UC San Diego

Discussion Section A01 & A02
Nov 2, 2022

Ashwin Mishra, TA

Jiesen Zhang, IA Lindsey Gu, IA

UC San Diego

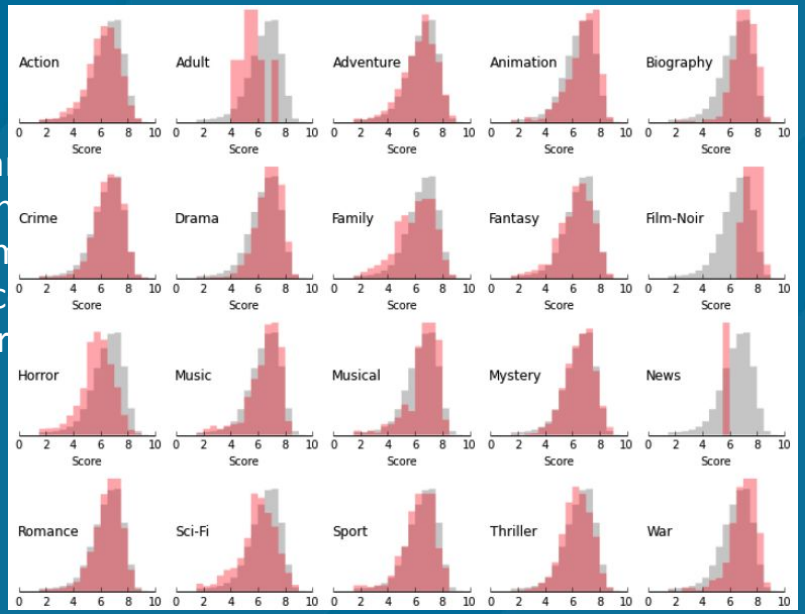
Visualization Goals

- Explanatory (Communicate)
  - Present data
  - Explain and inform
  - Provide evidence
  - Influence and persuade
- Exploratory (Analyze)
  - Explore the data
  - Assess a situation
  - Determine how to proceed
  - Decide what to do

UC San Diego

Exploratory Data Analysis (EDA)

- Exa
- Har
- Ren
- Enc
- Nor



UC San Diego

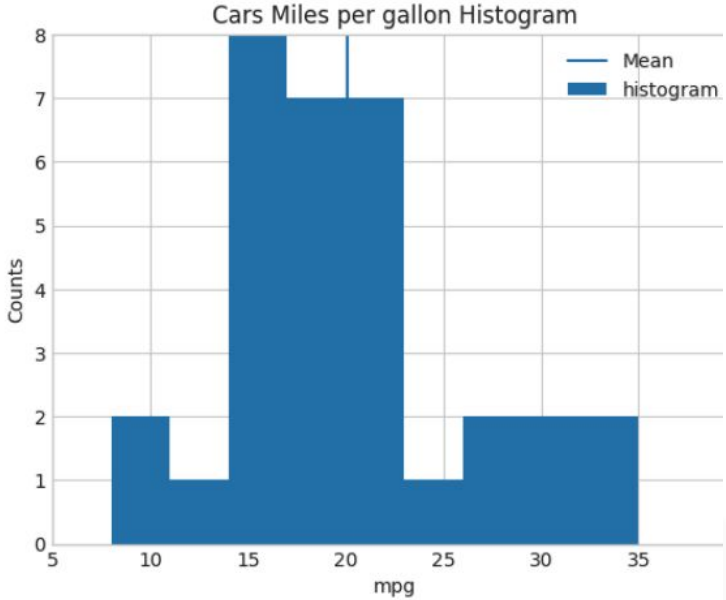
But how?

- **Build** a DataFrame (pandas) from the data
- **Clean** the DataFrame, i.e,
  - Each row describes a single object
  - Each column describes a property of that object
- Explore **summary** of the data (histograms, scatterplots, aggregation functions)
- Explore **subsets** of data (groupBy)

UC San Diego

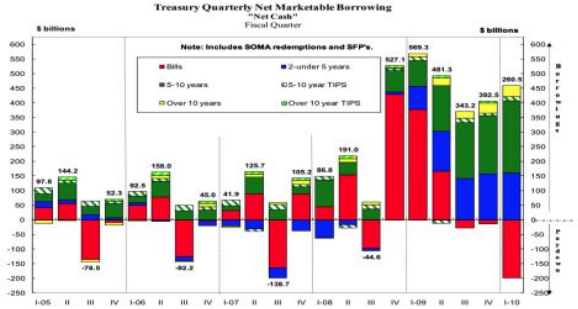
But how?

	name	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	maker
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	Mazda
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	Mazda
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	Datsun
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	Hornet
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	Hornet



UC San Diego

Not “Effective”



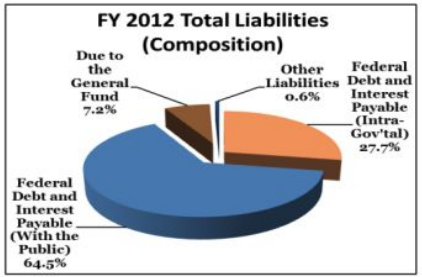
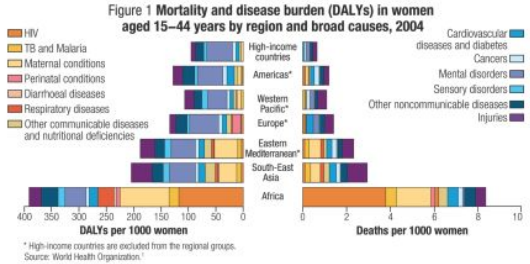
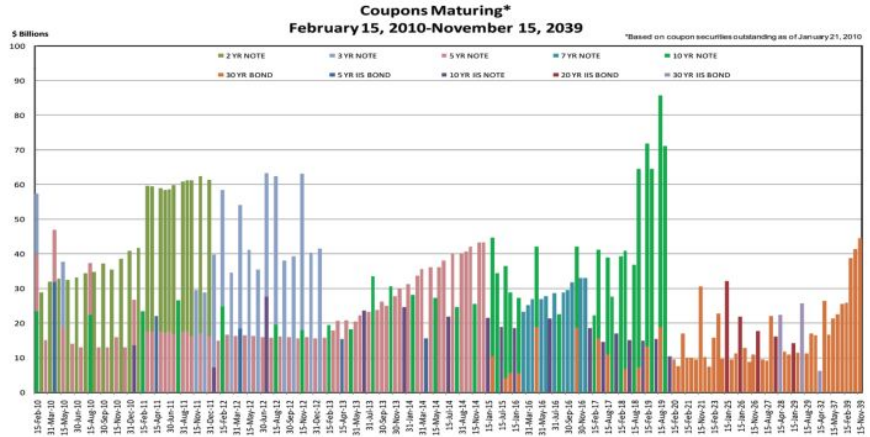


Figure 10





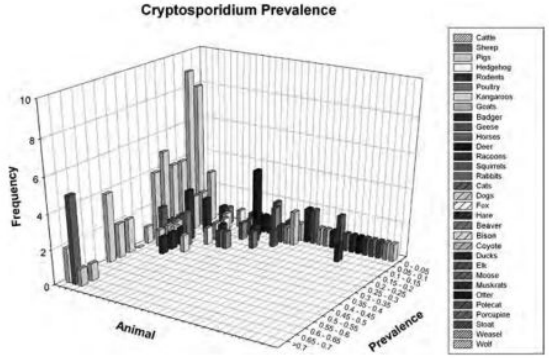


Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Sources: US Treasury and WHO reports

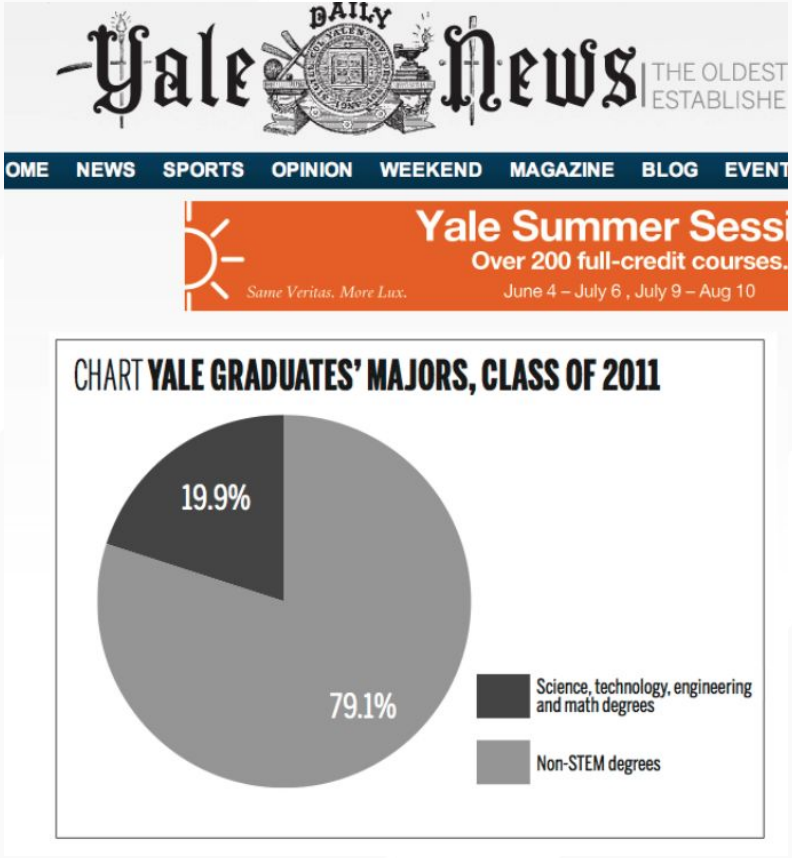
UC San Diego

Effective EDA Viz

- Have graphical integrity
- Keep it simple
- Use the right chart
- Use the right colors

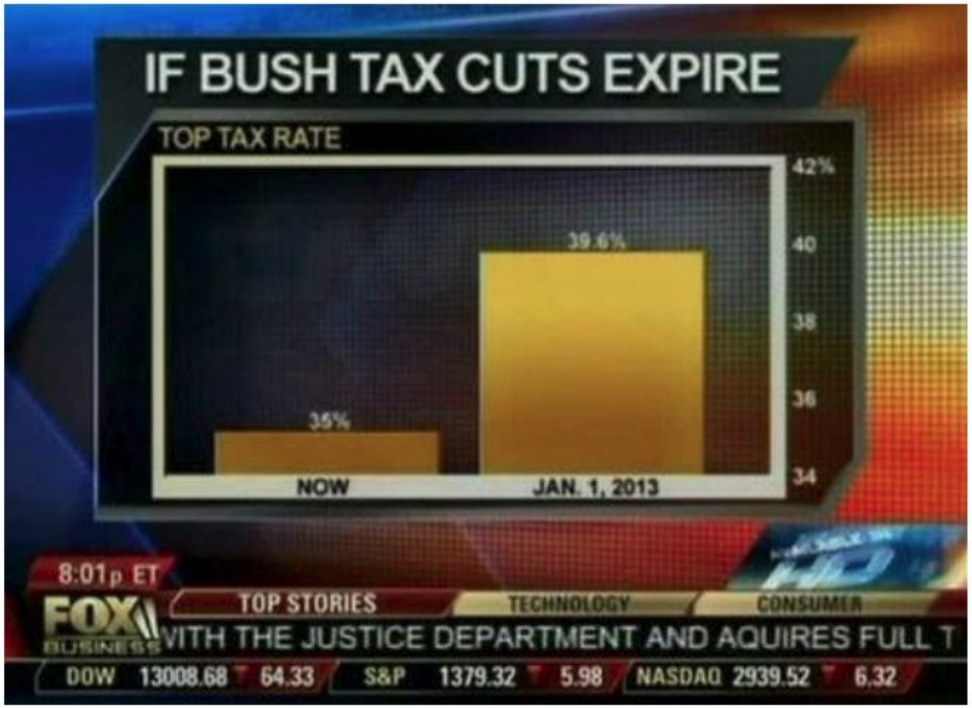
UC San Diego

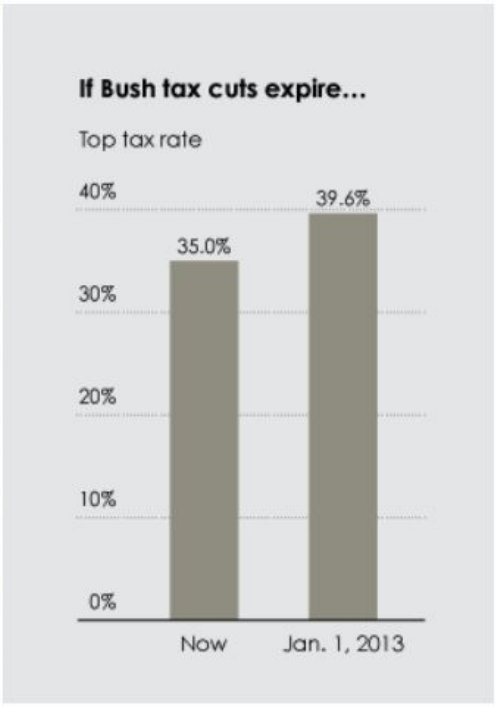
Graphical Integrity



UC San Diego

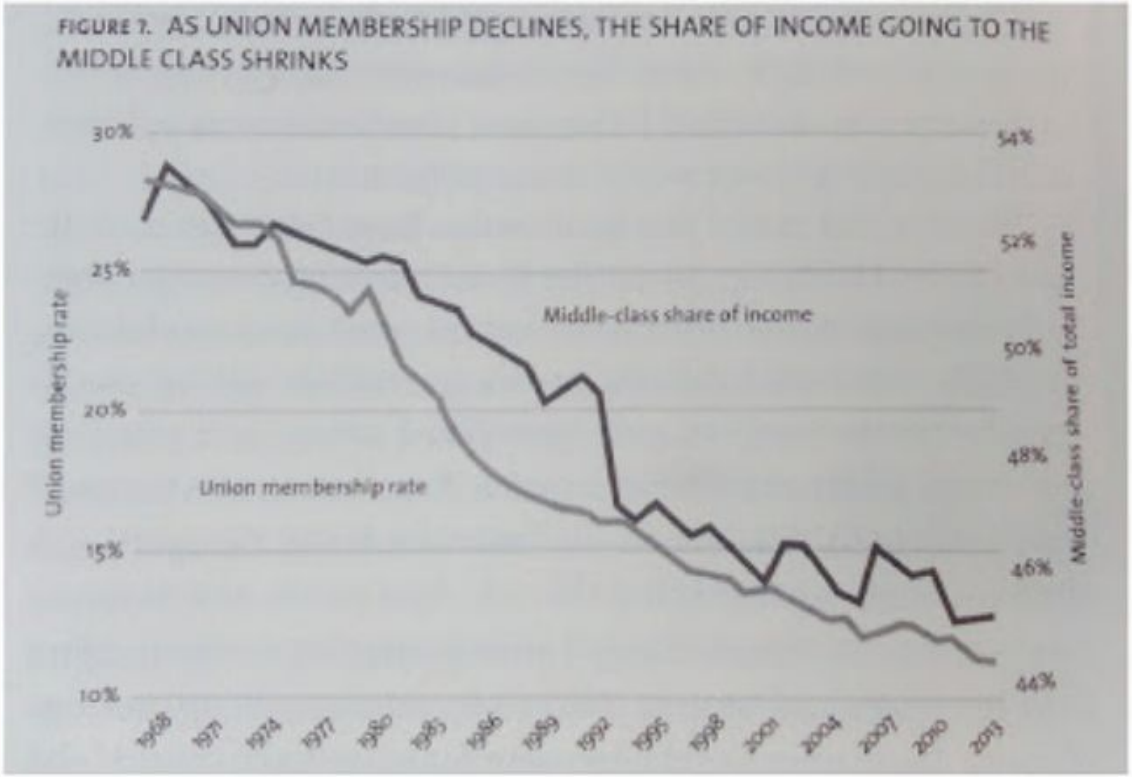
Graphical Integrity



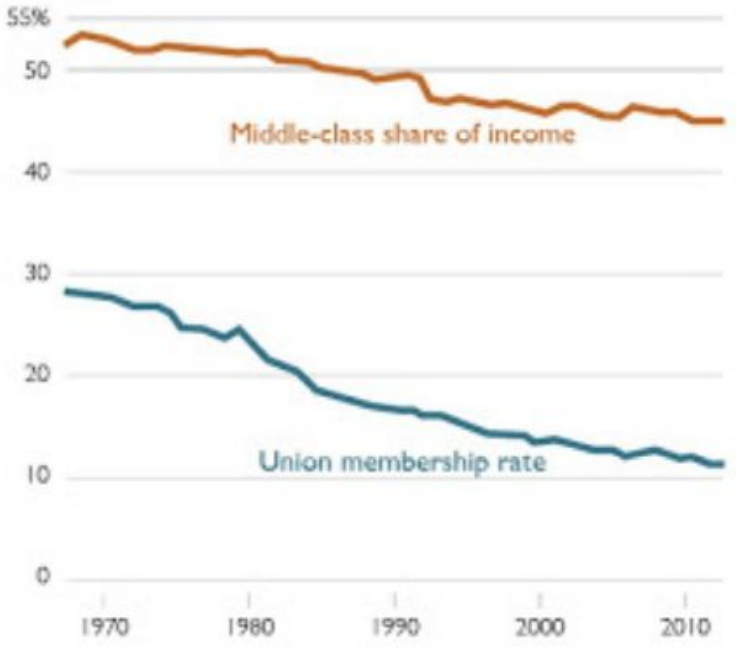


UC San Diego

Graphical Integrity

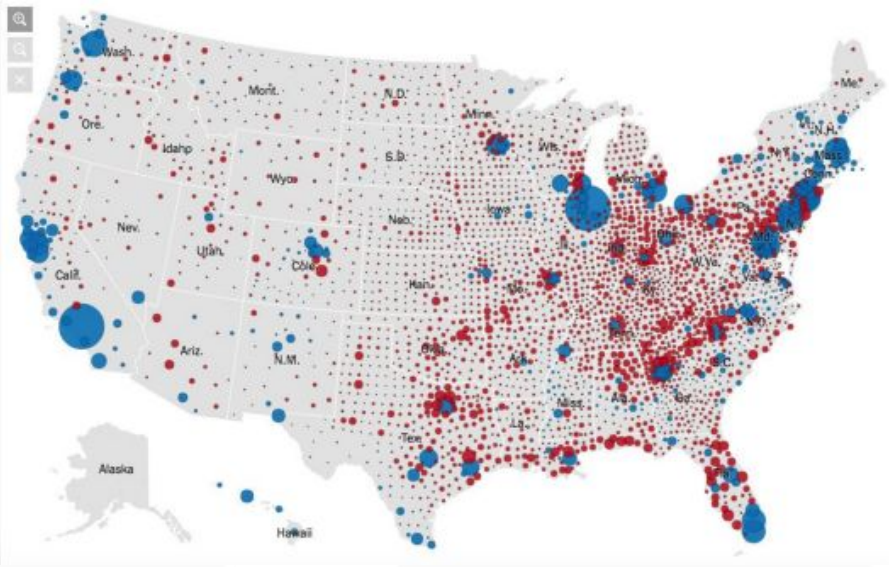


NEW VERSION



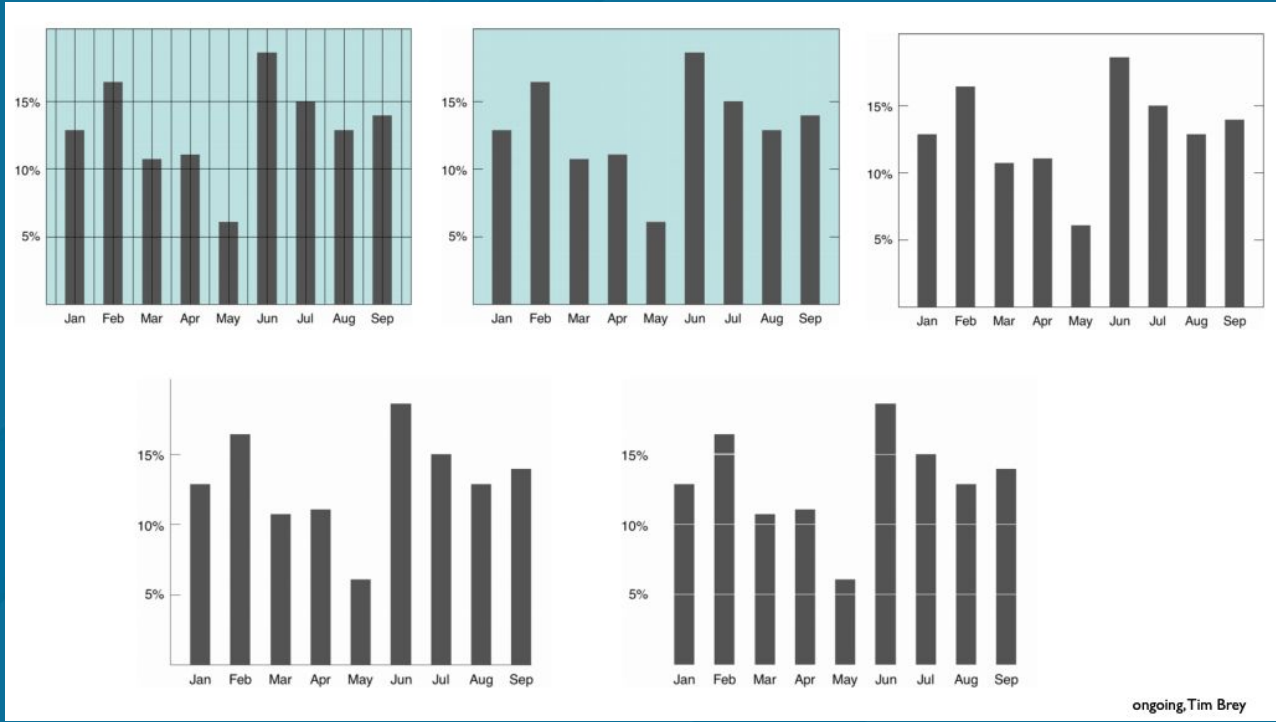
UC San Diego

Graphical Integrity



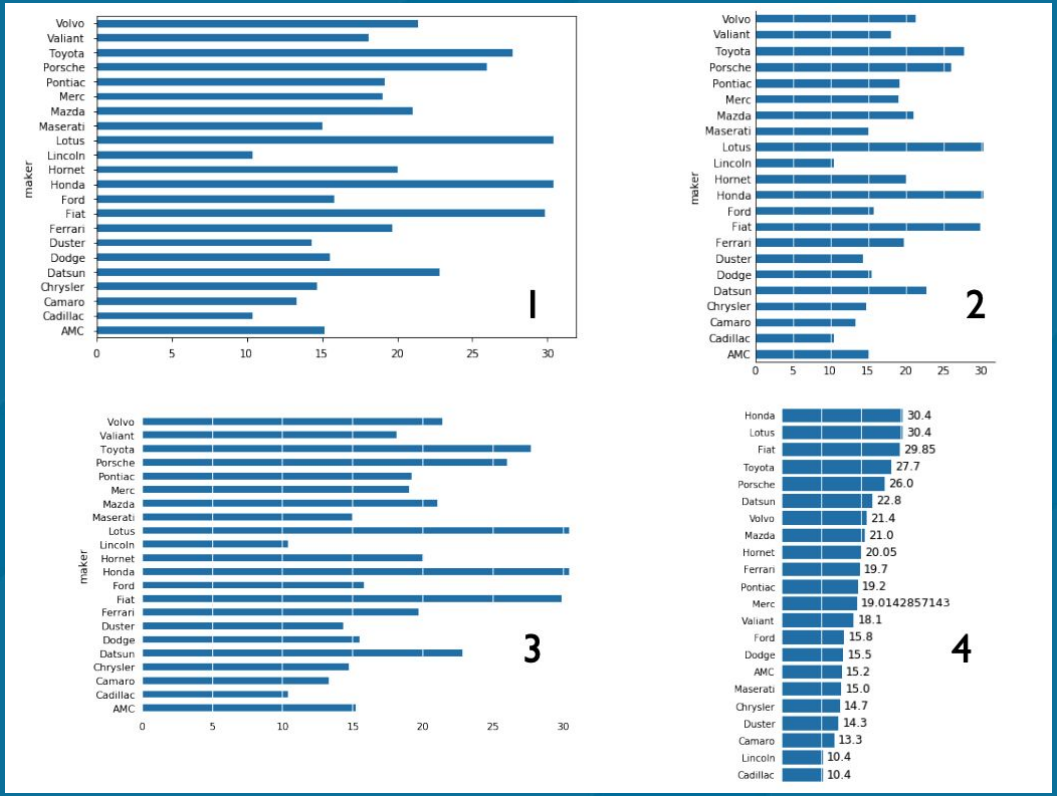
UC San Diego

Keep it simple



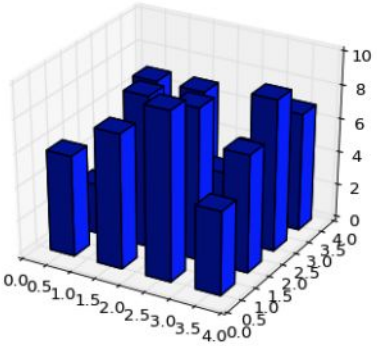
UC San Diego

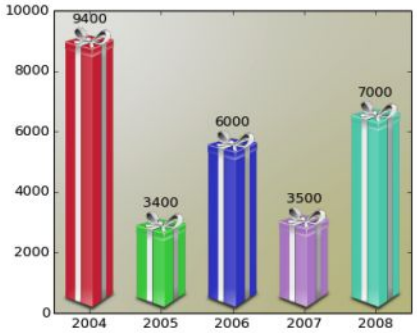
Keep it simple



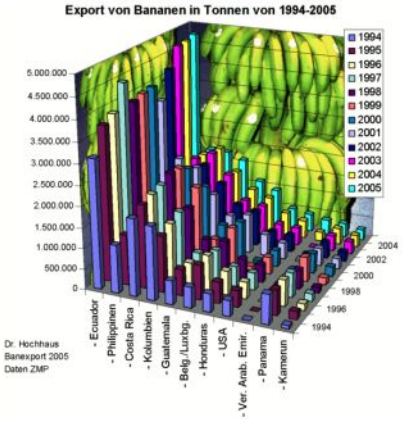
UC San Diego

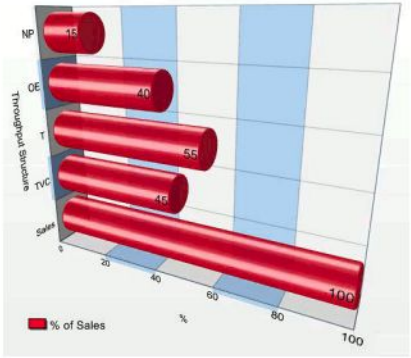
Nope…





matplotlib gallery



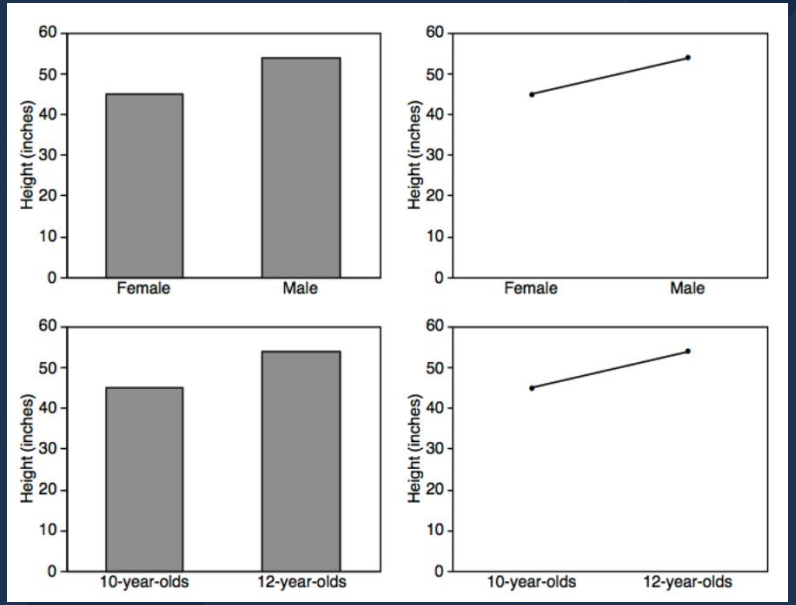


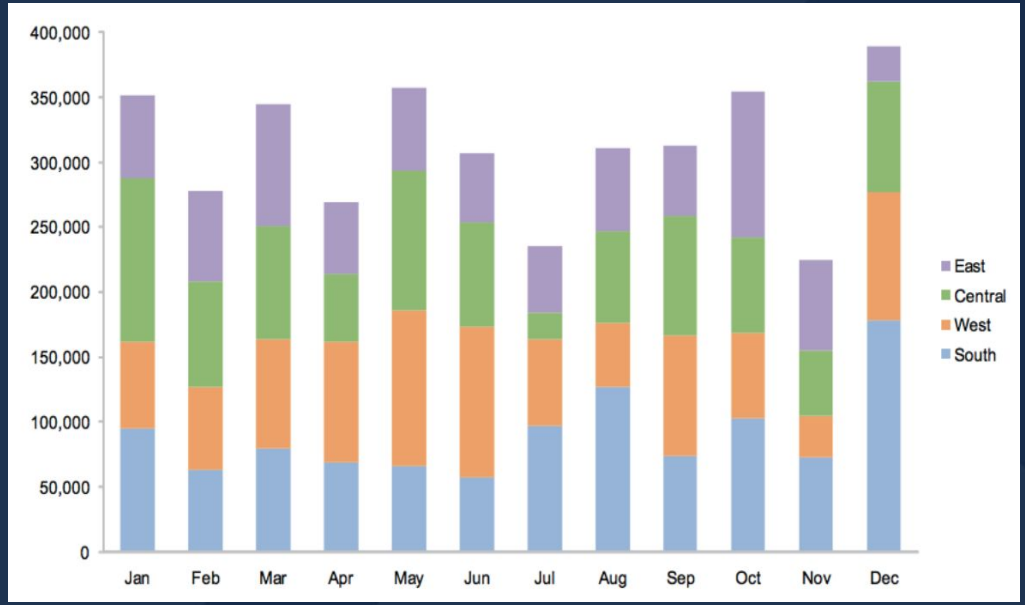
Excel Charts Blog

UC San Diego

Use the right chart

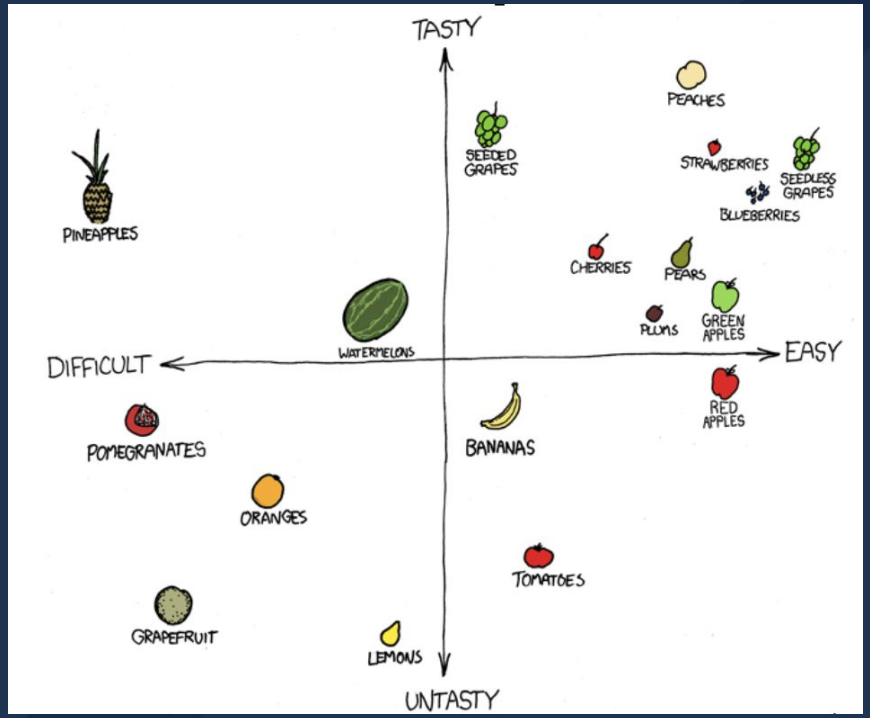
https://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf





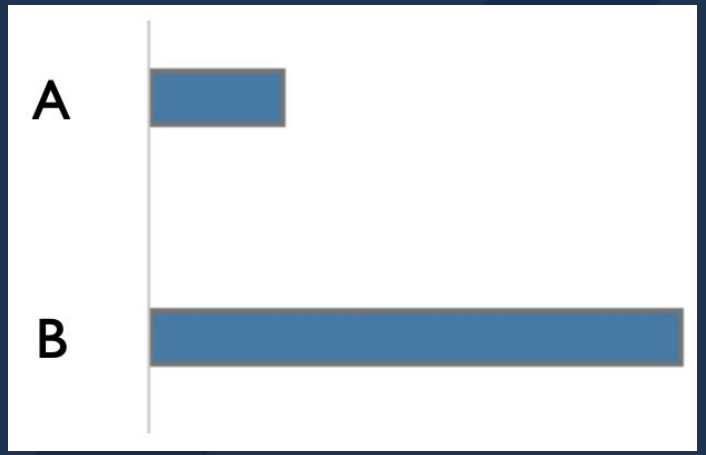
UC San Diego

Use the right chart



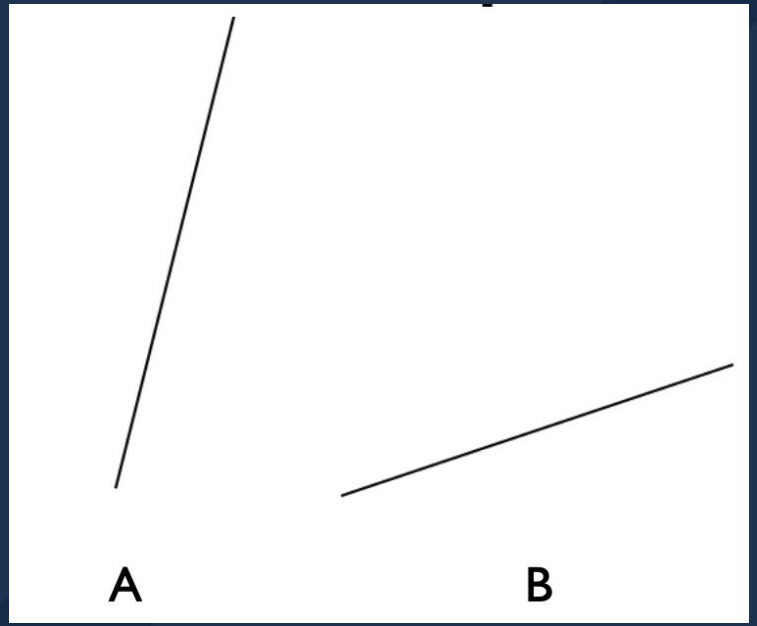
UC San Diego

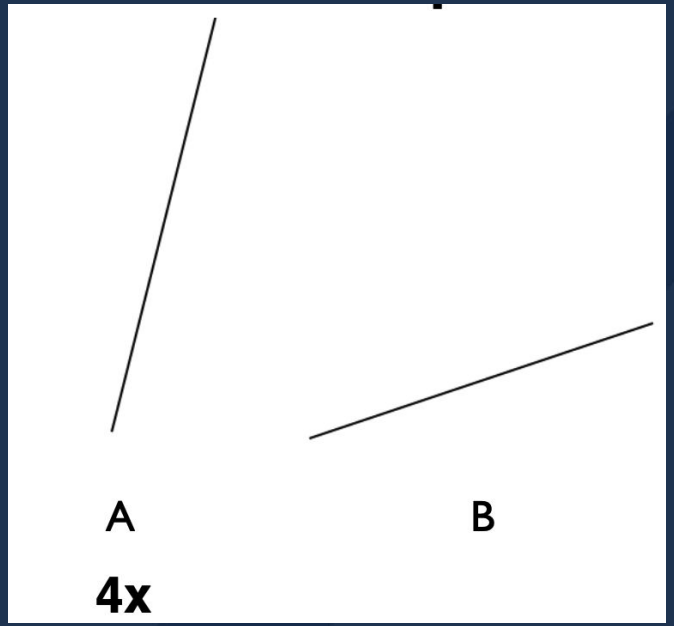
How much longer?



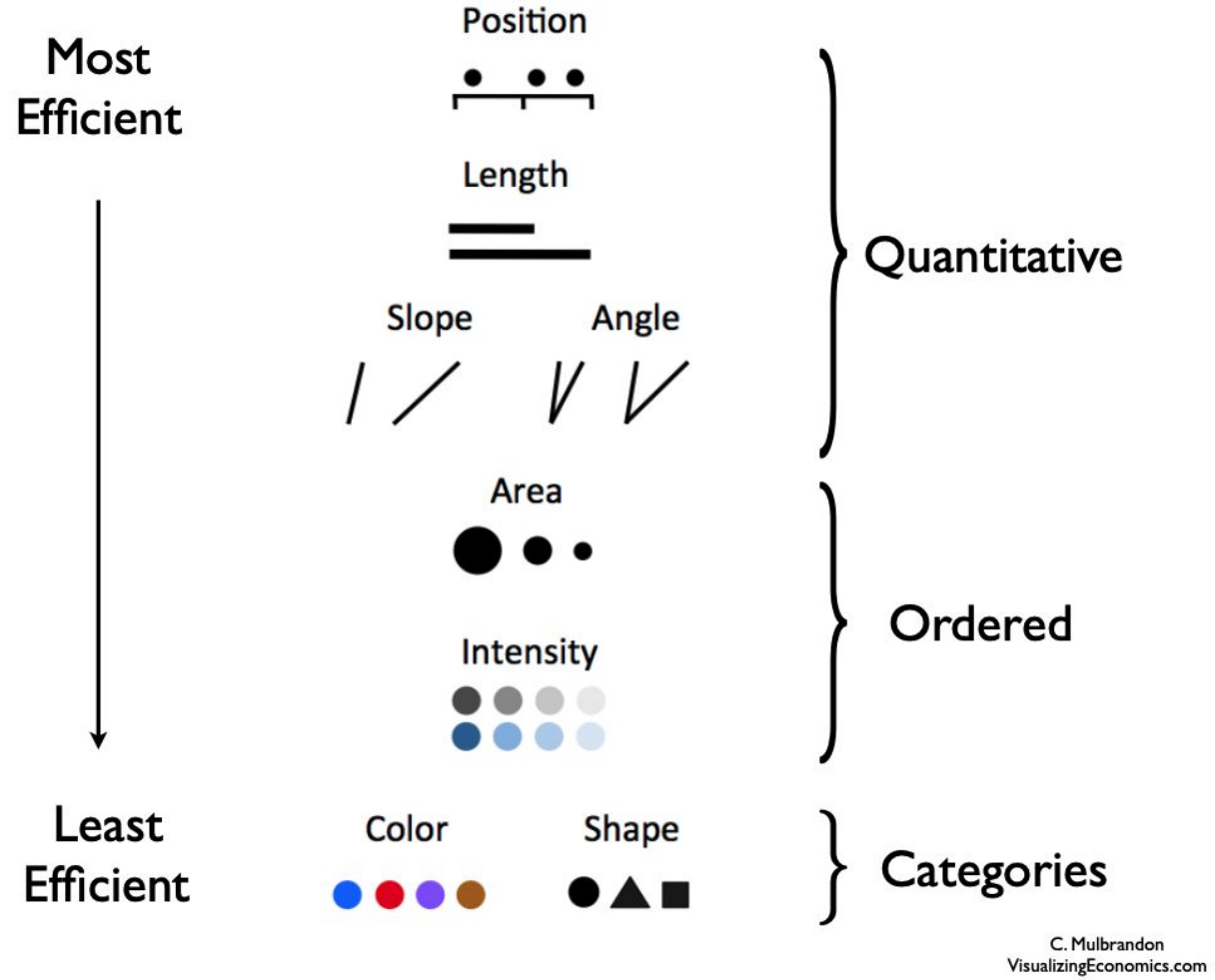
UC San Diego

How much steeper?



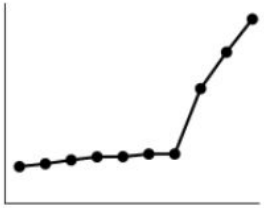


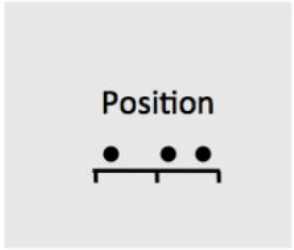
UC San Diego

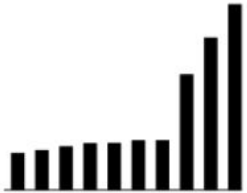

Most Efficient
Position
Length
Slope
Angle
Quantitative
Area
Intensity
Ordered
Least Efficient
Color
Shape
Categories
C. Mulbrandon
VisualizingEconomics.com

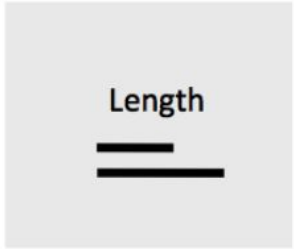
UC San Diego

Most Effective





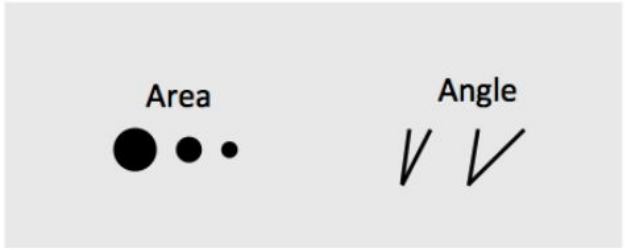


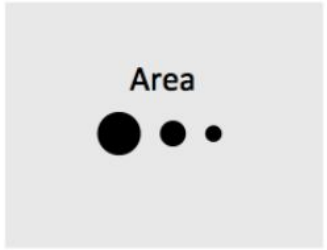


VisualizingEconomics.com

UC San Diego

Less Effective



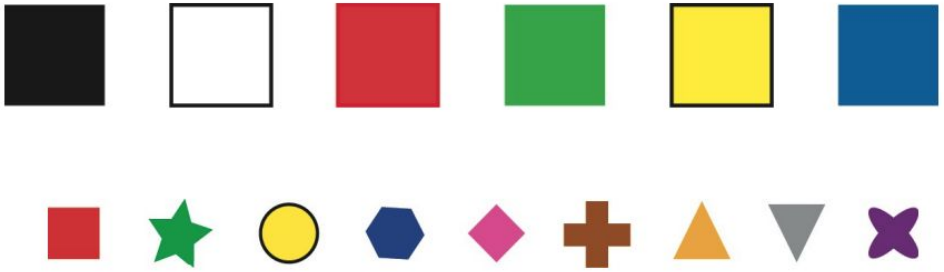


VisualizingEconomics.com

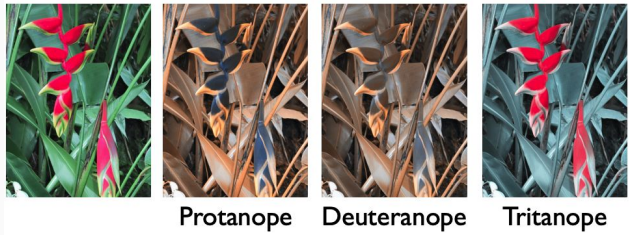
UC San Diego

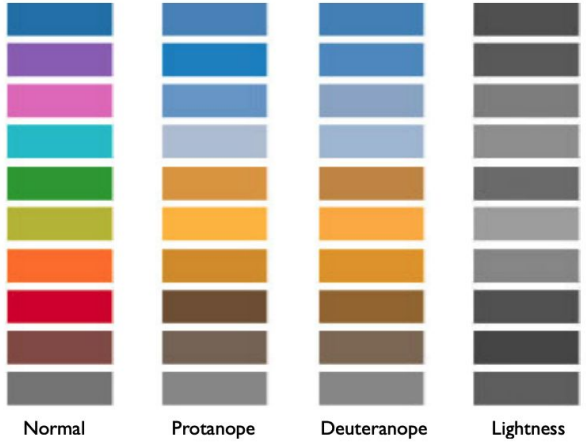
Use the right colors

Do not use more than 5-8 colors at once



Color Blindness





UC San Diego

Communicating

- Who:
  - Who is your audience?
  - What is your relationship to your audience?
- What:
  - What do you need your audience to know?
  - What do you want your audience to do?
  - What will your tone be?
- How:
  - How will you communicate this to your audience? Live or written?

UC San Diego

When NOT to do EDA?

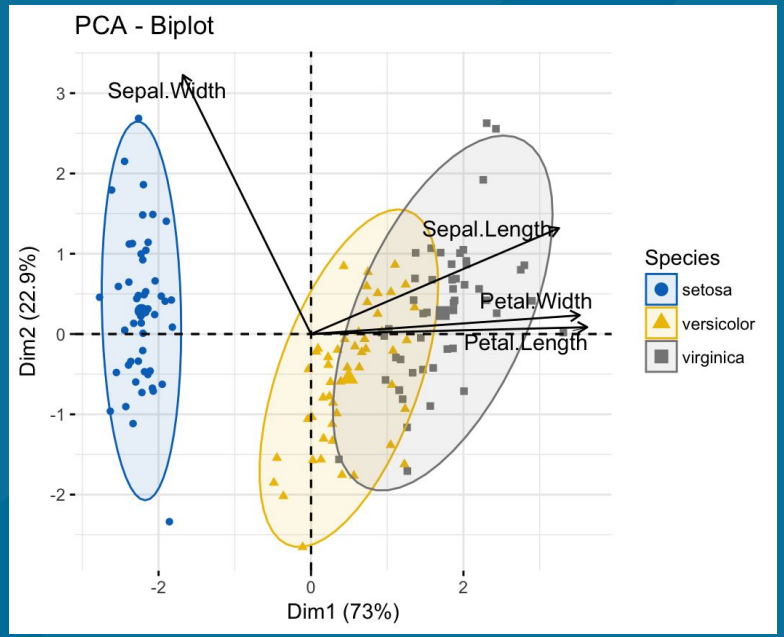
- Identify/remove samples after analyzing data
- After running statistical test and obtaining p-value
- After getting an answer you don't like
- To improve the correlation between variables

EDA is NOT a tool to get your data analysis to give you the results you want.

UC San Diego

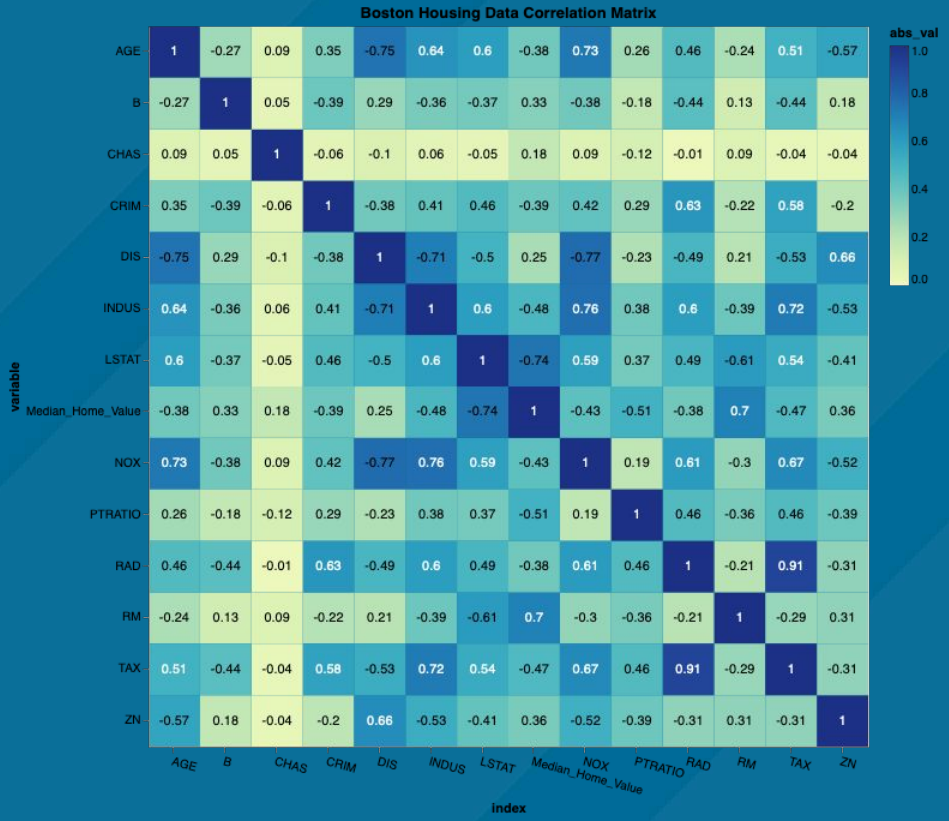
PCA Analysis

- Summarize and visualize the most important part of the data
- Pick the variables/features most closely related to the data



UC San Diego

Correlation Matrix



Source: https://harvard-iacs.github.io/2018-CS109A/lectures/lecture-3/presentation/lecture3.pdf

UC San Diego