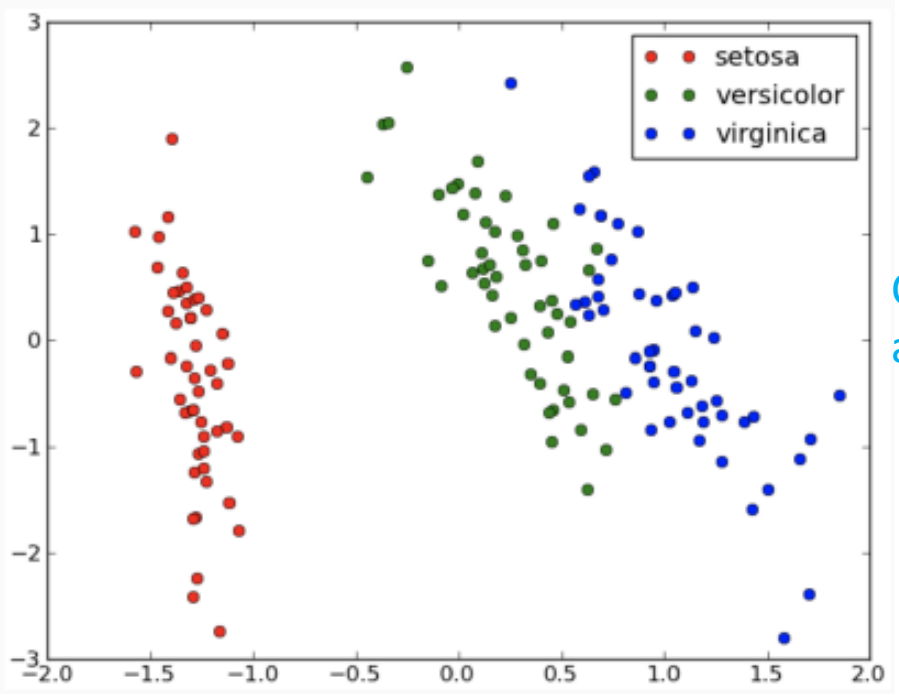
- For attributes more than 2, you may use a *scatter plot matrix*
- All the pairwise scatter plots of variables on a single page in a matrix format



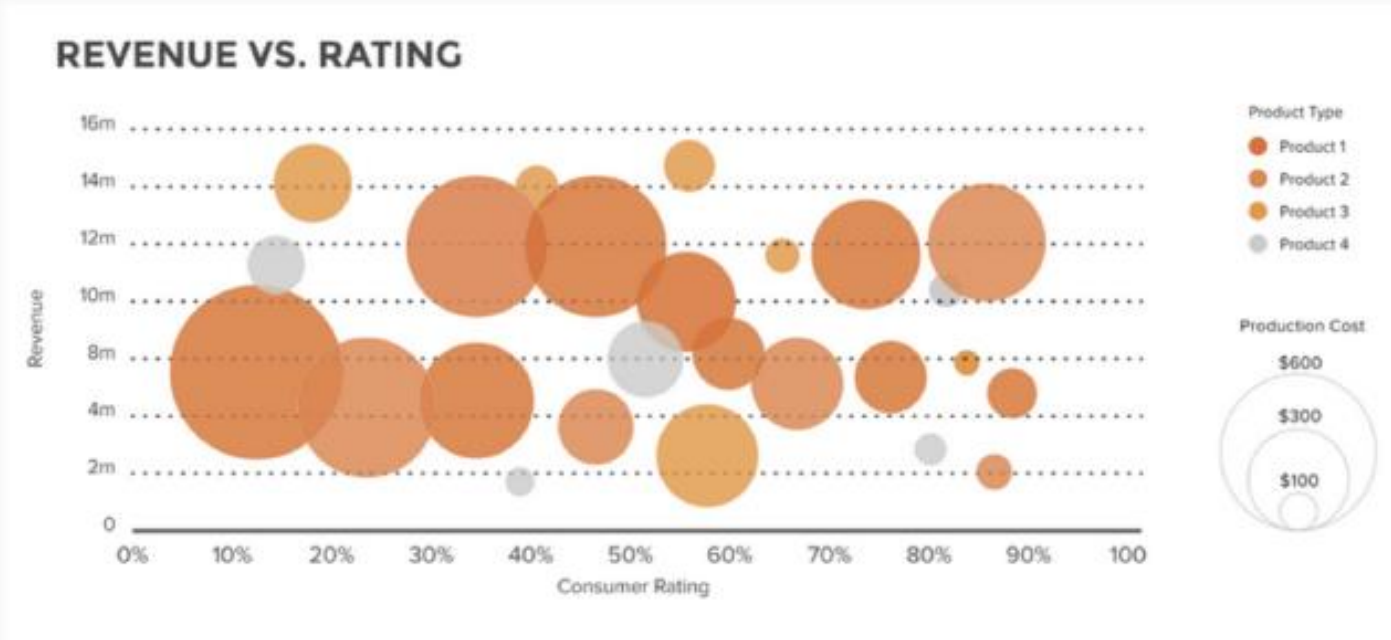
Colors and Bubbles



A quantitative attribute can be encoded by size in a bubble chart

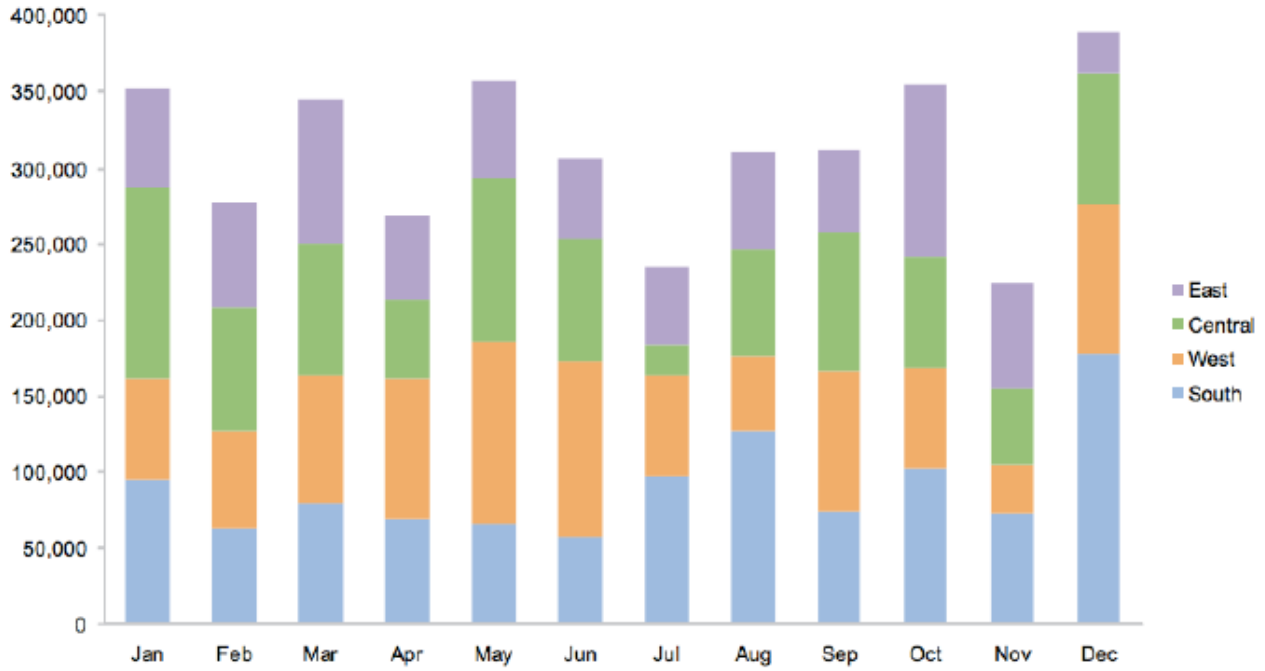


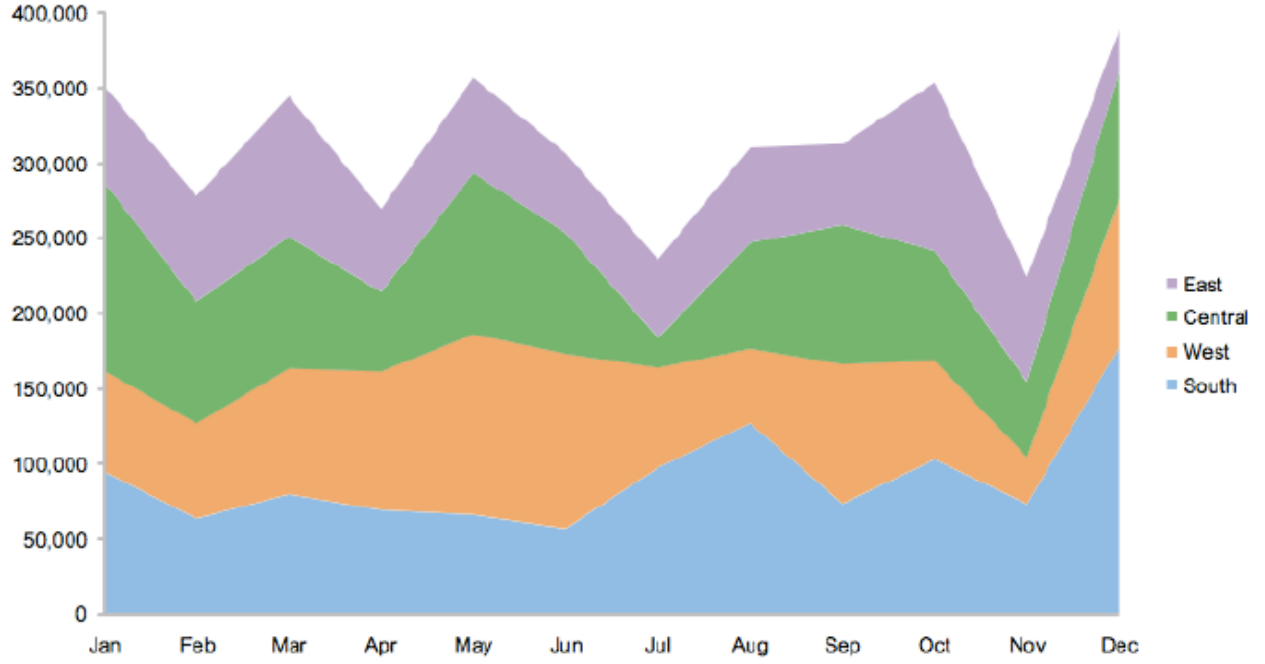
Color coding a categorical attribute can be effective



Stacked Bar and Area Charts

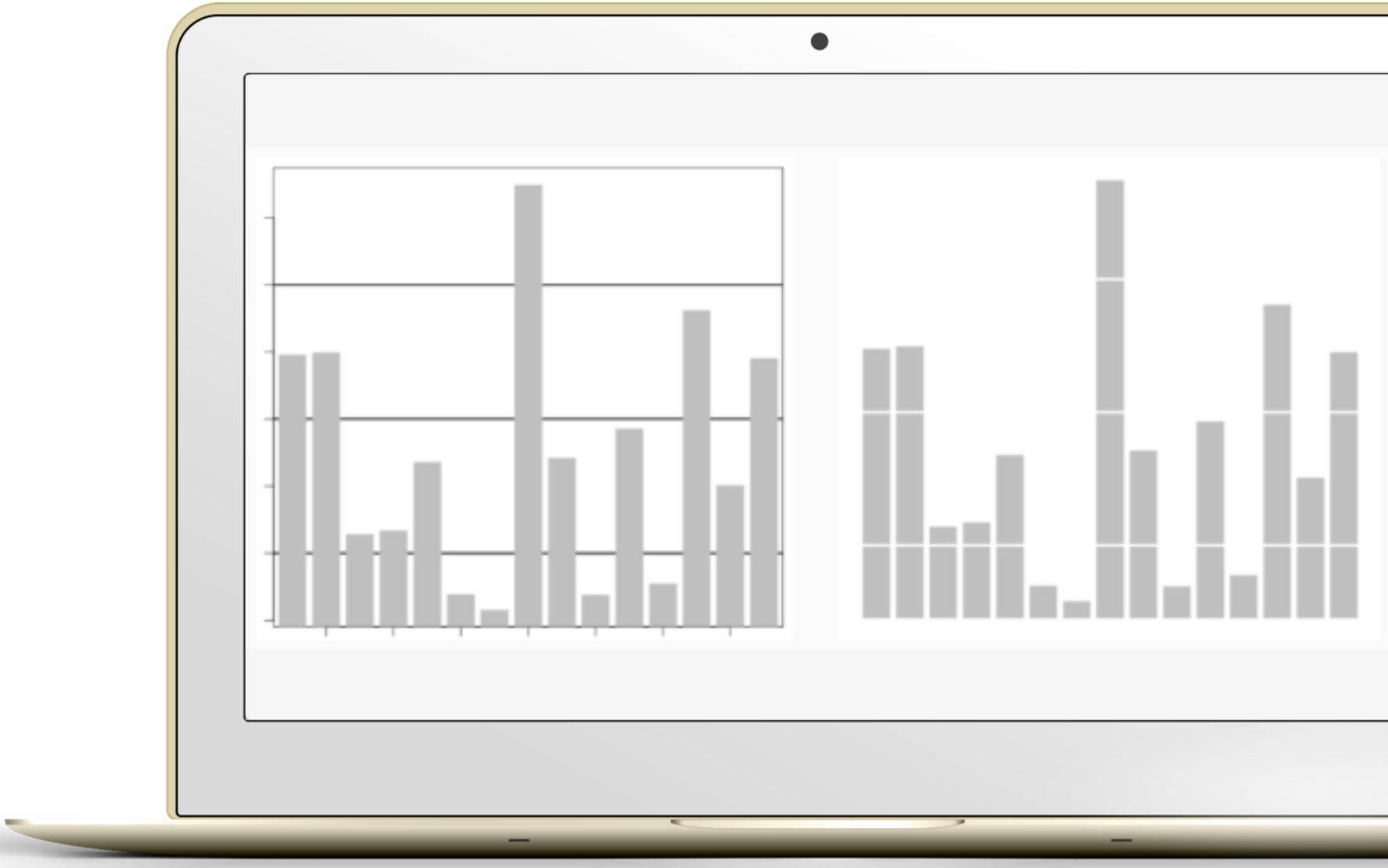
- A stacked bar or area graph is a way to visualize the composition of a group as it changes over time (or some other quantitative variable)





Principles of Visualizations

Maximize data to ink ratio: show the data

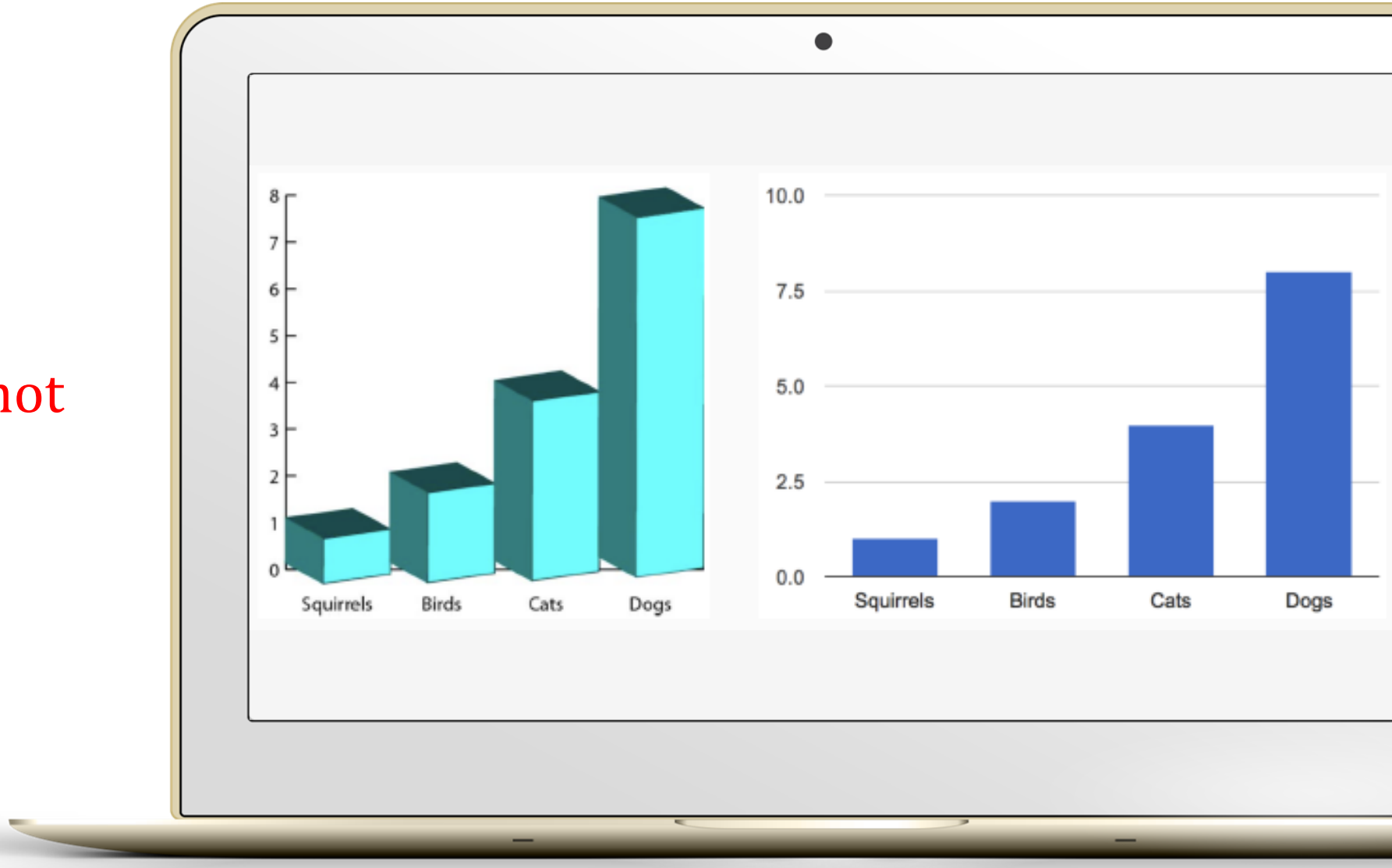


Principles of Visualizations

Don’t lie with scale

Principles of Visualizations

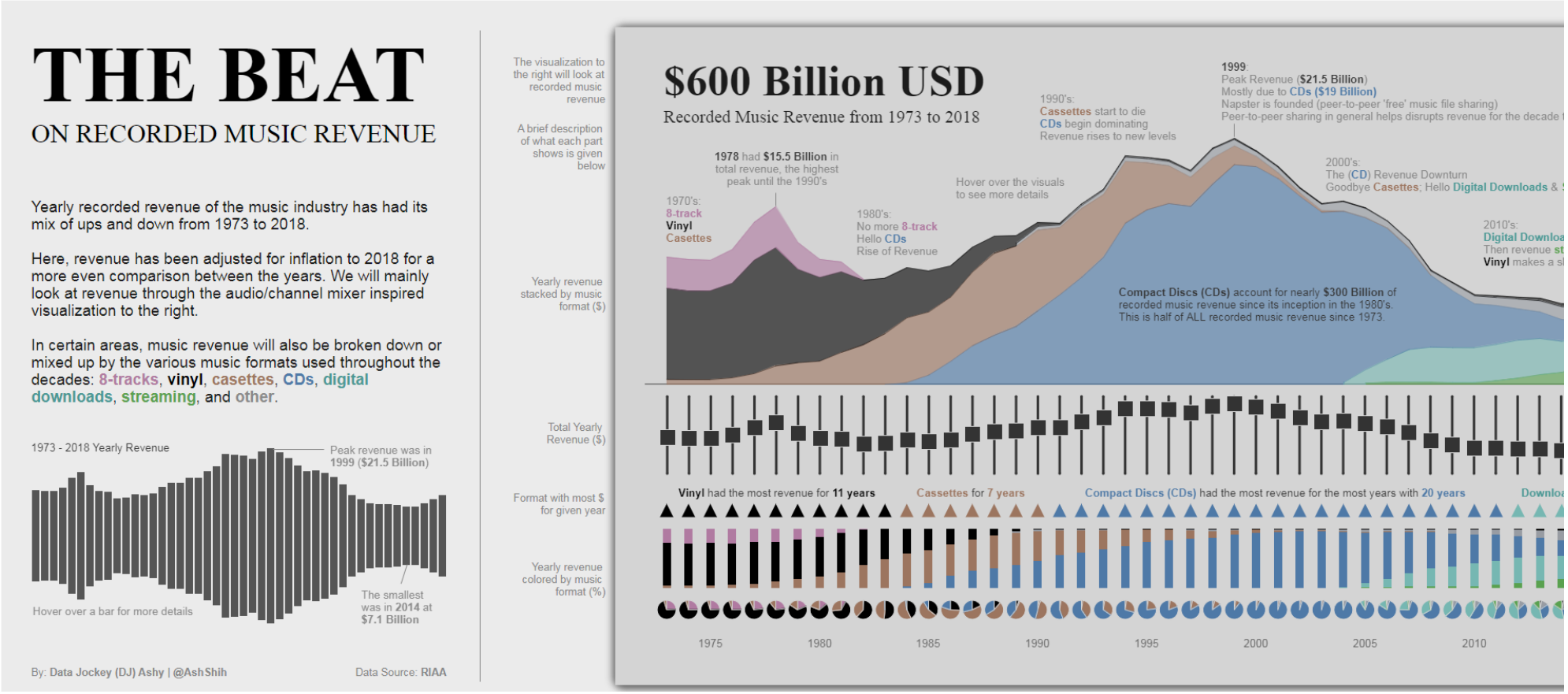
Minimize chart-junk: show data variation, not design variation



Take a look at this interactive visualization:

Yearly recorded revenue of the music industry from 1973 to 2018

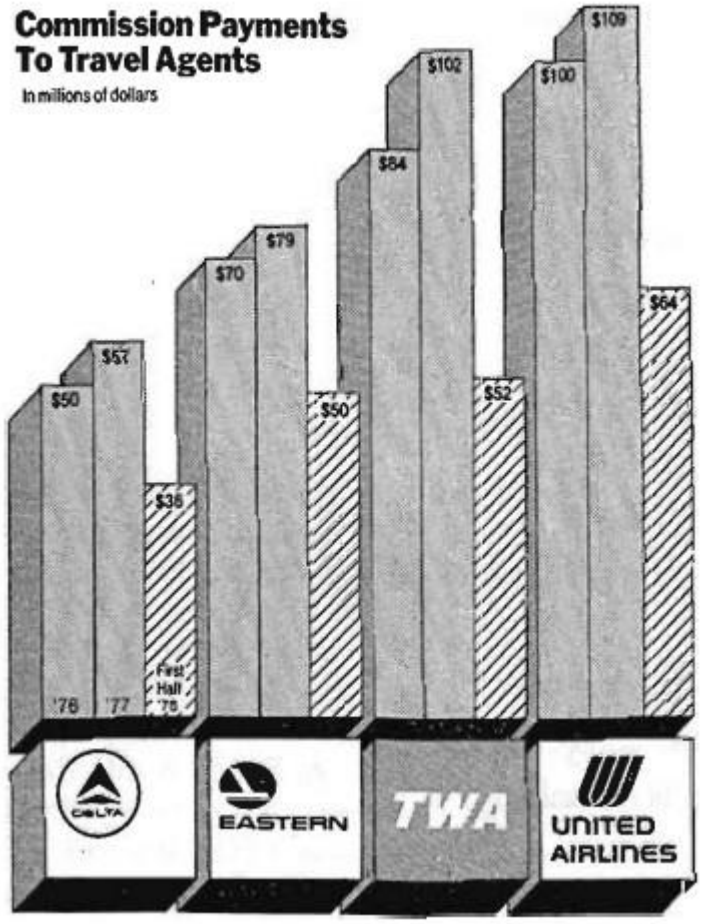
https://public.tableau.com/profile/ash.s.#!/vizhome/TheMusicIndustryBeat/TheMusicIndustryBeat



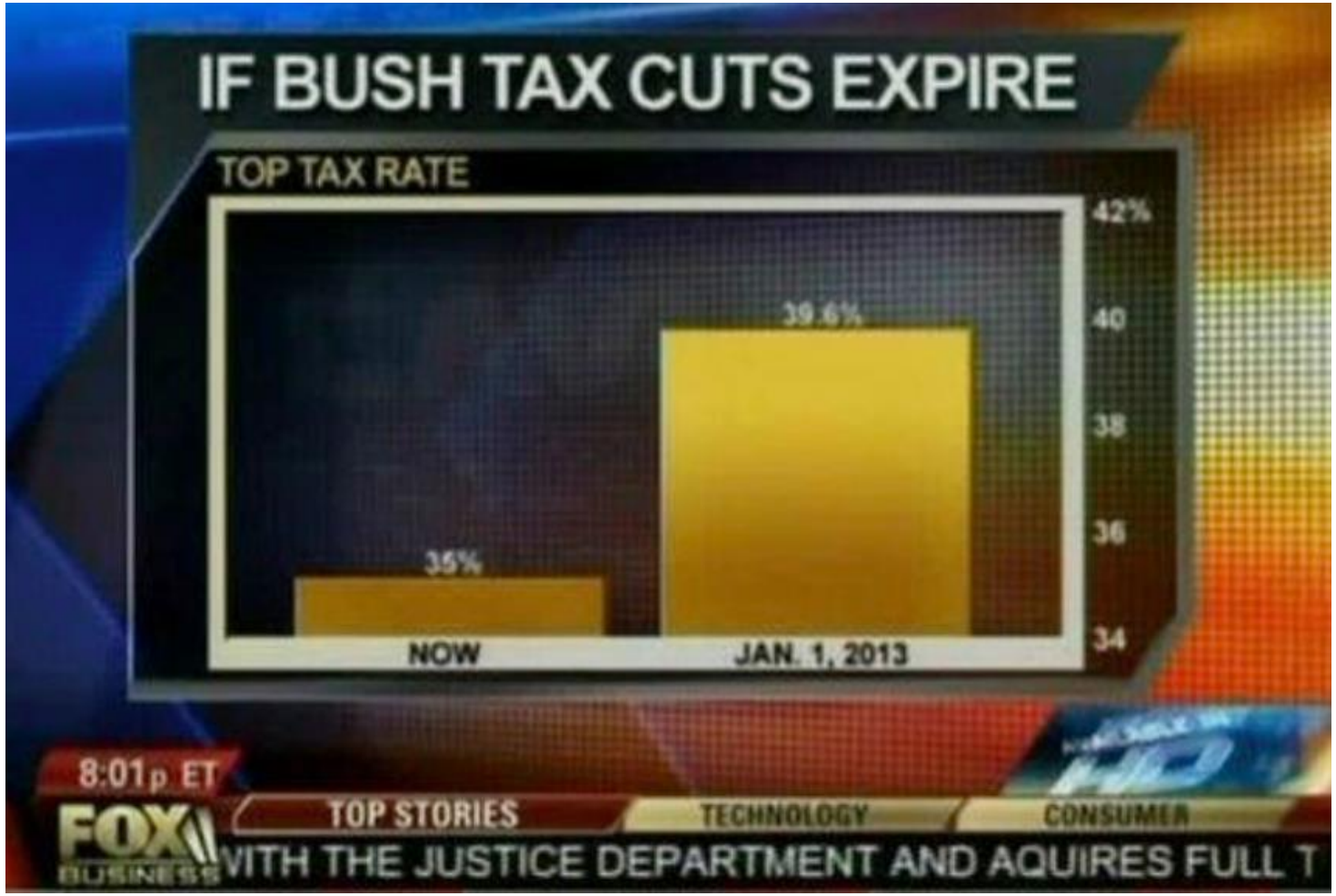
Let's Critique!

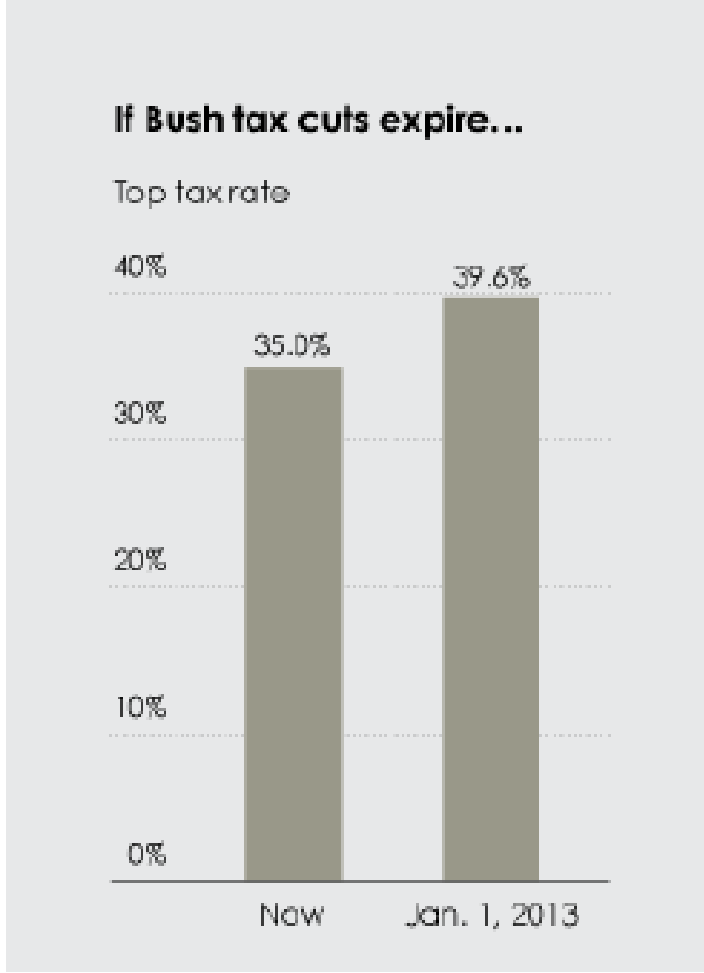
What do you think of the next visual representations?

What do you think of this visual representation?



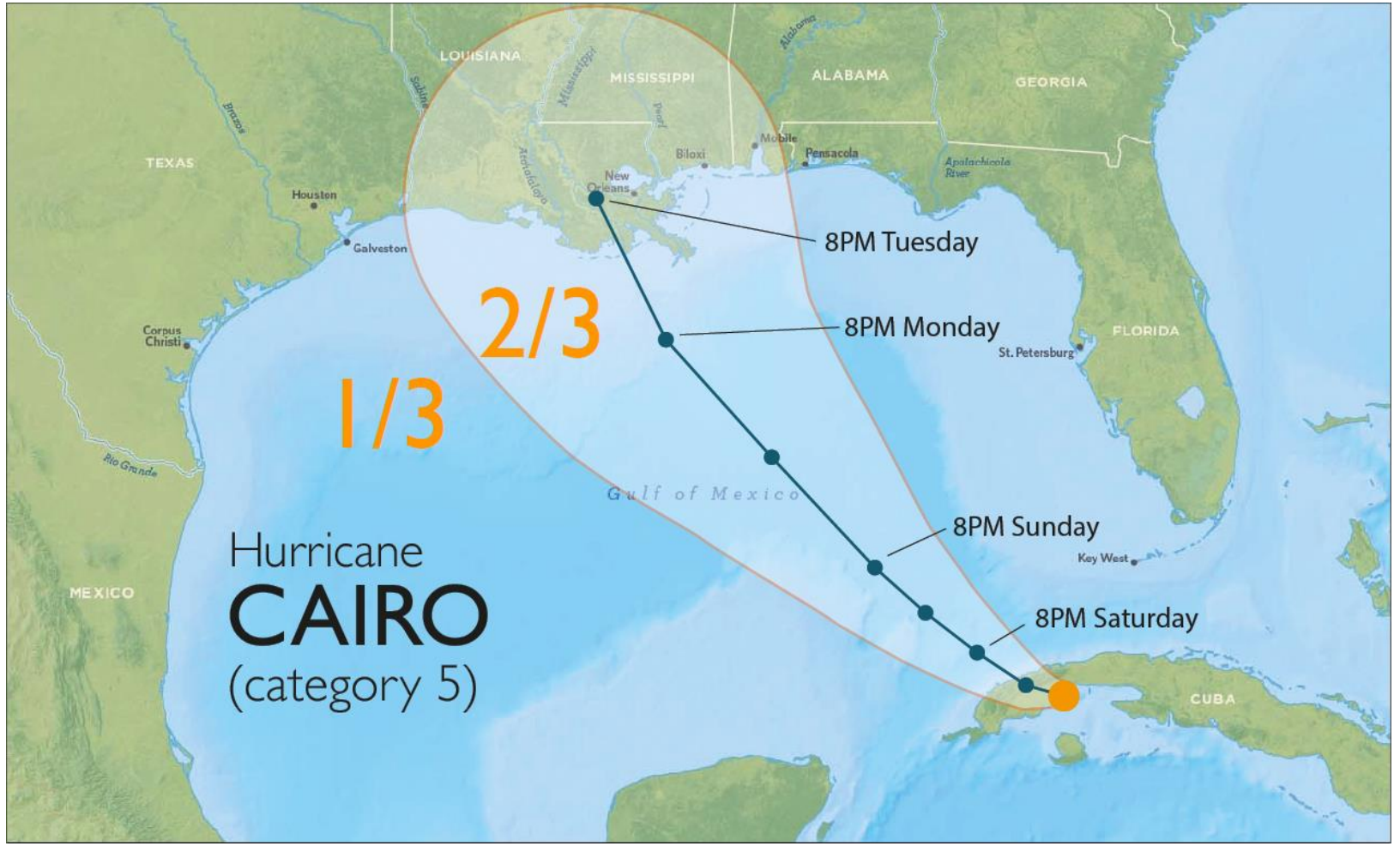
What do you think of this visual representation?





What do you think of this visual representation?

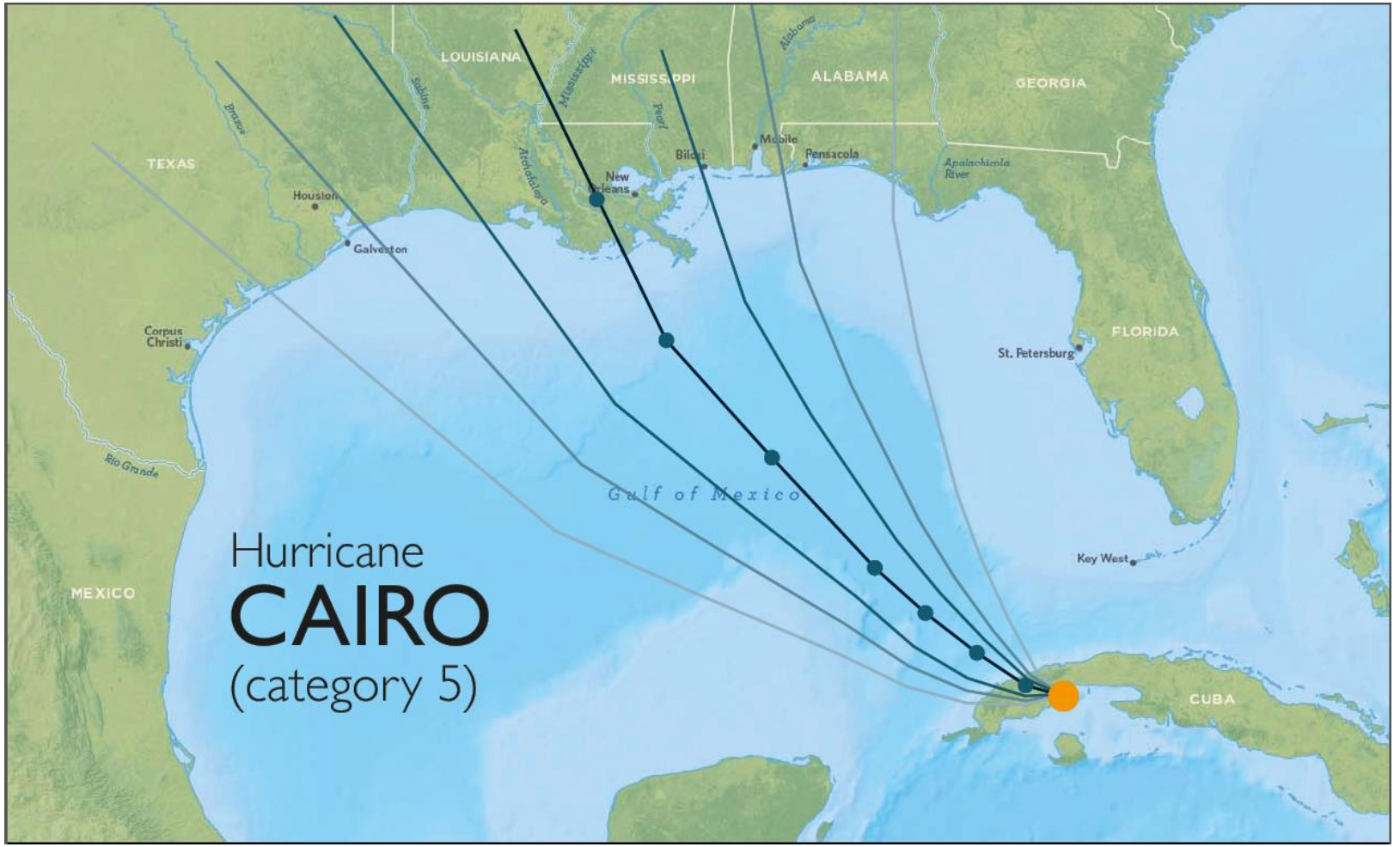
What do you think of this visual representation?



What non-scientists are not aware of (cone is just 66% probability)

Source: CS109 Stanford's Data Science Course

What do you think of this visual representation?

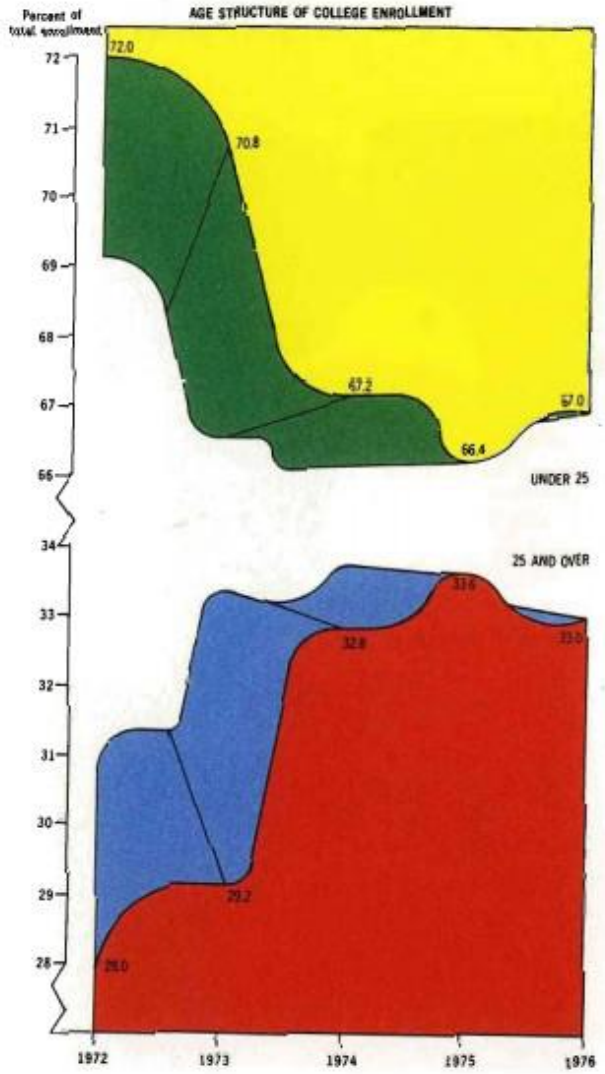


Incorporating uncertainty

Source: CS109 Stanford's Data Science Course

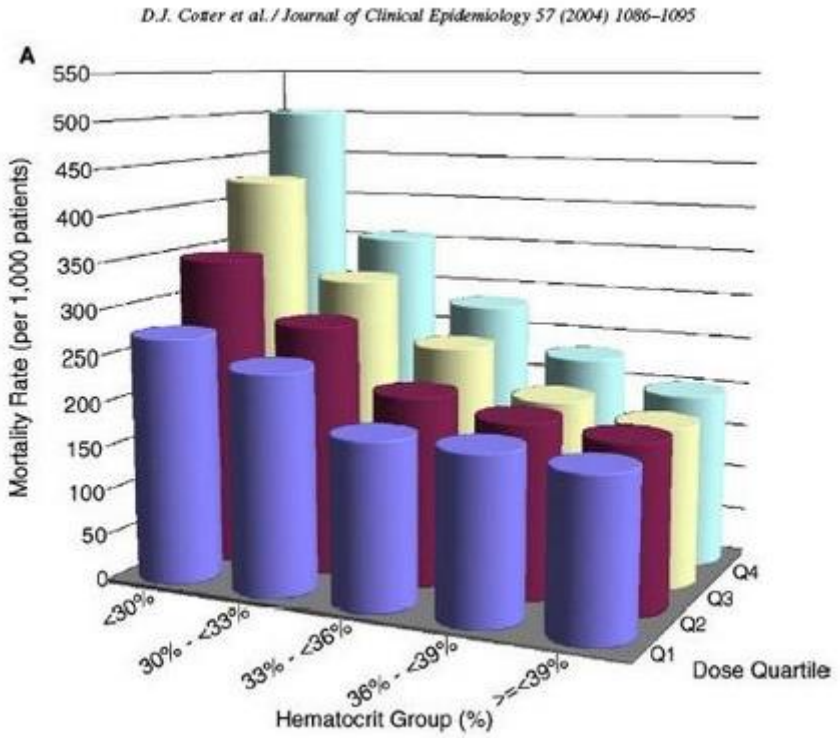
What do you think of this visual representation?

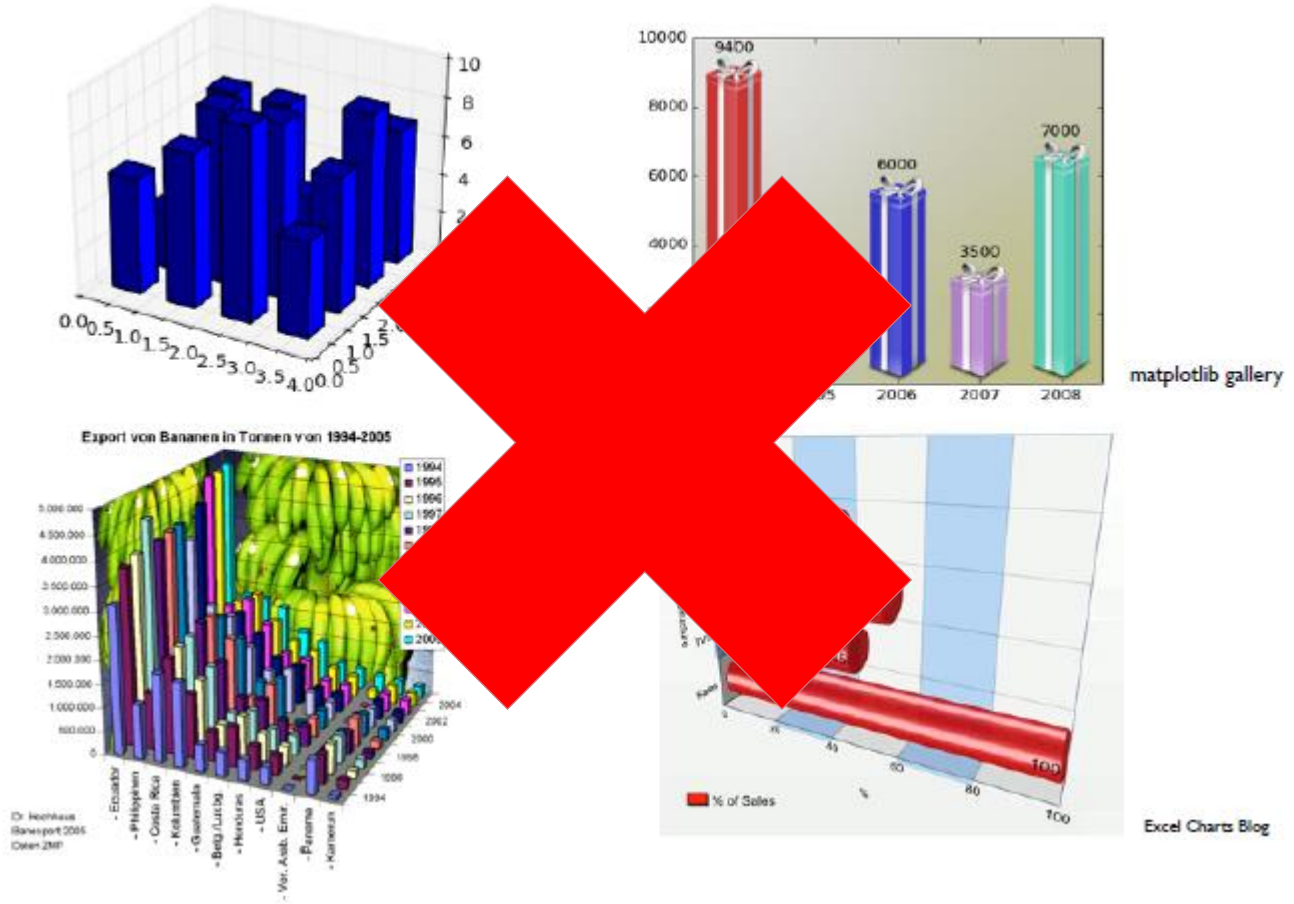
Tufte, in his seminal book "*The Visual Display of Quantitative Information*" (1983, p.118) says, "*This may well be the worst graphic ever to find its way into print.*"



Let's Play!

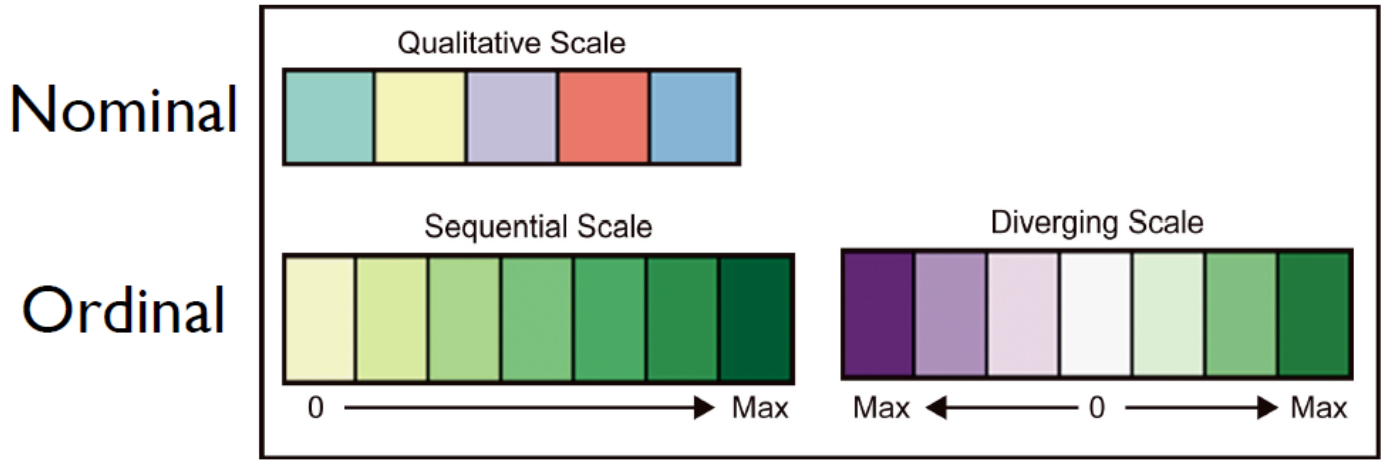
What do you think of this visual representation?





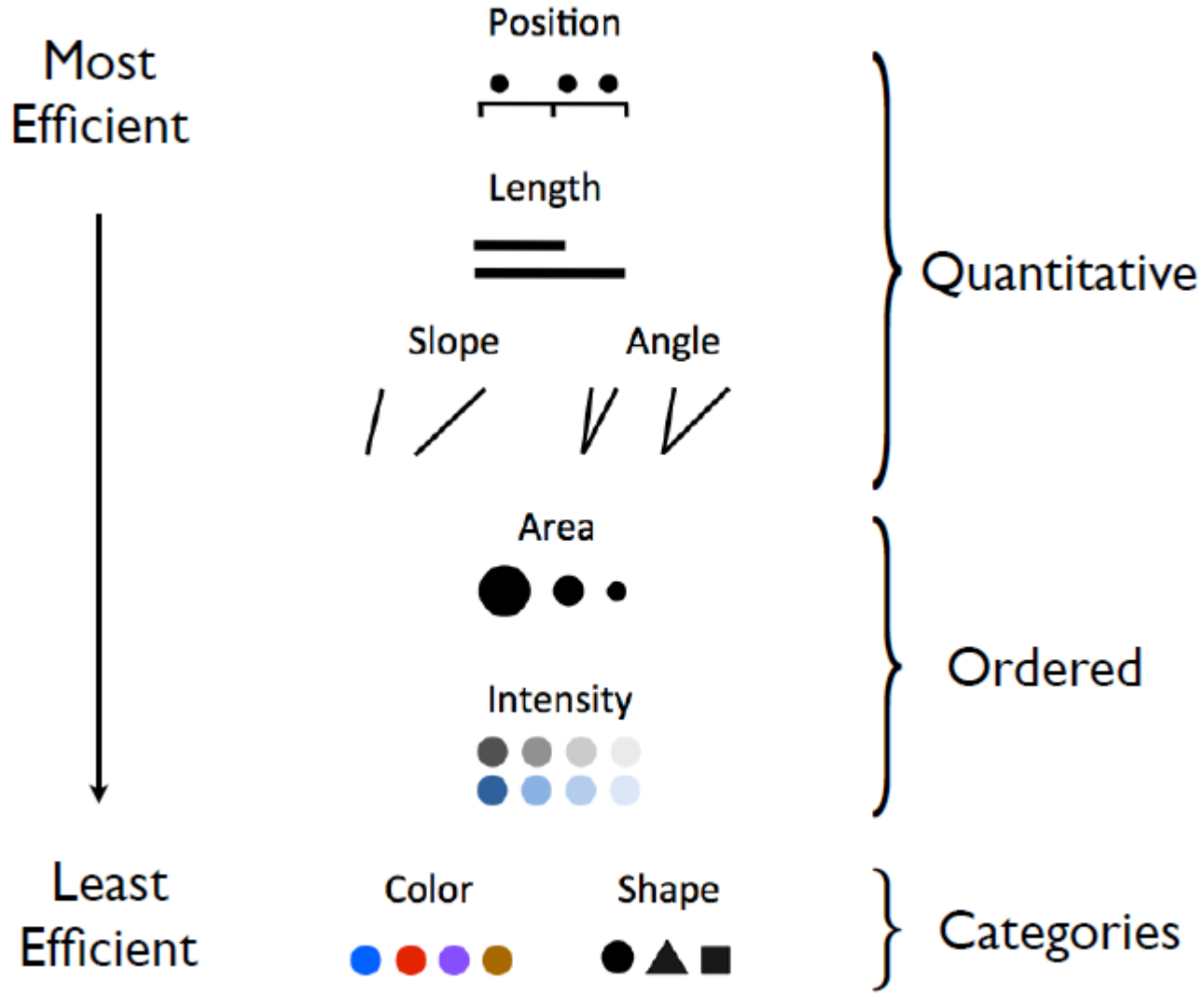
To Sum up – For Effective Visualizations:

- Have graphical integrity
  - Do not distort scale
  - Include uncertainty
  - Plot all the data
- Keep it simple
- Use the right chart
- Use color sensibly
  - Use different colors for categories (5-8 at most)
  - Use same color with varying luminance or saturation for ordinal data
    - If scale starts at 0, use sequential colors, if scale diverges from 0, use diverging scale



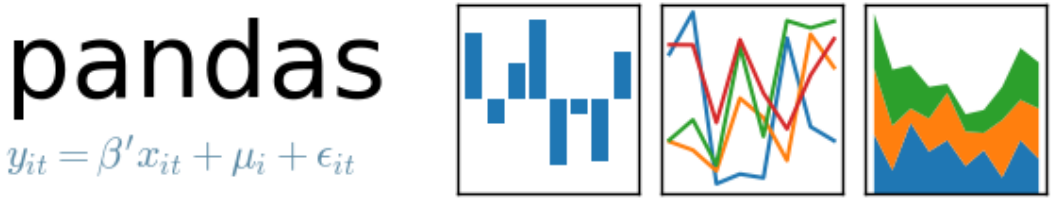
Source: CS109 Stanford's Data Science Course

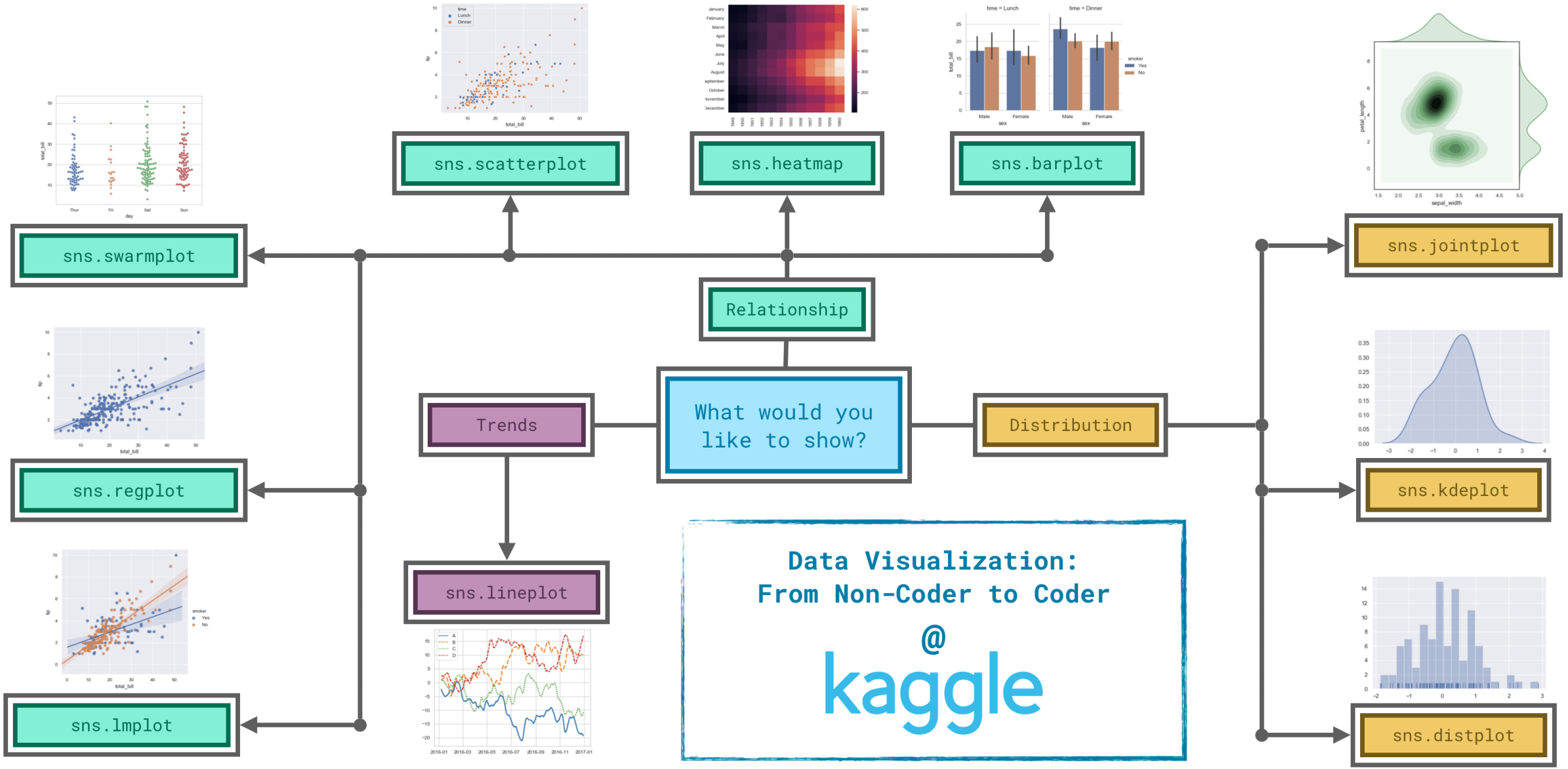
Perceptual Effectiveness



Source: CS109 Stanford's Data Science Course

Preparing for Next Week's Practice Sessions





Functions we will use in Seaborn

GUC

Thank You