Final Project and Proposal

https://iu.instructure.com/courses/2216627/assignments/15873434

Data Visualization

for interactive (e.g., Plotly) **and** static (e.g., Seaborn) visualizations

What is Data Visualization

“The use of computer-generated, interactive, ***visual representations of data*** *to amplify cognition.”*

“The ***transformation of data into visual representations*** *to aid people in the analysis, exploration, and communication of that data”*

Card et al. 1999 and Jeff Heer

Data

What are we gathering?

Types of Data

- Healthcare
- Financial Markets
- Scientific Data
- Social Media
- etc.

How much data are we gathering

- Data on Wikipedia *alone* in 2023 = **439.17 TB**
  - Over 100 million files.
- This is just one (relatively lightweight) server
- Extrapolate this to the whole of the internet . . .
- 1000 “Wikipedias” would take up 420 petabytes
  - 420,... with 16 zeros . . . bytes (e.g., an integer takes up ~4 bytes)

https://commons.m.wikimedia.org/wiki/Special:MediaStatistics

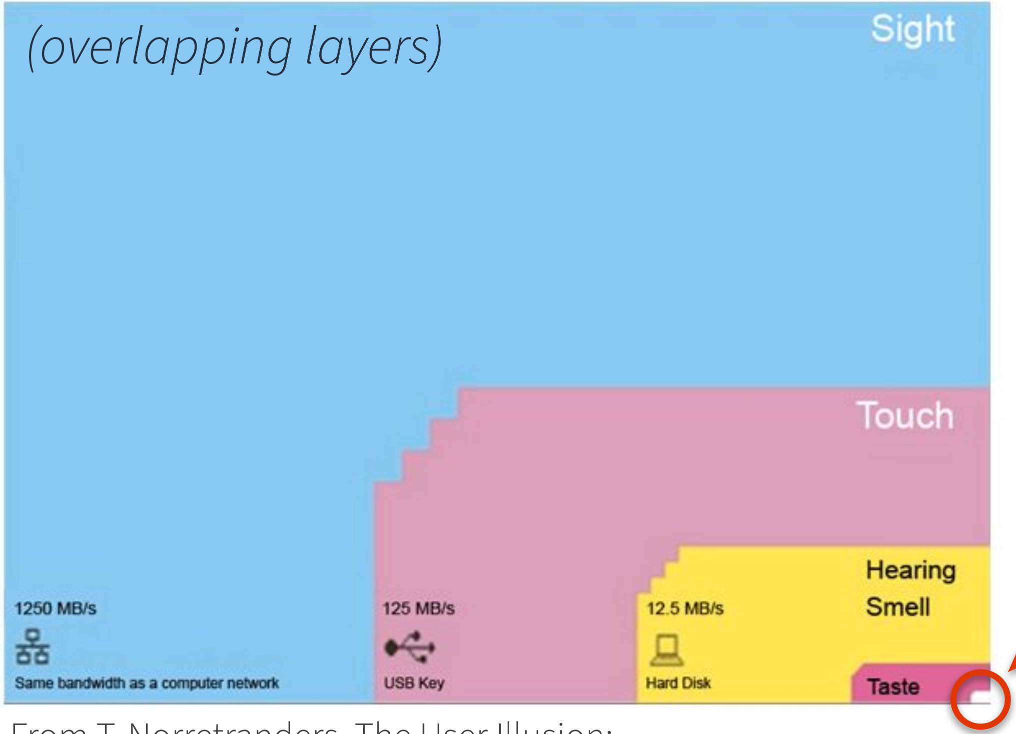
How to interpret it?

Sight,
Hearing,
Touch,
Smell,
Taste,
Telepathy

How to interpret it?

Sight,
Hearing,
Touch,
Smell,
Taste,
Telepathy

Sensory Bandwidth



From T. Norretranders, The User Illusion: *Cutting Consciousness Down to Size*, 1999

This small white box is what we’re actually conscious of . . .

How many “X” letters are here?

345OIJDFG98C90U5ET09VBKK23490XIVBCIBJ0345T09U

2G84GDF09U34590IDFK90345I-09345K90FU90DF90JDF

34T09X90DFJG90J34T09J34509J3459DFG08JKLSTJP435

DFDFG45OJERPOTJ45OPIJFDGLKM34T5XJSCTYY7K456

POJ345OIJLGJKOPE390UVFHUDGH9345H9R4N97HWTIO

MADSIOPEJDFGPJ4309UT509345PODFGX093490823JFD

PWDEIJ3408UDFMV984385Y0834N92384YU8DFB0H3T4N

345J09JDFG09J345X98U5Y09JGFB089H34509UJ45TM0IG

P5JDGIOEGWJPIO345U345OPIJDTOPI3458345JPODFG09

45POJ34X09345J08EFJ825HJDFSJIPADOPQWIXERWNVF

Based on a slide by John Stasko

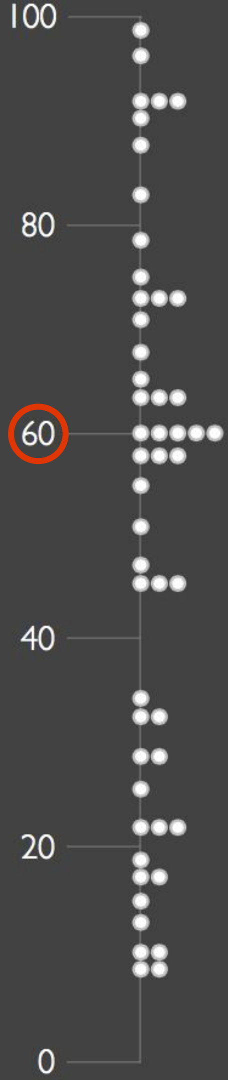
345OIJDFG98C90U5ET09VBKK23490**X**IVBCIBJ0345T09U
2G84GDF09U34590IDFK90345I-09345K90FU90DF90JDF
34T09**X**90DFJG90J34T09J34509J3459DFG08JKLSTJP435
DFDFG45OJERPOTJ45OPIJFDGLKM34T5**X**JSCTYY7K456
POJ345OIJLGJKOPE390UVFHUDGH9345H9R4N97HWTIO
MADSIOPEJDFGPJ4309UT509345PODFG**X**093490823JFD
PWDEIJ3408UDFMV984385Y0834N92384YU8DFB0H3T4N
345J09JDFG09J345**X**98U5Y09JGFB089H34509UJ45TM0IG
P5JDGIOEGWJPIO345U345OPIJDTOPI3458345JPODFG09
45POJ34**X**09345J08EFJ825HJDFSJIPADOPQWI**X**ERWNVF

Based on a slide by John Stasko

15	19	60
33	11	75
57	34	79
18	51	92
73	22	13
71	60	22
17	10	68
73	18	55
65	46	29
60	73	22
46	92	97
10	58	46
57	17	83
26	99	33
88	92	60
91	29	57
96	12	47

Given these 50 numbers
what number appears most often?

Slide by Miriah Meyer, University of Utah
http://www.cs.utah.edu/~miriah/teaching/cs6630/



Given these 50 numbers
what number appears most often?

Slide by Miriah Meyer, University of Utah
http://www.cs.utah.edu/~miriah/teaching/cs6630/

Visuals in our Culture

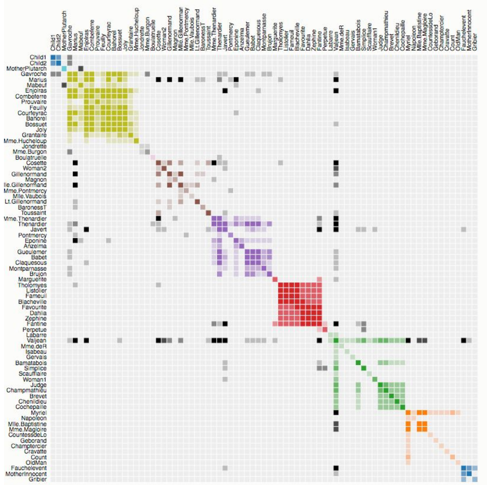
- “I see what you’re saying”
- “Seeing is believing”
- “I now see the big picture”
- “A picture is worth a thousand words”

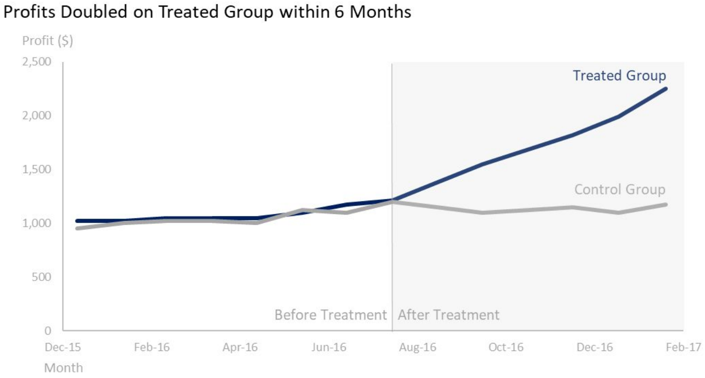
Computers vs. People

- Computers
    - Process large quantities of information quickly
    - Computing a specific task
    - Re-use of code for different datasets
- Humans
    - Devising questions when we don’t know what to look for
    - We can use visuals to help with this

Building Visualizations

Spectrum of Visualization Use





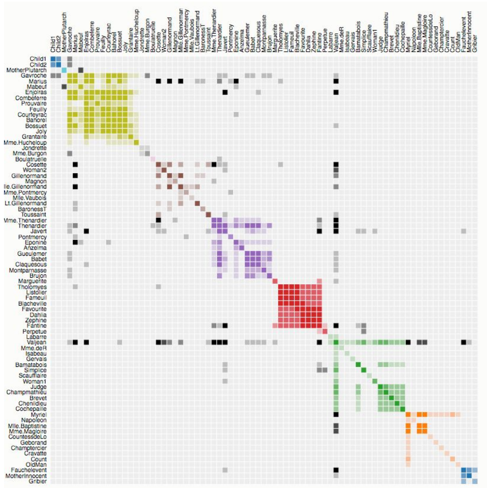
Analysis

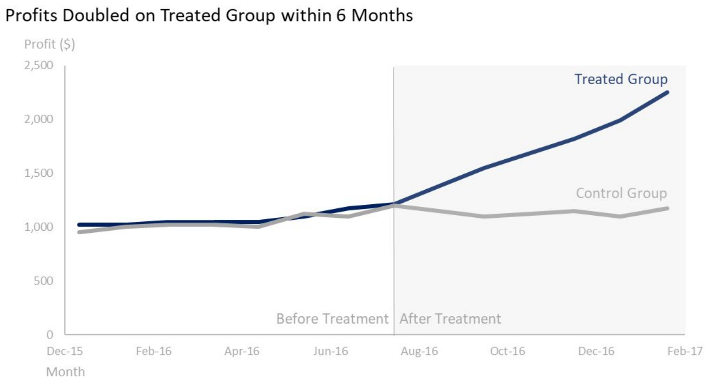
Better understand data

Communication

tell a story with data

Spectrum of Visualization Use





Analysis

ad hoc; for your eyes only

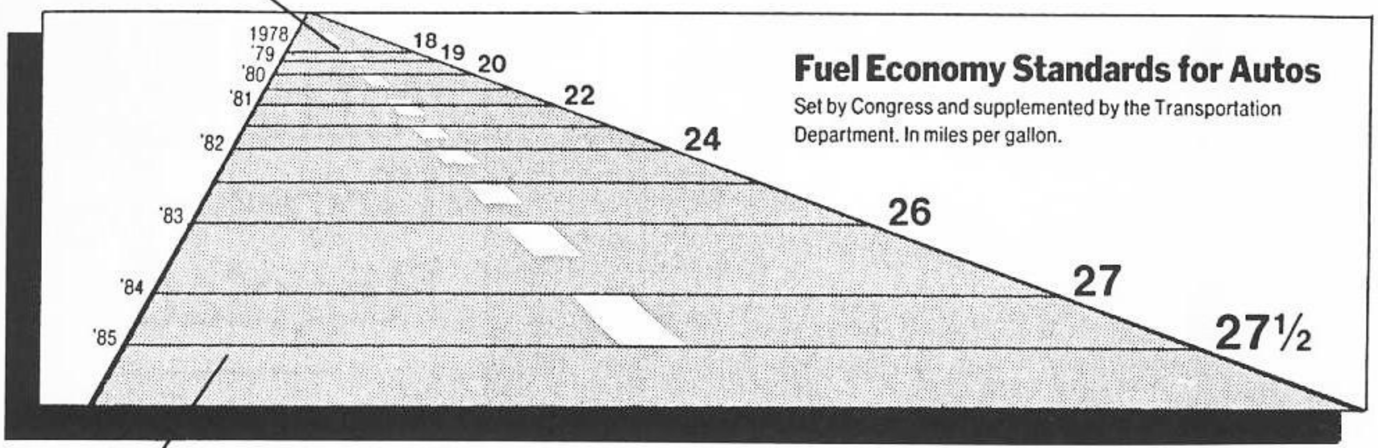
Communication

more presentable!

Visual Integrity: **not to lie with data**

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented.

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

$$\text{lie factor} = \frac{\text{graphic effect}}{\text{numeric effect}}$$

$$= \frac{(5.3 - 0.6)/0.6}{(27.5 - 18)/18} = \frac{7.8}{0.5} = 15.6$$

Plot Types

Big Numbers

- Great to display simple information
- Tips:
  - Include a baseline for comparison (e.g., an average)

Page Views:

5,567

Daily Average: 3,625

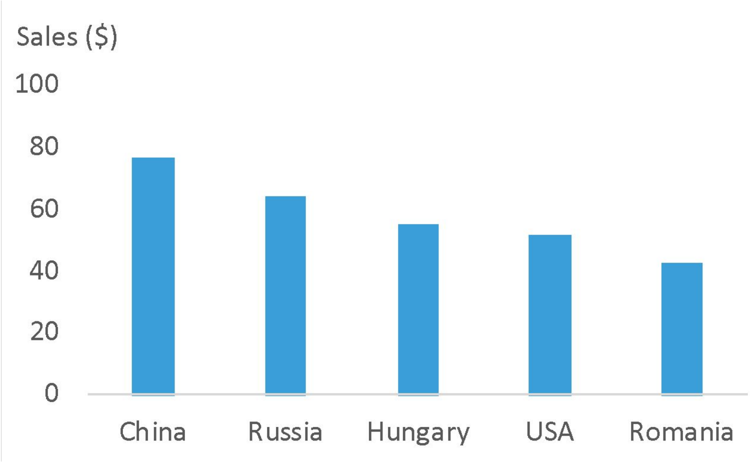
Tables

- Conveys comparisons across categories
- Tips:
  - Too big is less effective
  - Use background color sparingly
  - Bold what's important (e.g., the data)
  - Keep to less than 3 x 3

	Likes Chocolate	Dislikes Chocolate
Children	**65**	**5**
Adults	**40**	**30**

Bar Charts

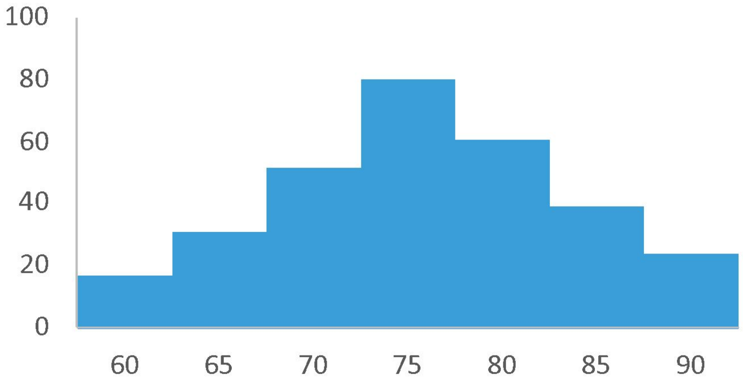
- Compare different groups or categories
- Tips:
  - Sort based on something meaningful (e.g. length, alphabetical)
  - **Start scale at 0 (for all plots, really)**
  - Horizontal, vertical and stacked bars are commonly used



Histograms

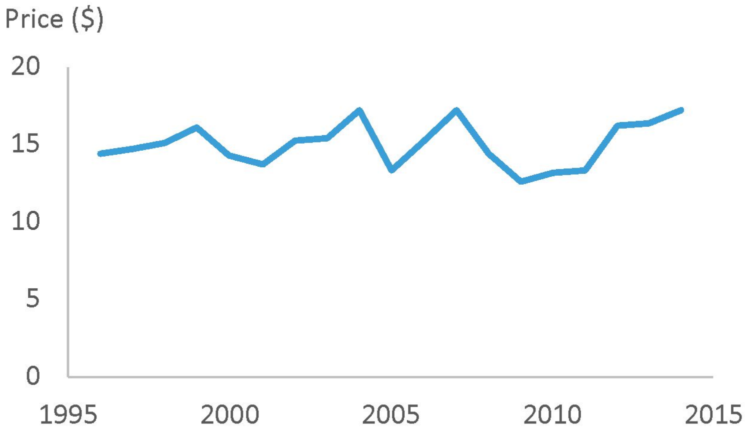
- Represents the distribution of data
- Tips:
  - Range of values must be binned
  - Can be normalized

    (sum of bar heights is 1)



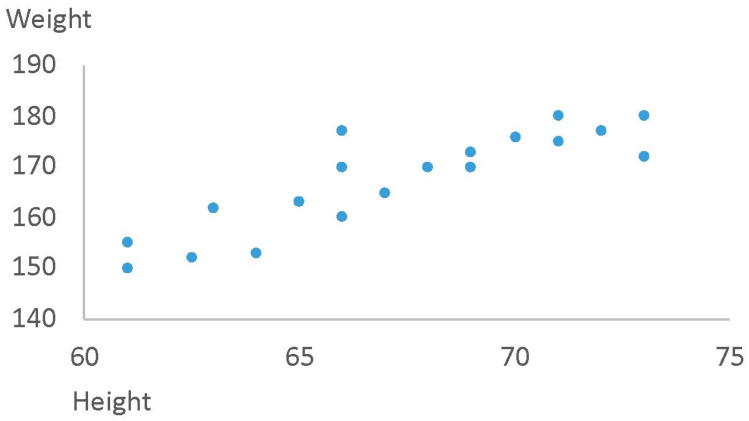
Line Charts

- Track changes over time
- Height and slope lets us see trends
- Tips:
  - x-axis should be *continuous* data
  - Time is represented from left to right
  - **Do not use if x-axis is not ordered**



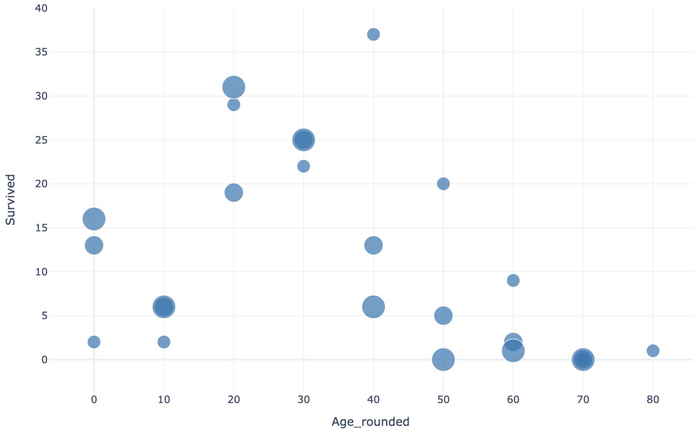
Scatter Plots

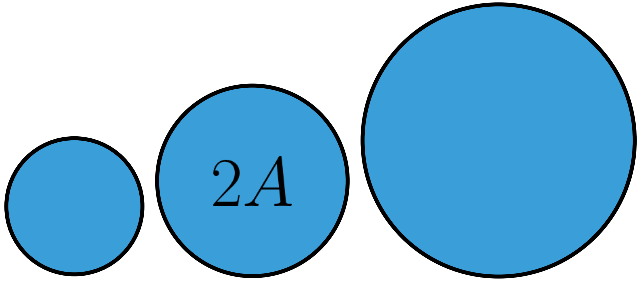
- Displays relationship between two measures
- Tips:
  - Best for continuous data
  - Avoid using with qualitative data



Bubble Charts

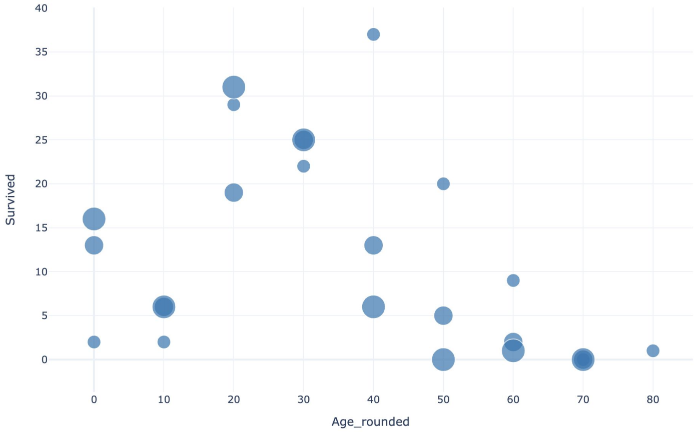
- Used for scatter plots with 3 variables
- Tips:
  - Area causes confusion
  - Use diameter

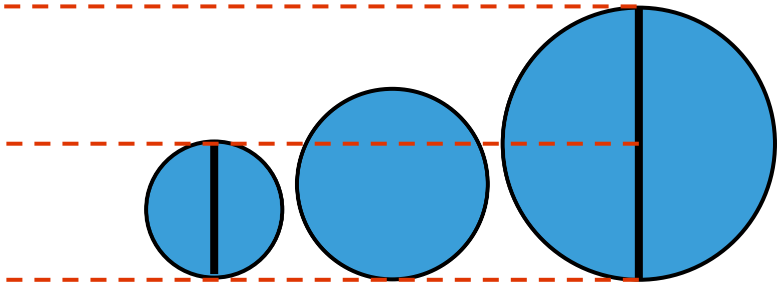




Bubble Charts

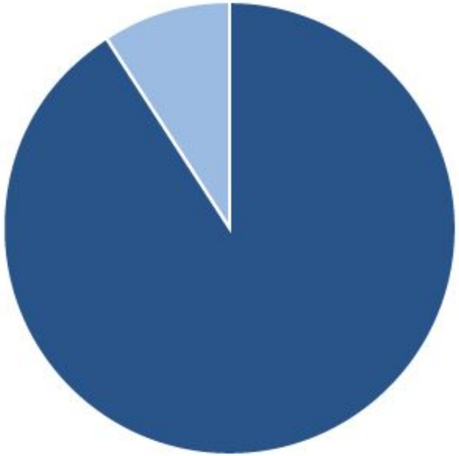
- Used for scatter plots with 3 variables
- Tips:
  - Area causes confusion
  - Use diameter





Pie Charts (think “slice”)

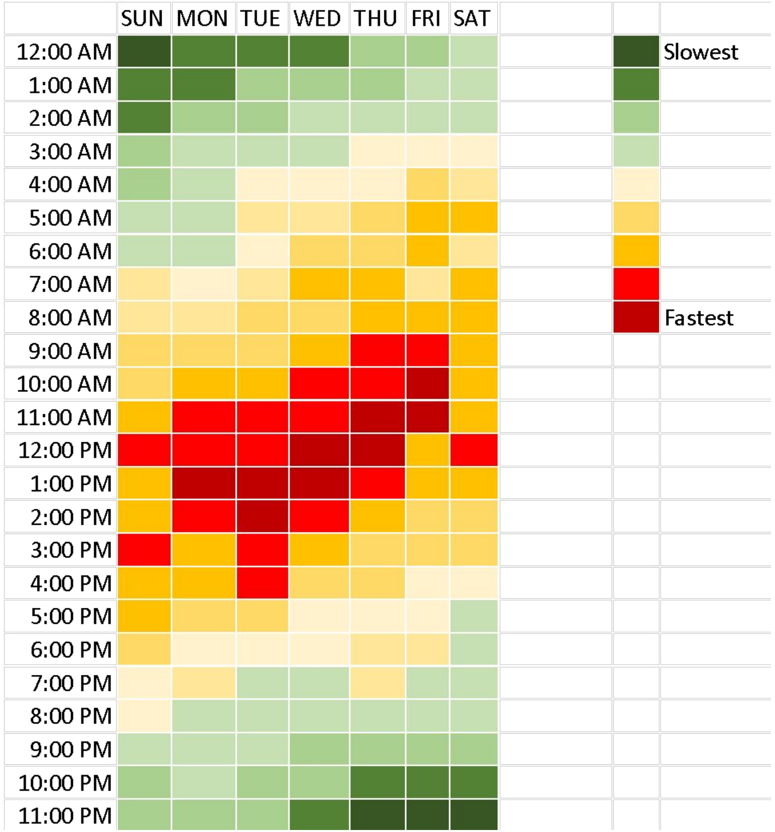
- Compare parts of a whole
- Tips:
  - All parts must sum to 100%
  - Best for binary data
  - **Only use if one is very small**



Heat Maps

- Use color to visualize a matrix
- Tip:
  - **Best when x *and* y variables are ordinal**

    (E.g., weekday, low to high, distance)
  - Use sensible color maps!
  - Color *by-column* or *by-row*
- Very useful for binary data (e.g., missing)



Design

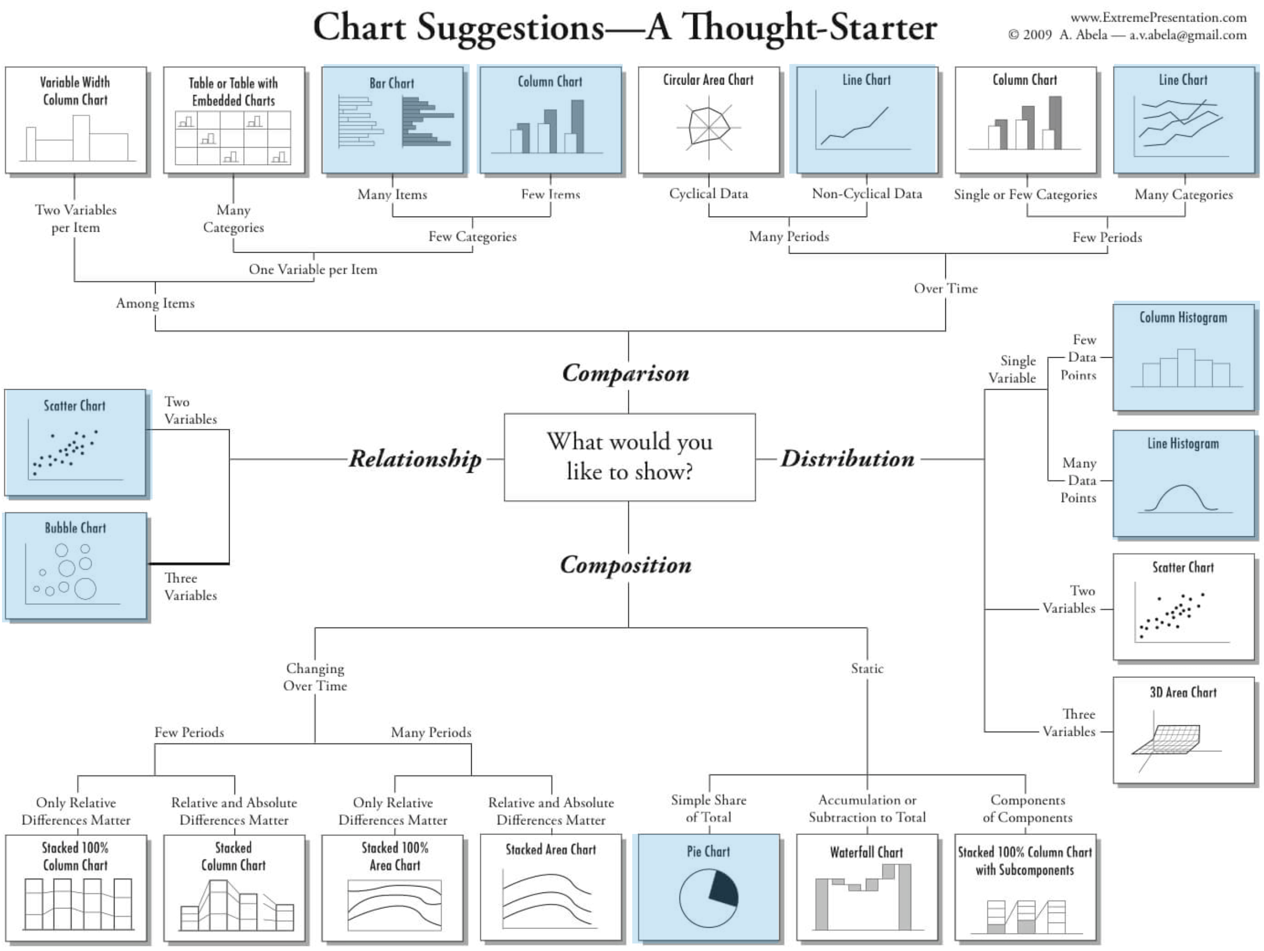
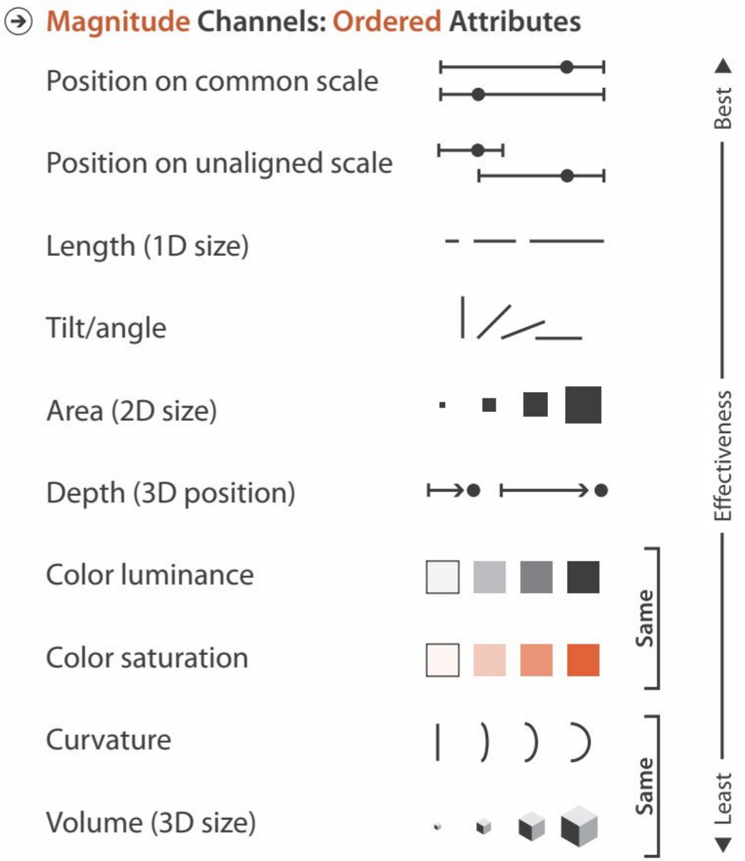


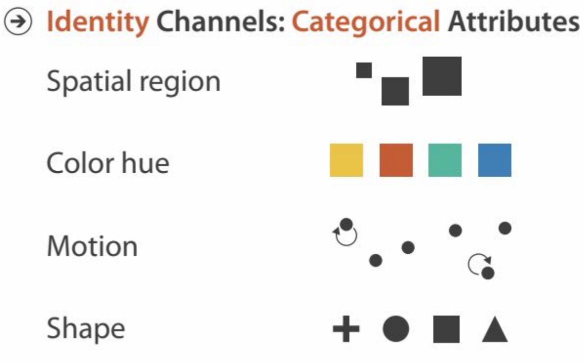
Chart Suggestions—A Thought-Starter

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

Visual Principles





- **expressiveness principle**
  - match channel and data characteristics
- **effectiveness principle**
  - encode most important attributes with highest ranked channels

Cr: Tamara Munzner,
https://www.cs.ubc.ca/~tmm/talks.html#vad21biomedvis

Text Orientation

Vertical text orientation requires
some mental effort!

Text Orientation

Vertical text orientation requires mental effort!

So does this!

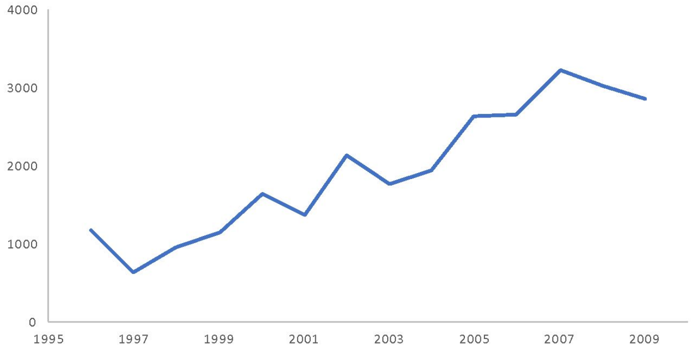
Text Orientation

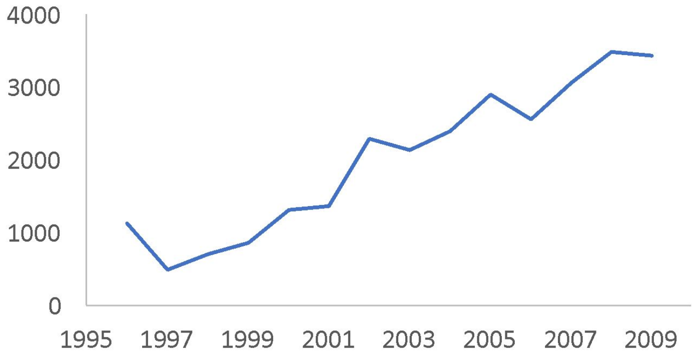
Vertical text orientation requires mental effort!

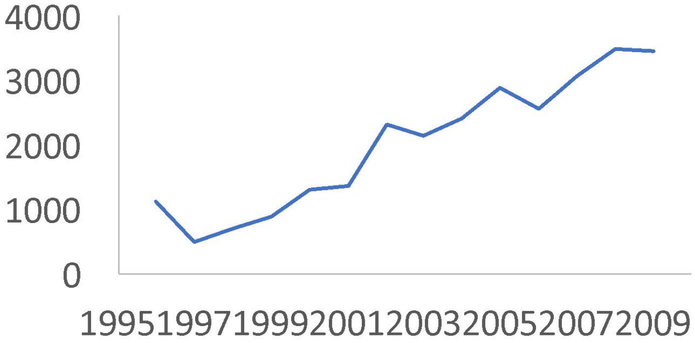
So does this!

Horizontal text is easier to read!

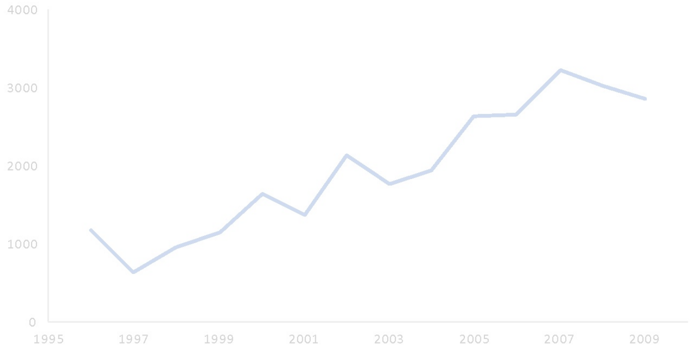
Font Size

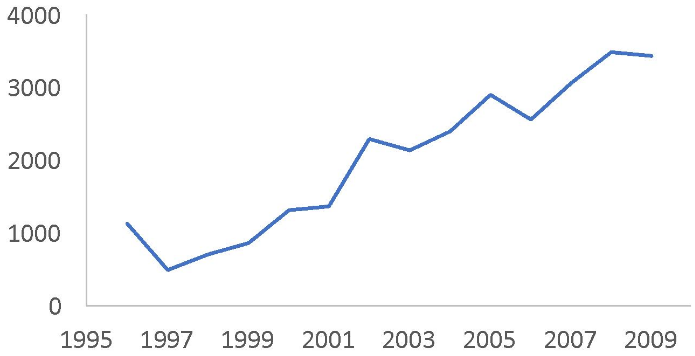

4000
3000
2000
1000
0
1995
1997
1999
2001
2003
2005
2007
2009

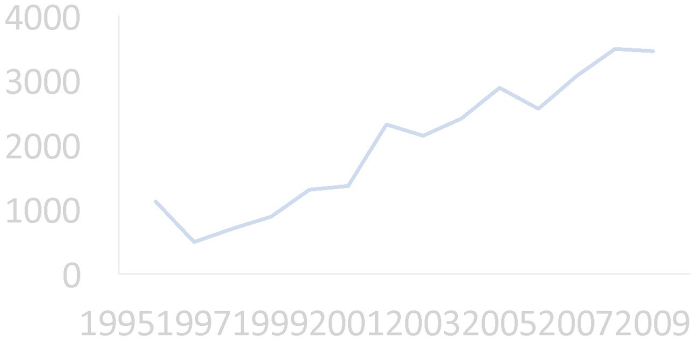

4000
3000
2000
1000
0
1995
1997
1999
2001
2003
2005
2007
2009


4000
3000
2000
1000
0
19951997199920012003200520072009

Font Size


4000
3000
2000
1000
0
1995
1997
1999
2001
2003
2005
2007
2009

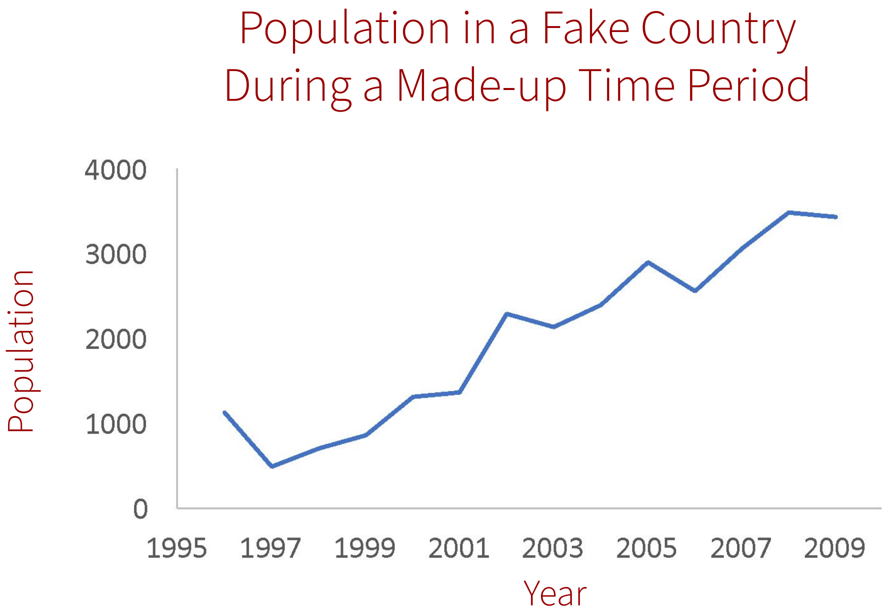

4000
3000
2000
1000
0
1995
1997
1999
2001
2003
2005
2007
2009


4000
3000
2000
1000
0
19951997199920012003200520072009

Labels

Always include all descriptive labels:

- Title of Plot
- Axis labels
- Legend (color, line, etc.)
- Highlights and Callouts

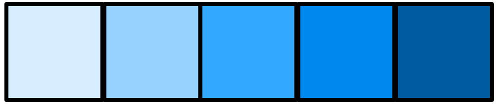


Leave nothing ambiguous or unclear!

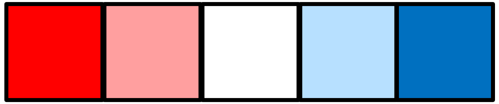
Color

Color

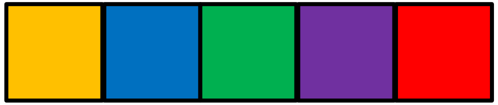
Sequential



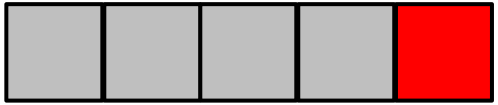
Diverging



Categorical

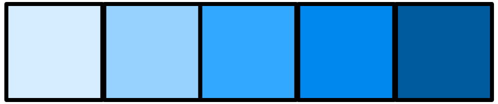


Highlight



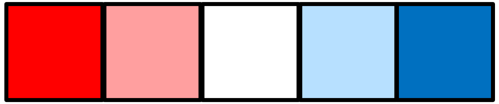
Color

Sequential



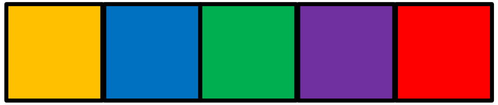
Pick **one** hue; captures “amount”

Diverging



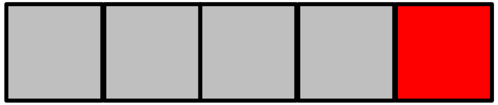
Middle point *must* hold **meaning** (e.g., 0)

Categorical



No color should stand out

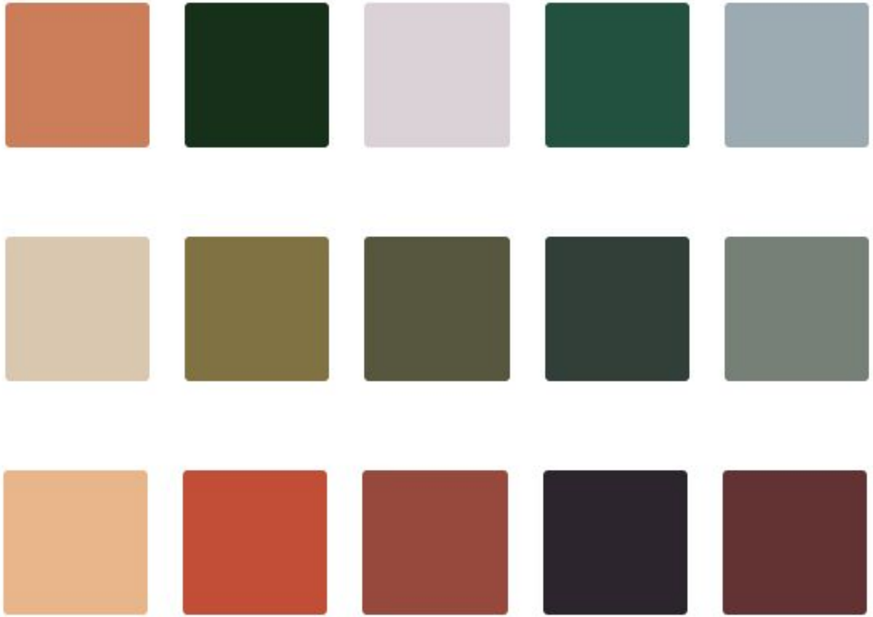
Highlight



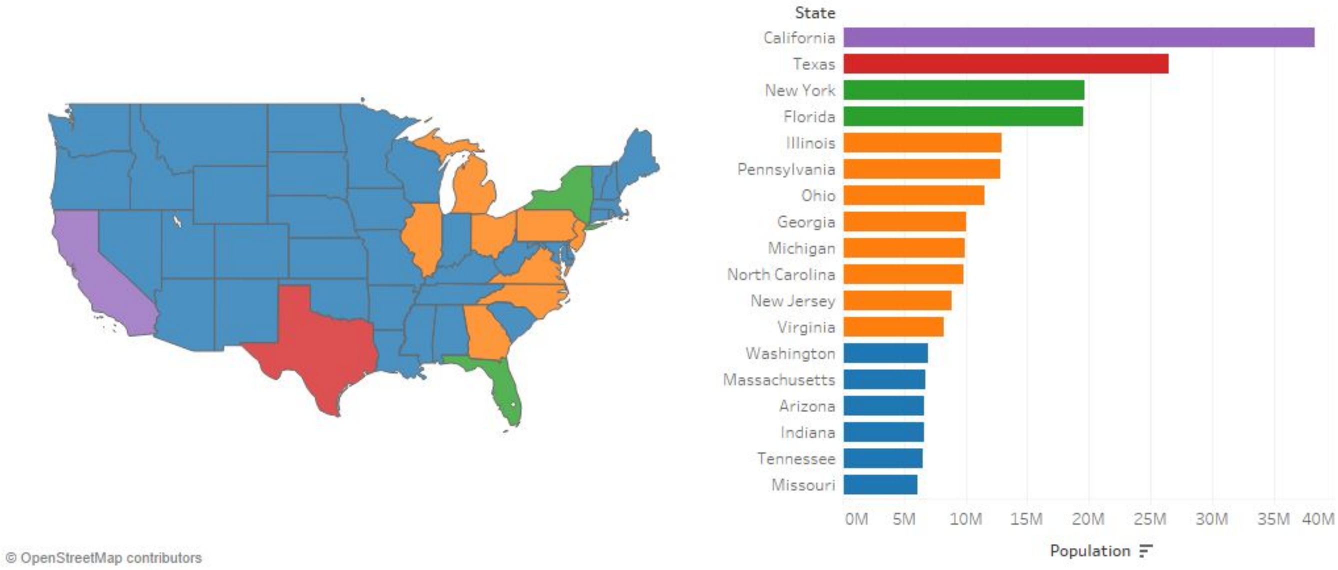
Communicate w/ color (e.g., red = bad)

Color Palette

- **Limit categorical colors to 5**
  - 7 at the absolute most.



Color consistency across charts

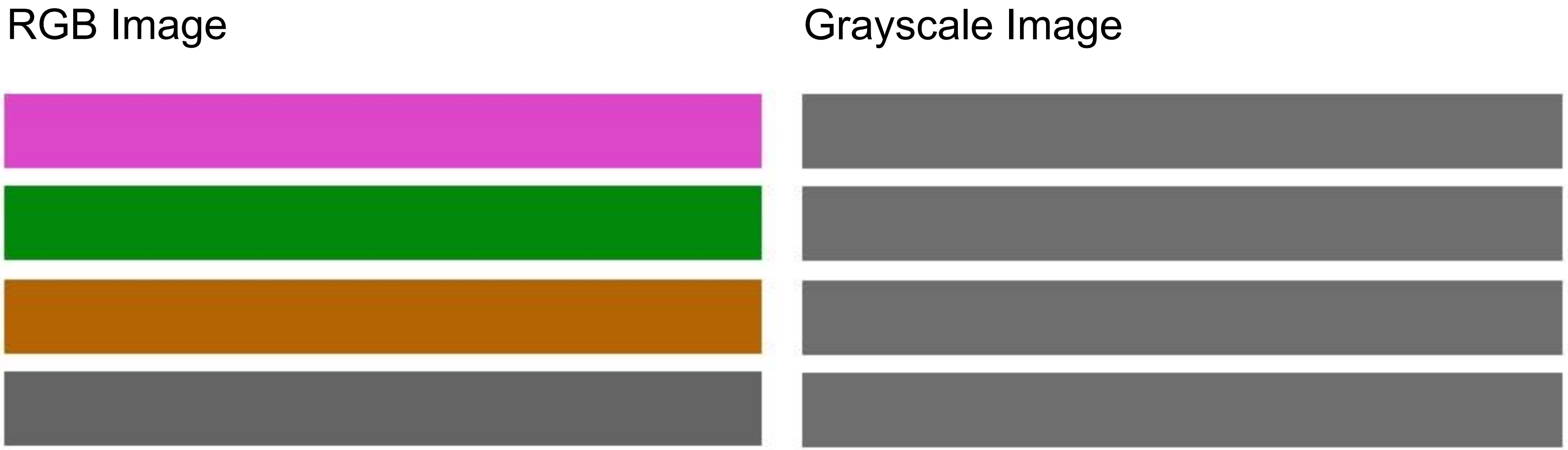


© OpenStreetMap contributors

No color should stand out more than any other.

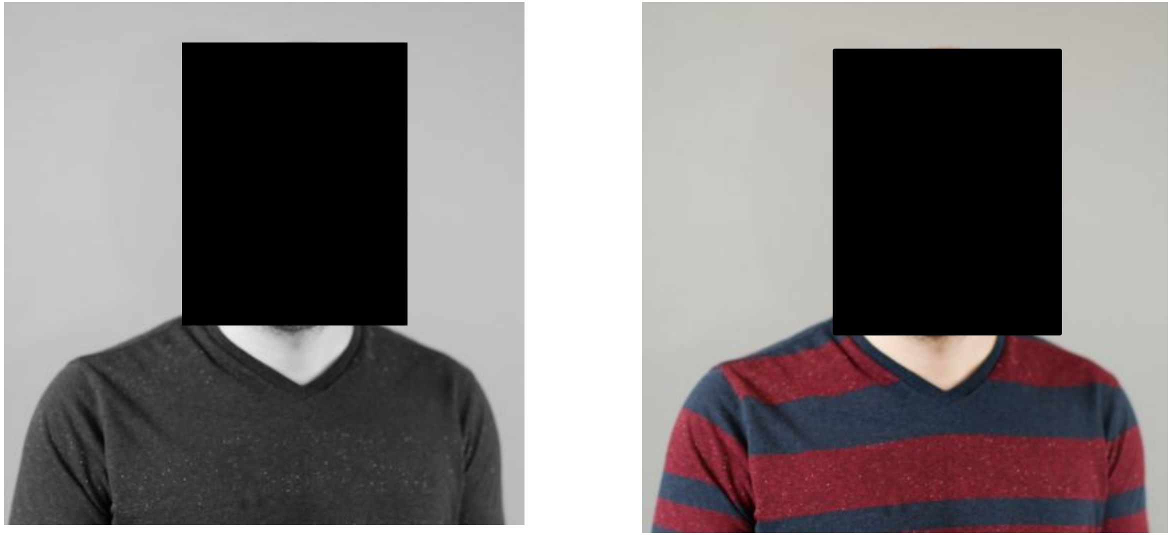
RGB to grayscale

Grayscale conversion “equates” some hues …



RGB to grayscale

Grayscale conversion “equates” some hues . . .



[redacted]

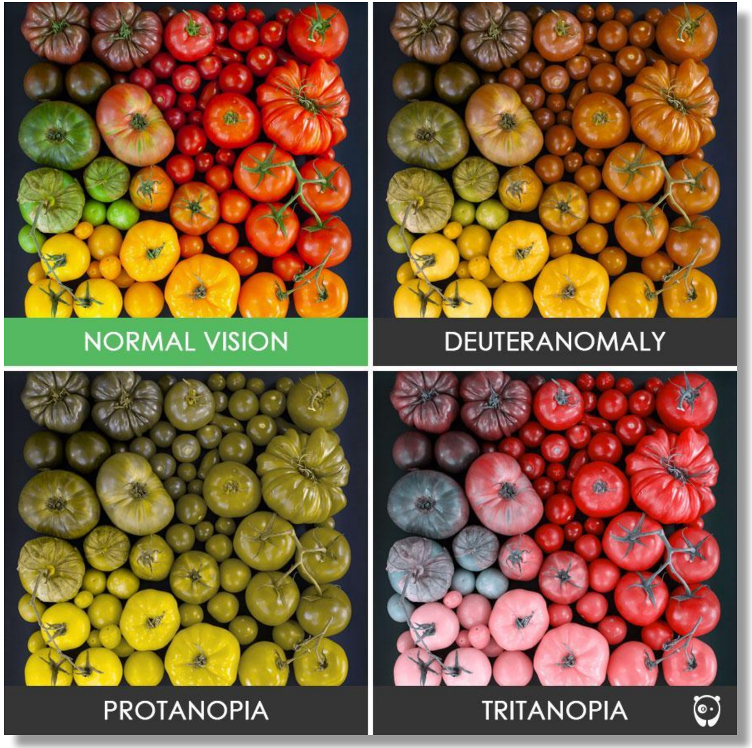
[redacted]

Color blindness

0.5% 8.0%

Instead of red and green
use blue and orange.

Consider color-blind color scale
(e.g., “viridis”)



Count the 2s

5498731840

4893128612

2634085106

1592059852

3854634876

Count the 3s

5498731840

4893128612

2634085106

1592059852

3854634876

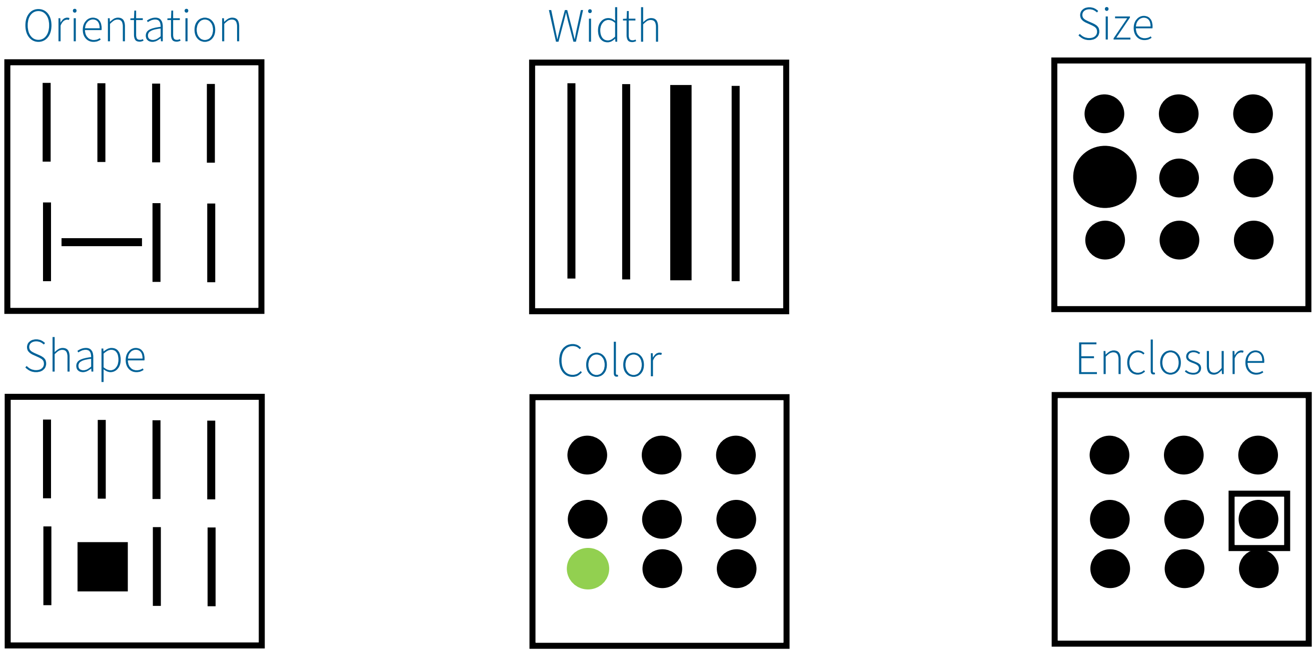
Balance background and foreground

ACVLSIGBSLWUHKAJSLHV

ACVLSIGBSLWUHKAJSLHV

ACVLSIGBSLWUHKAJSLHV

Pre-attentive attributes



Orientation

Width

Size

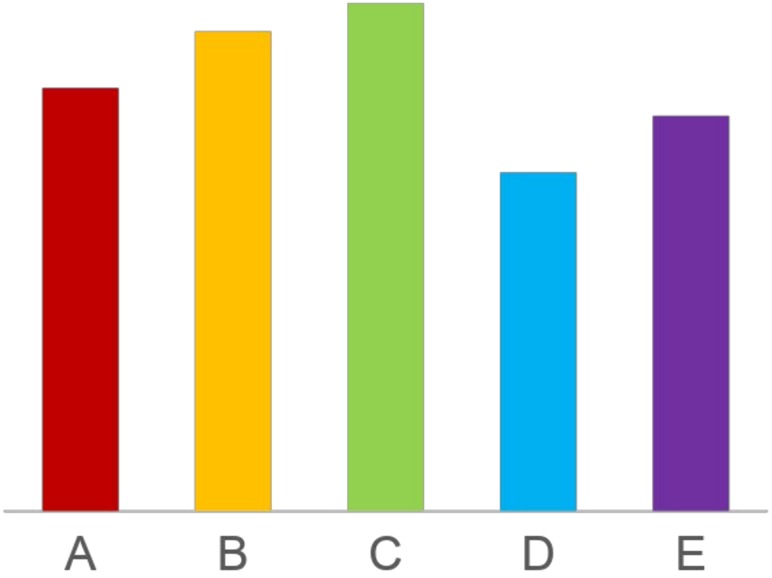
Shape

Color

Enclosure

We perceive these differences unconsciously before we are aware of them

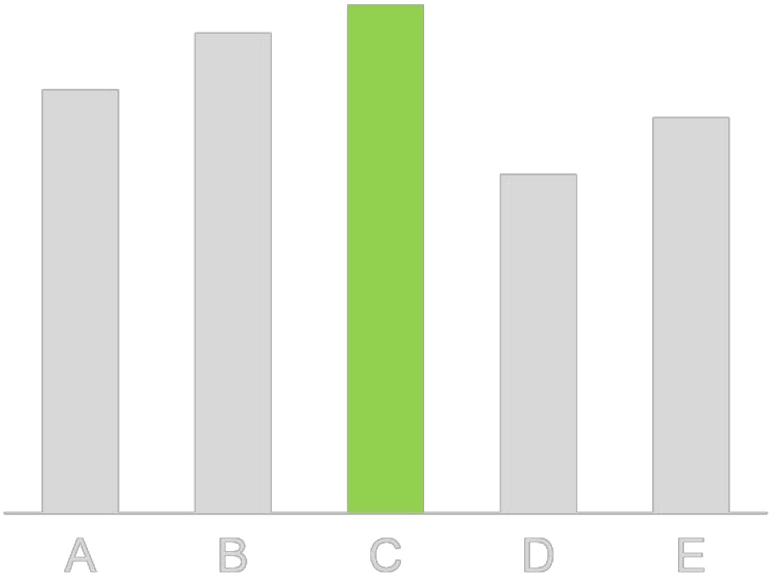
What is the main point?



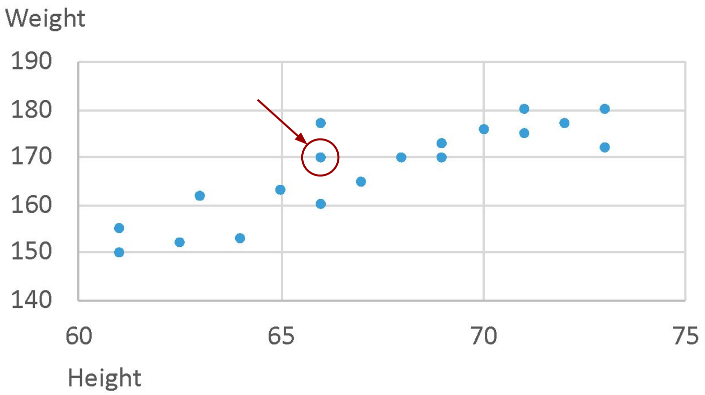
What is the main point?

Highlight your conclusion

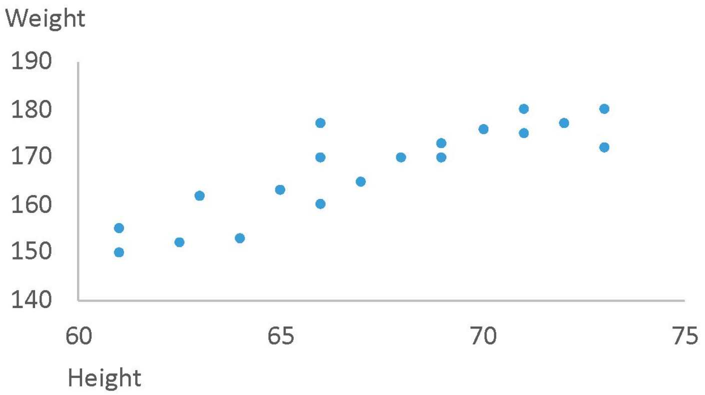
Non-critical information should be removed or put in the background



Gridlines



Keep when illustrating specific numeric values

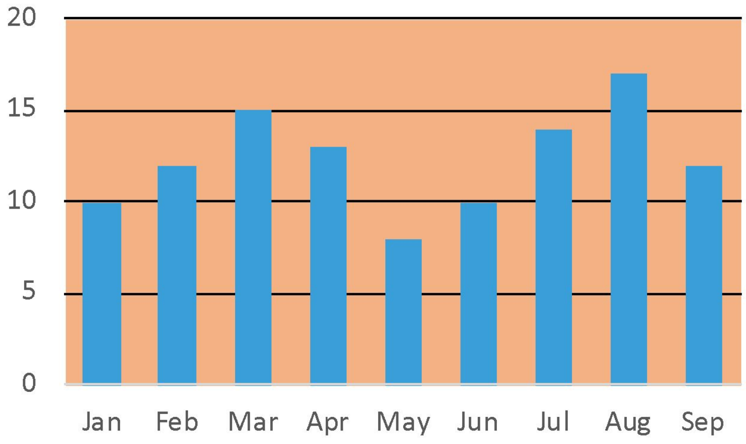


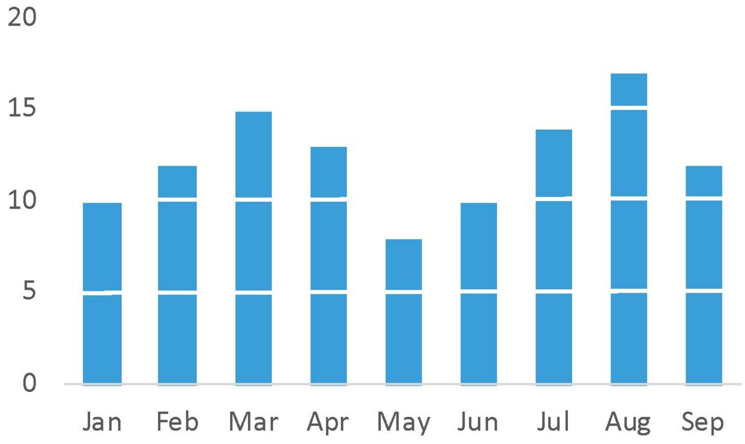
Remove when the “message” is the trend itself

Increase Data-Ink Ratio

Aim for a high Data-to-Ink Ratio:

- Maximize visual elements related explicitly to data
- Minimize extra pixels not related to data (e.g., background color, grids)



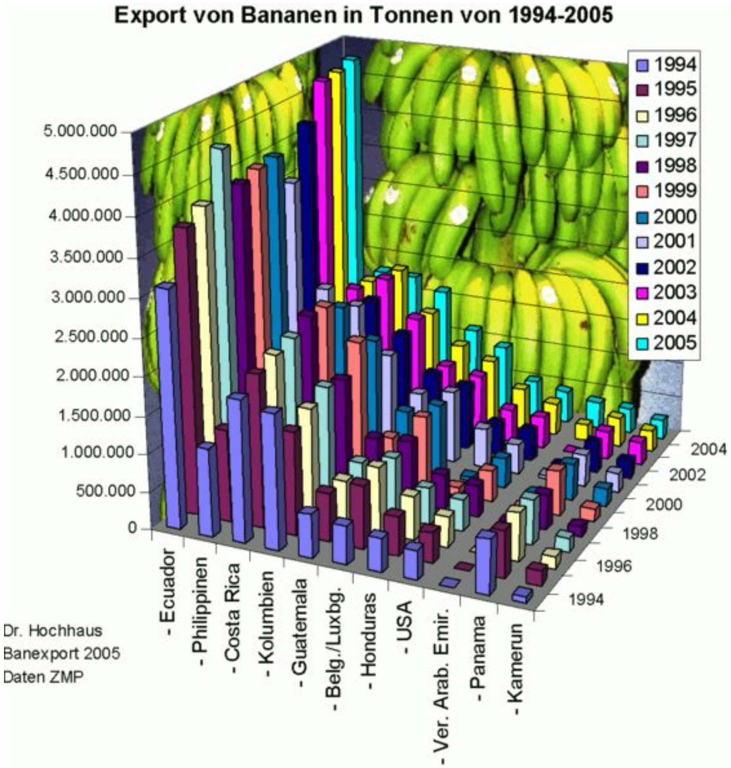


30 second rule

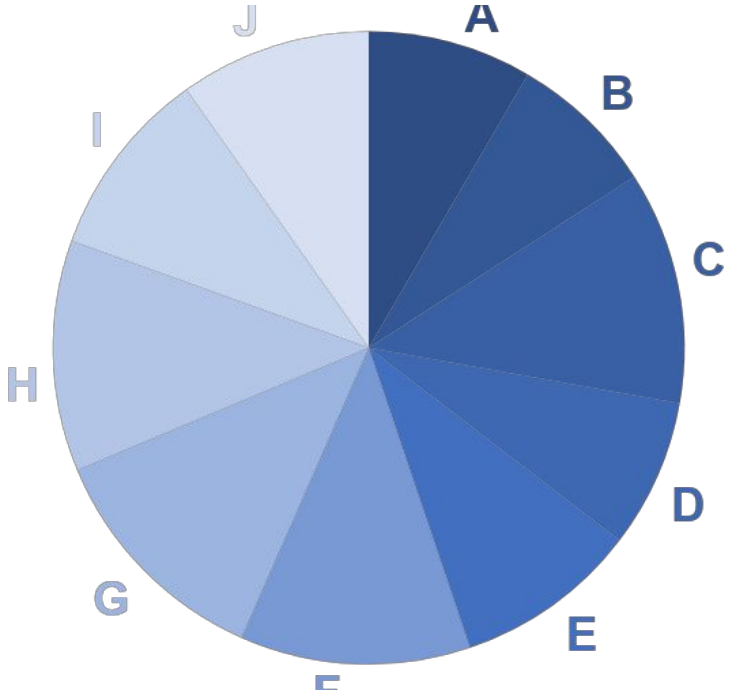
A viewer should be able to interpret the message of a visualization within 30 seconds

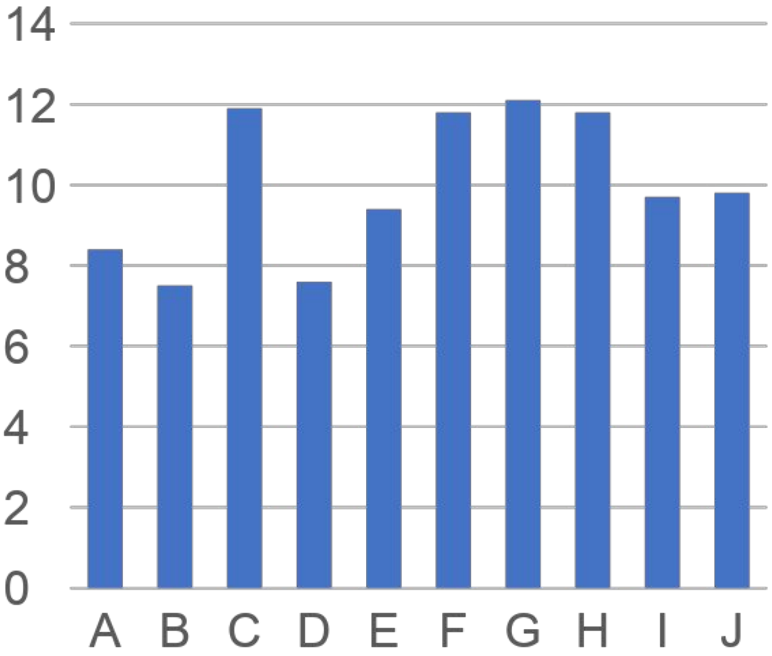
Common Mistakes

Avoid 3D


Export von Bananen in Tonnen von 1994-2005
5.000.000
4.500.000
4.000.000
3.500.000
3.000.000
2.500.000
2.000.000
1.500.000
1.000.000
500.000
0
- Ecuador
- Philippinen
- Costa Rica
- Kolumbien
- Guatemala
- Belg./Luxbg.
- Honduras
- USA
- Ver. Arab. Emir.
- Panama
- Kamerun
1994
1996
1998
2000
2002
2004
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Dr. Hochhaus
Banexport 2005
Daten ZMP

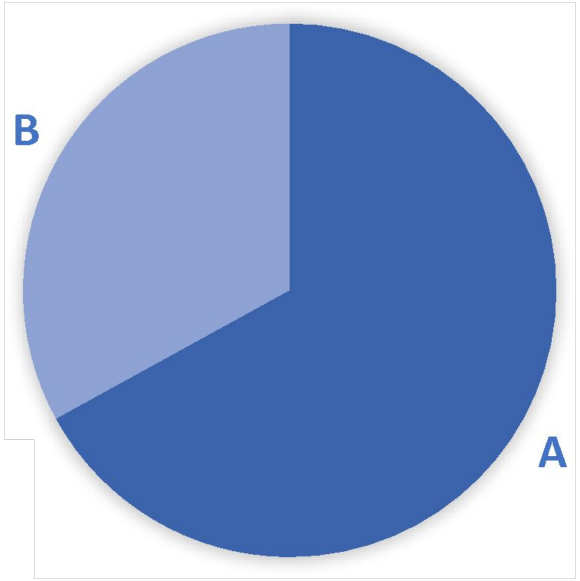
Which is the biggest slice?



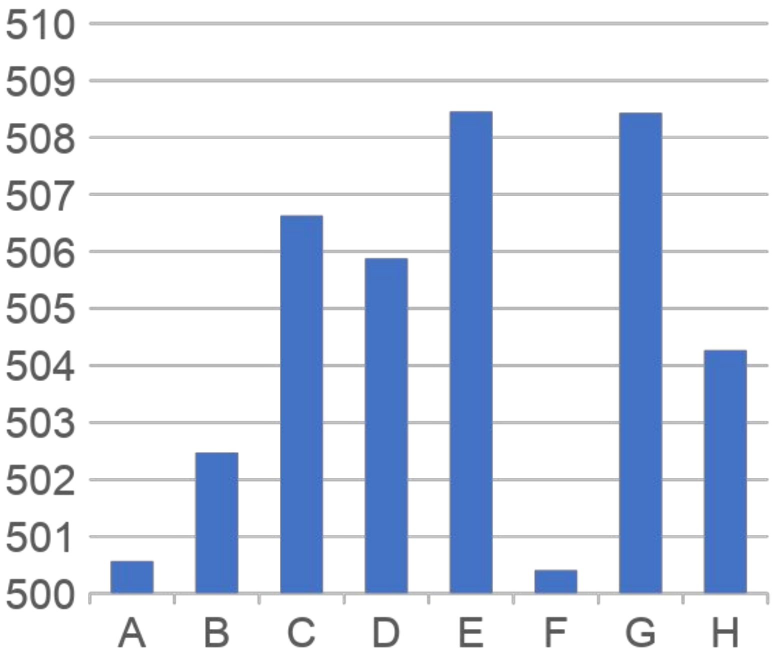


Sometimes pie charts work well

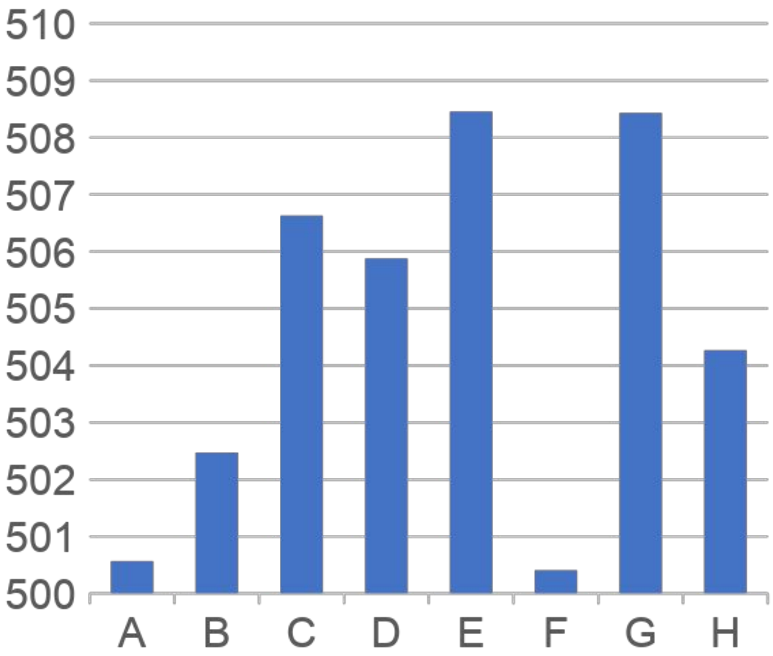
Keep to 2 slices, maximum

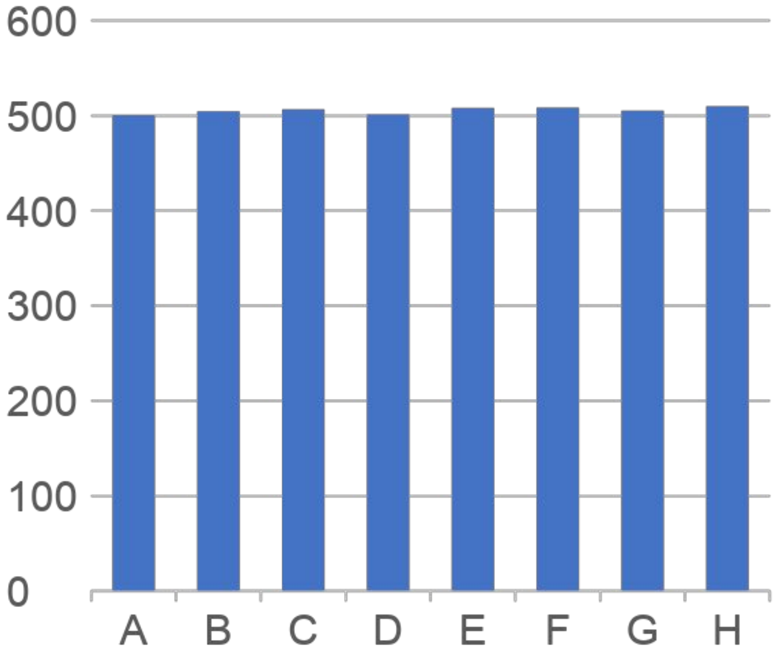


“Scale” requires a zero value



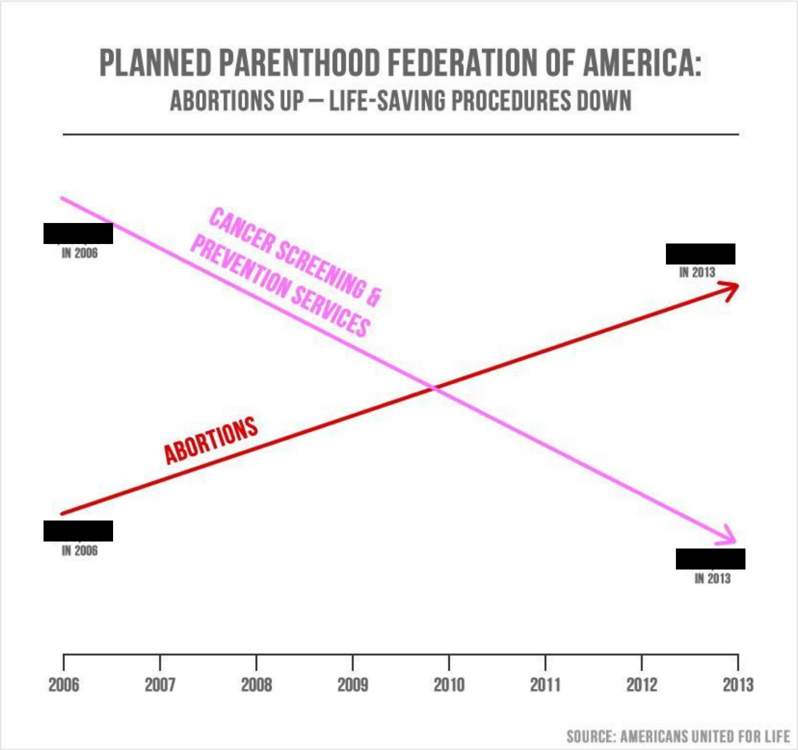
“Scale” requires a zero value





Start numerical axes at zero (or include a zero line)

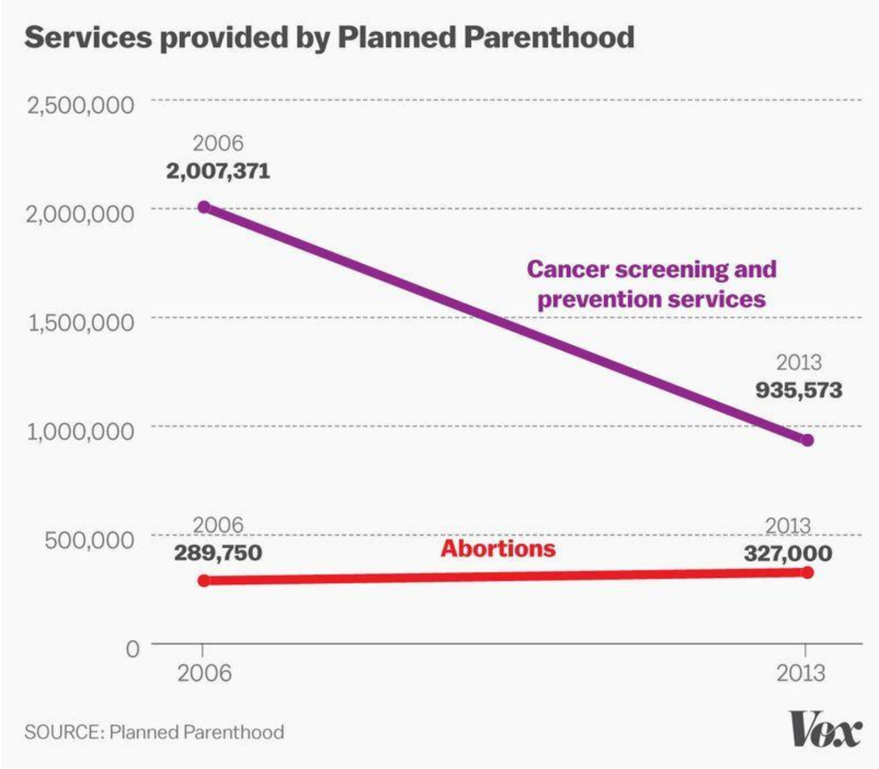
Example



PLANNED PARENTHOOD FEDERATION OF AMERICA:

ABORTIONS UP — LIFE-SAVING PROCEDURES DOWN

SOURCE: AMERICANS UNITED FOR LIFE

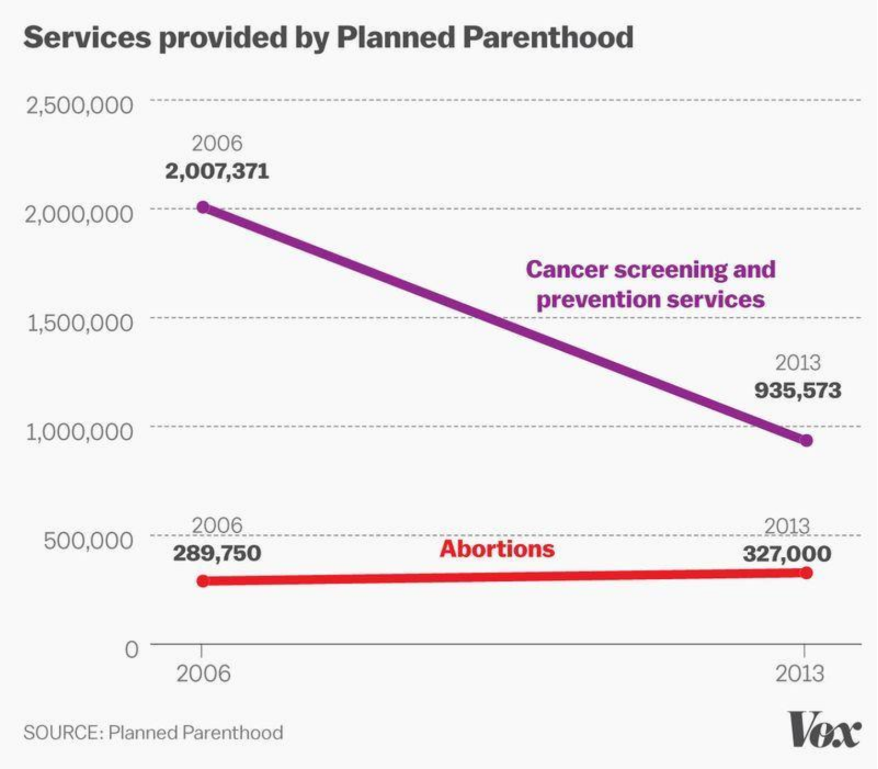


Services provided by Planned Parenthood

SOURCE: Planned Parenthood

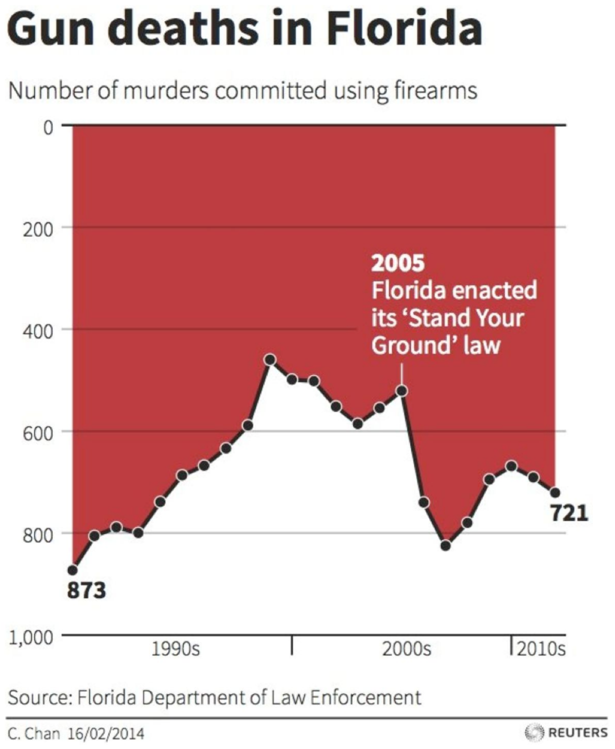
Example

Synchronize the scale of **both** axes



Example

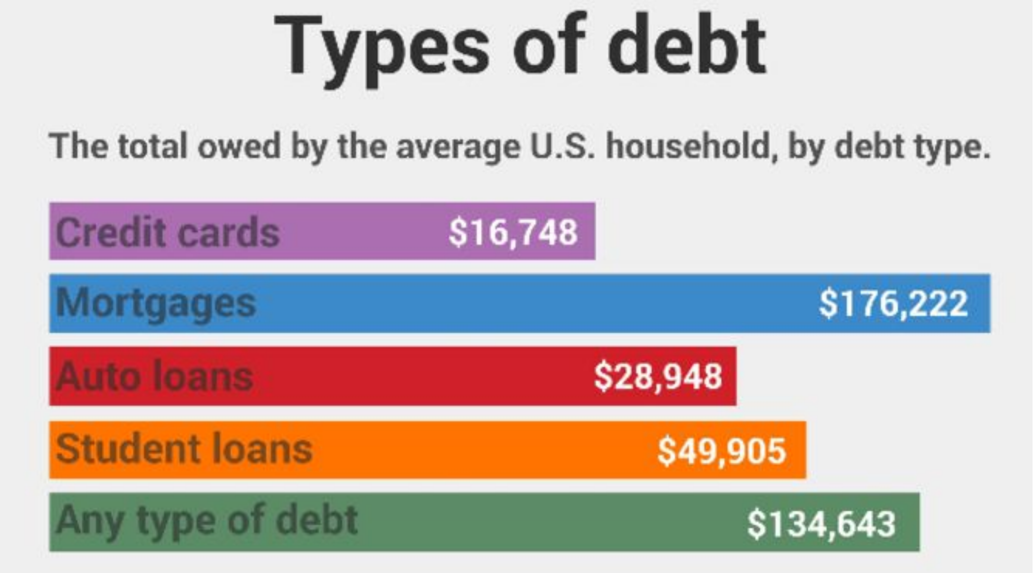
Avoid inverted axes



Bad Examples

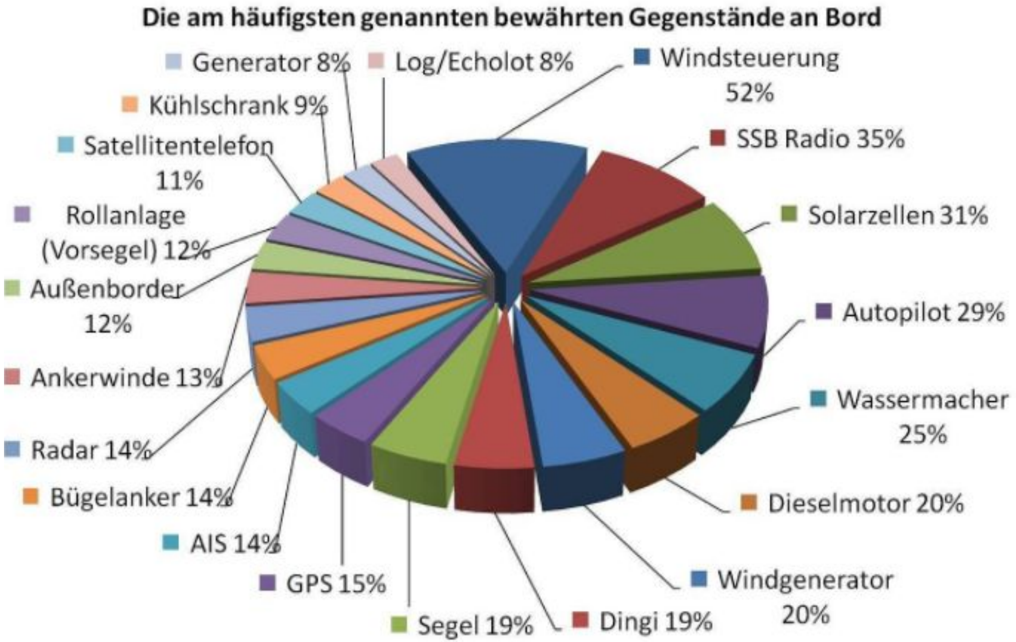
What is wrong with this visual?

- Readability
- Colors



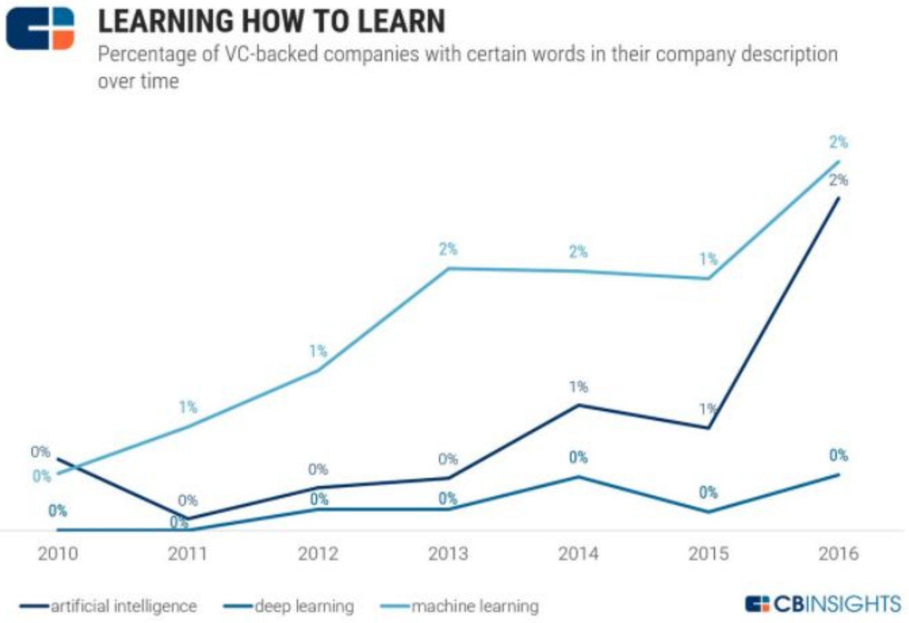
What is wrong with this visual?

- 3D
- Too many slices
- >100%



What is wrong with this visual?

- Decimals
- y-axis

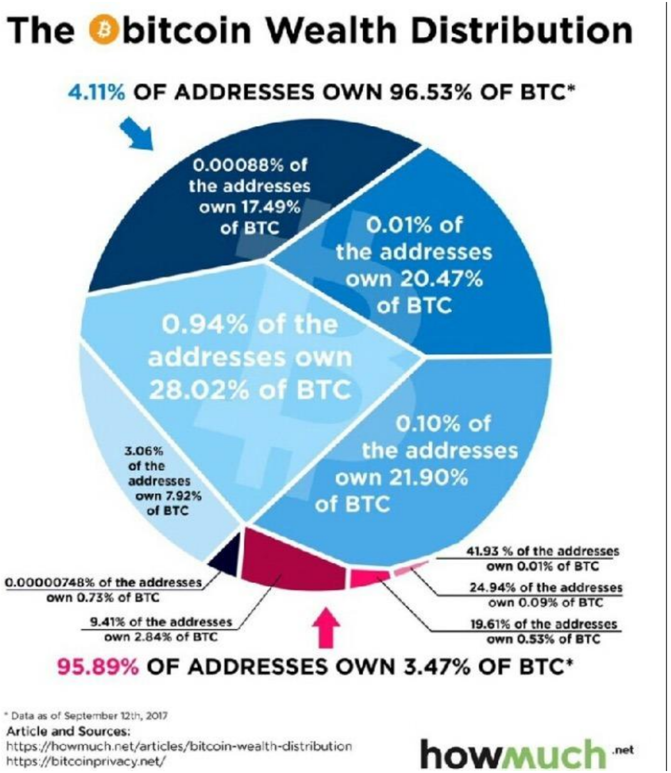


LEARNING HOW TO LEARN

Percentage of VC-backed companies with certain words in their company description over time

What is wrong with this visual?

- 30-second rule
- … What is happening?



The bitcoin Wealth Distribution

* Data as of September 12th, 2017

Article and Sources:
https://howmuch.net/articles/bitcoin-wealth-distribution
https://bitcoinprivacy.net/

Oh no ...

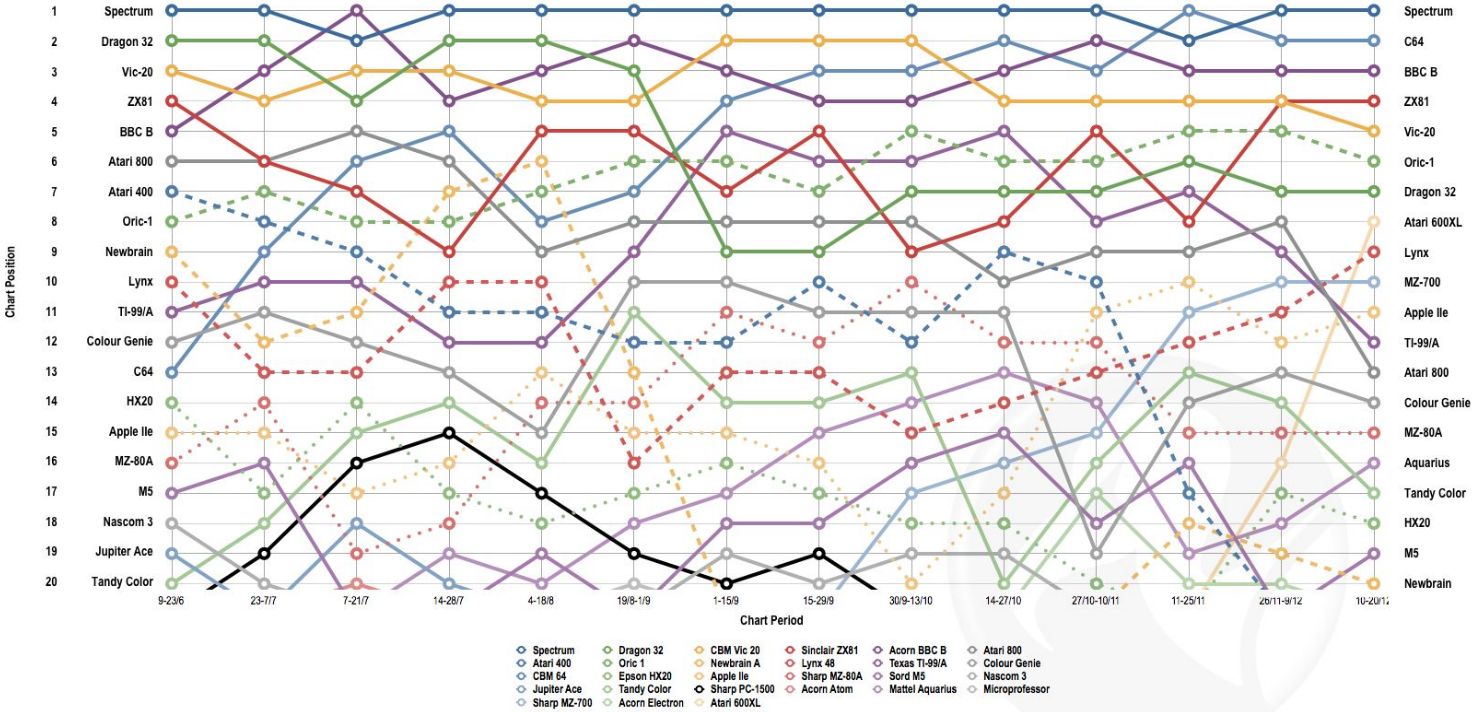


Chart Position
1 Spectrum
2 Dragon 32
3 Vic-20
4 ZX81
5 BBC B
6 Atari 800
7 Atari 400
8 Oric-1
9 Newbrain
10 Lynx
11 TI-99/A
12 Colour Genie
13 C64
14 HX20
15 Apple IIe
16 MZ-80A
17 M5
18 Nascom 3
19 Jupiter Ace
20 Tandy Color
Spectrum
C64
BBC B
ZX81
Vic-20
Oric-1
Dragon 32
Atari 600XL
Lynx
MZ-700
Apple IIe
TI-99/A
Atari 800
Colour Genie
MZ-80A
Aquarius
Tandy Color
HX20
M5
Newbrain
9-23/6
23-7/7
7-21/7
14-28/7
4-18/8
19/8-1/9
1-15/9
15-29/9
30/9-13/10
14-27/10
27/10-10/11
11-25/11
26/11-9/12
10-20/1
Chart Period
Spectrum
Atari 400
CBM 64
Jupiter Ace
Sharp MZ-700
Dragon 32
Oric 1
Epson HX20
Tandy Color
Acorn Electron
CBM Vic 20
Newbrain A
Apple IIe
Sharp PC-1500
Atari 600XL
Sinclair ZX81
Lynx 48
Sharp MZ-80A
Acorn Atom
Acorn BBC B
Texas TI-99/A
Sord M5
Mattel Aquarius
Atari 800
Colour Genie
Nascom 3
Microprofessor

Tools

Visualization Tools

	Excel	Seaborn	Plotly	Tableau	D3
Interactivity	Static	Static	Interactive	Interactive	Interactive
Difficulty	Easy	Medium	Medium	Easy	Difficult
Pros	Popular	Python	Python	Beautiful	Custom
Cons	Basic	Python	Python	$$$	Time Consuming

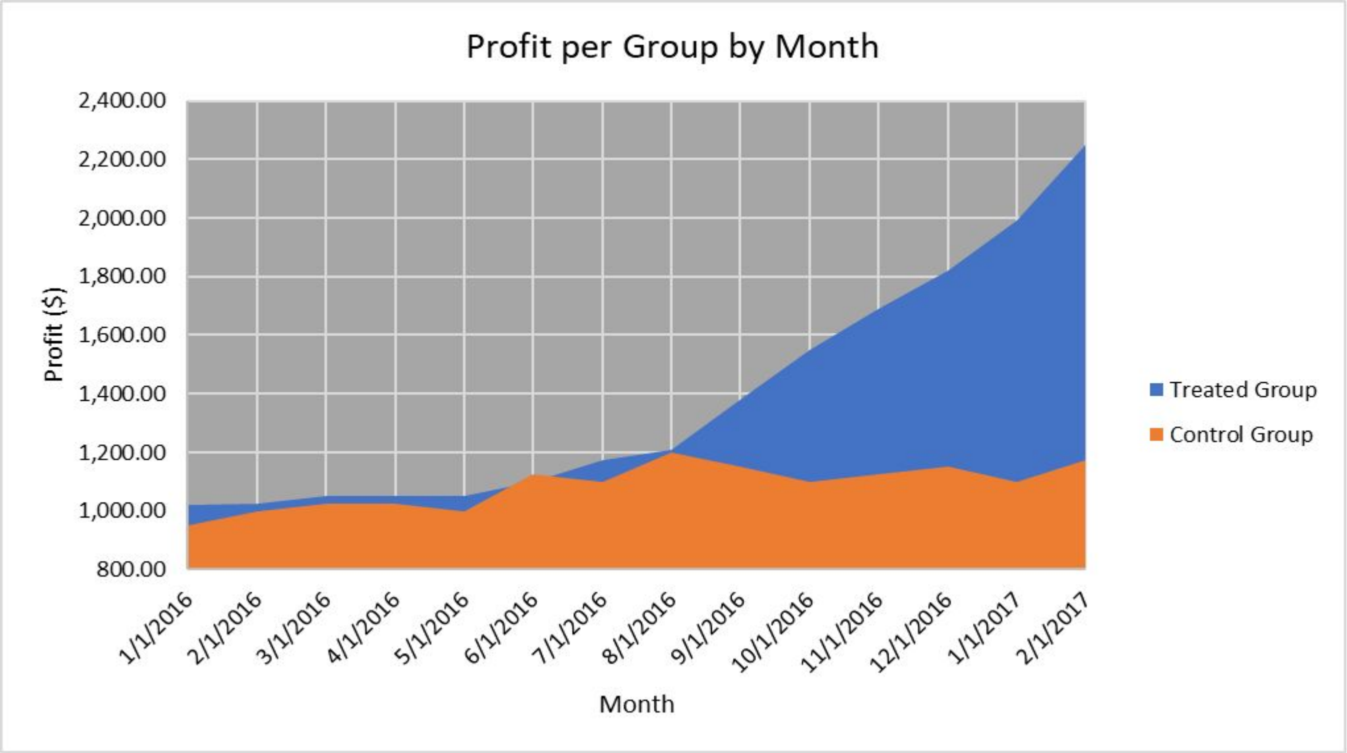
Visualization Tools

	Excel	Seaborn	Plotly	Tableau	D3
Interactivity	Static	Static	Interactive	Interactive	Interactive
Difficulty	Easy	Medium	Medium	Easy	Difficult
Pros	Popular	Python	Python	Beautiful	Custom
Cons	Basic	Python	Python	$$$	Time Consuming

Visualization Makeover

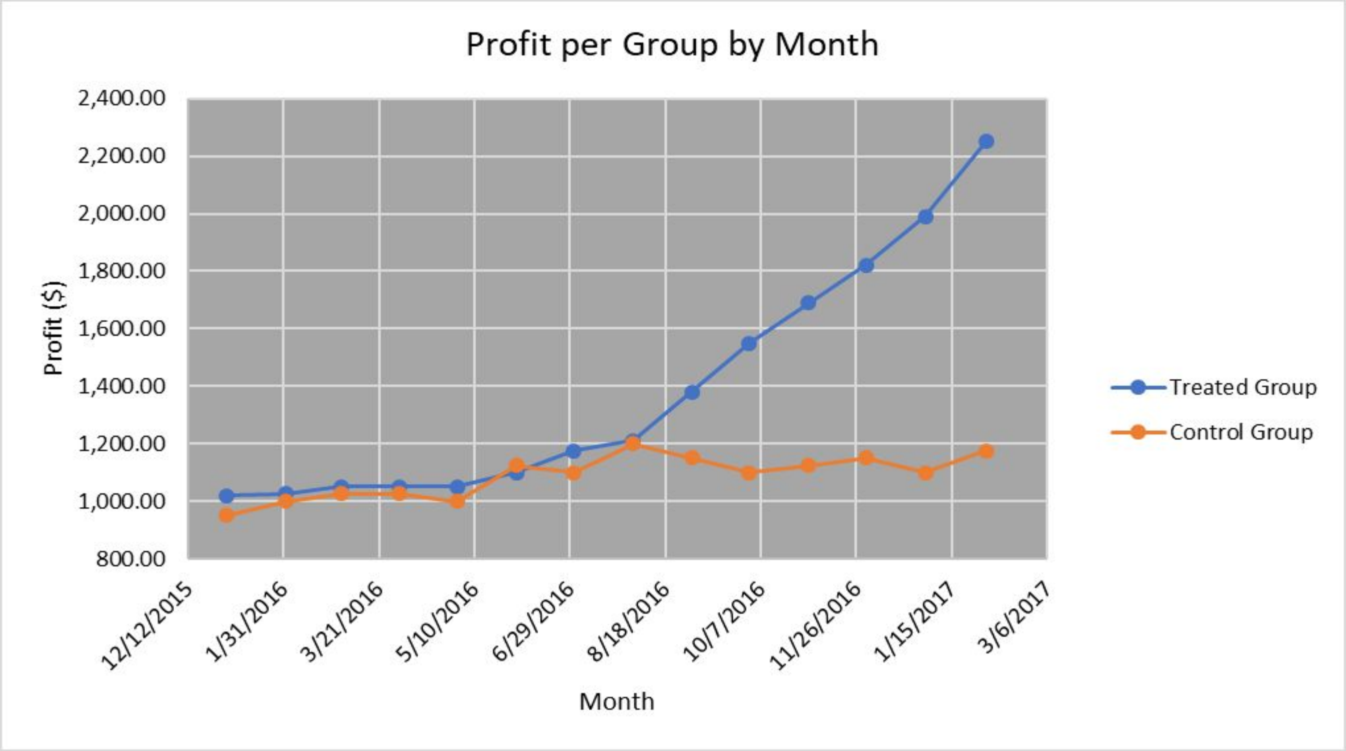
Example

Use line chart!



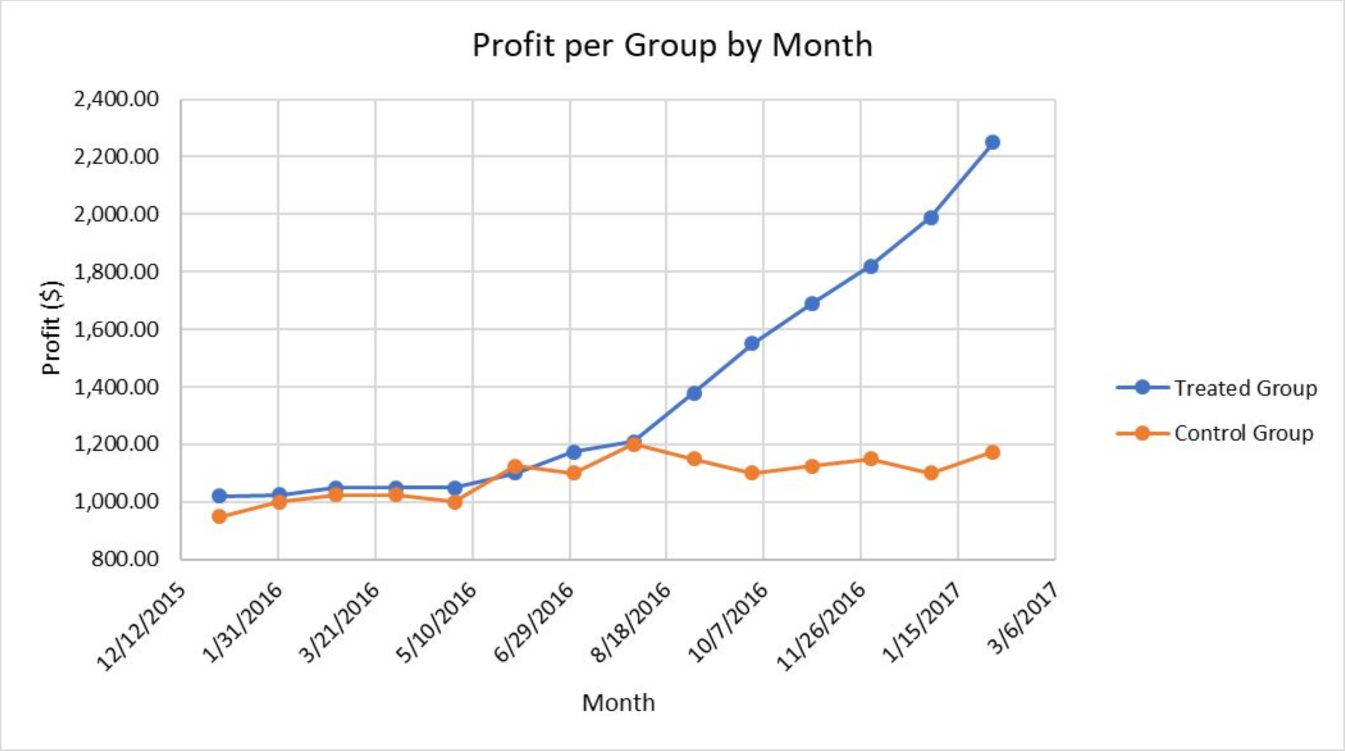
Example

Remove background color!



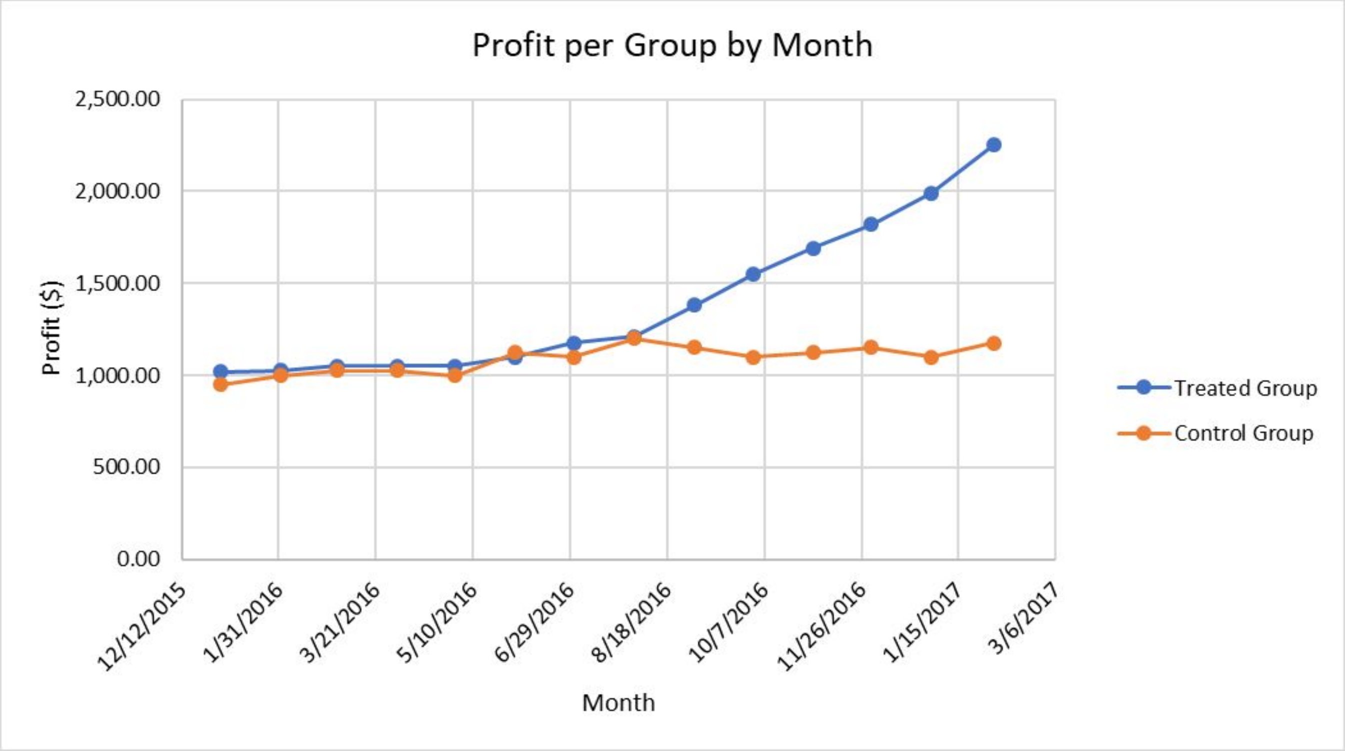
Example

Start axis at 0!



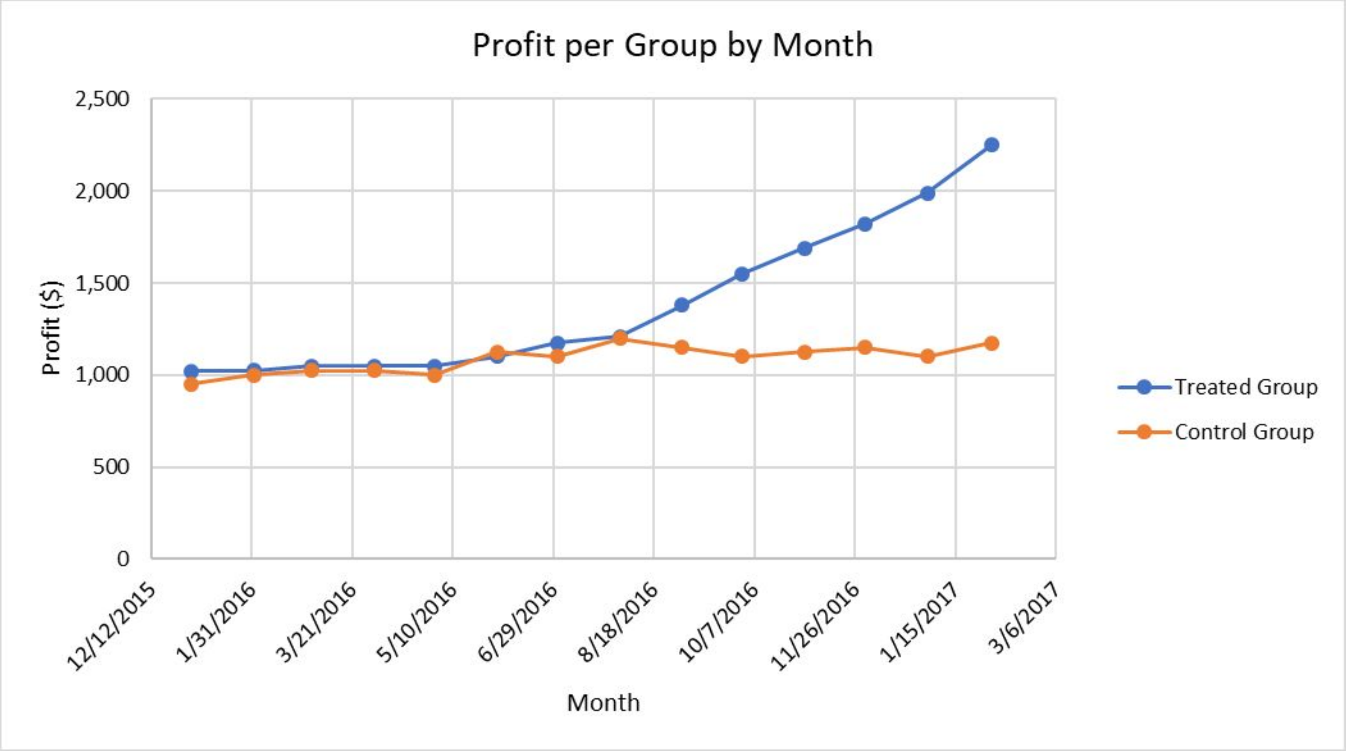
Example

Decimal points!



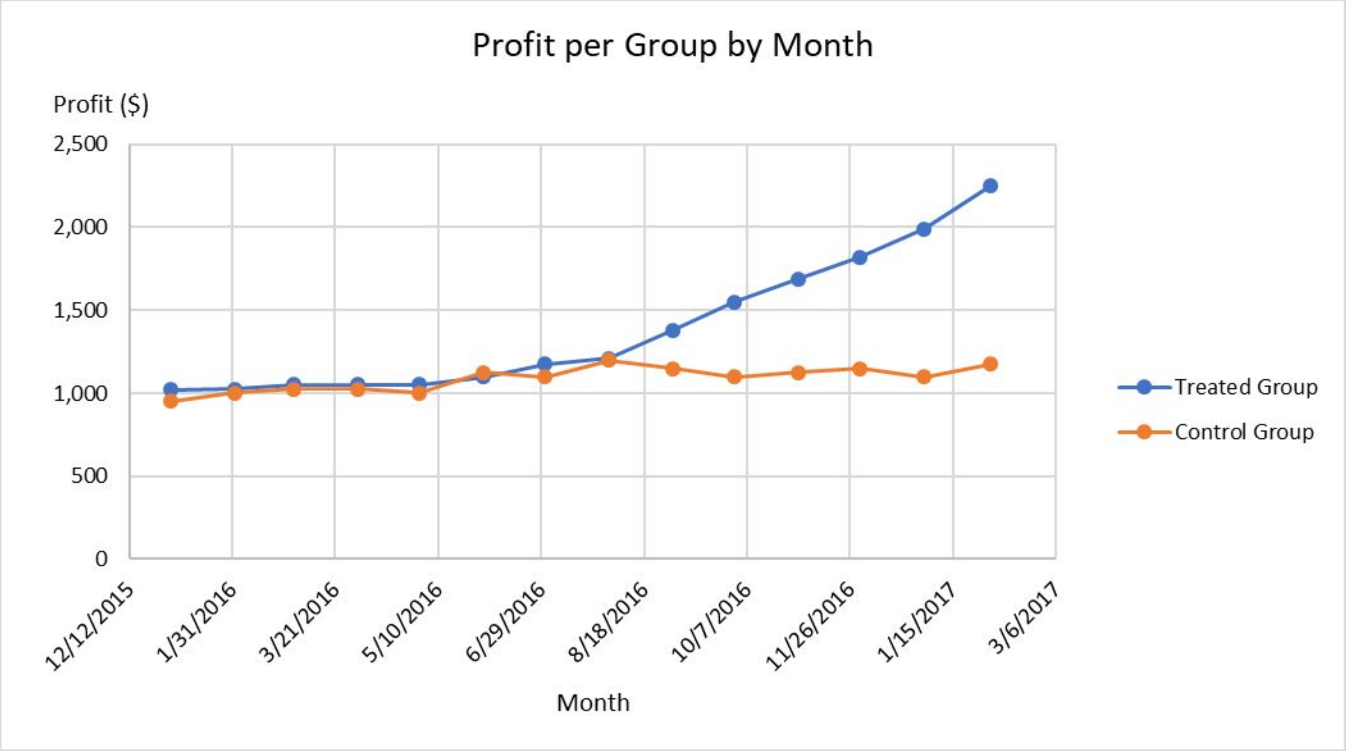
Example

Vertical “Profit” label!



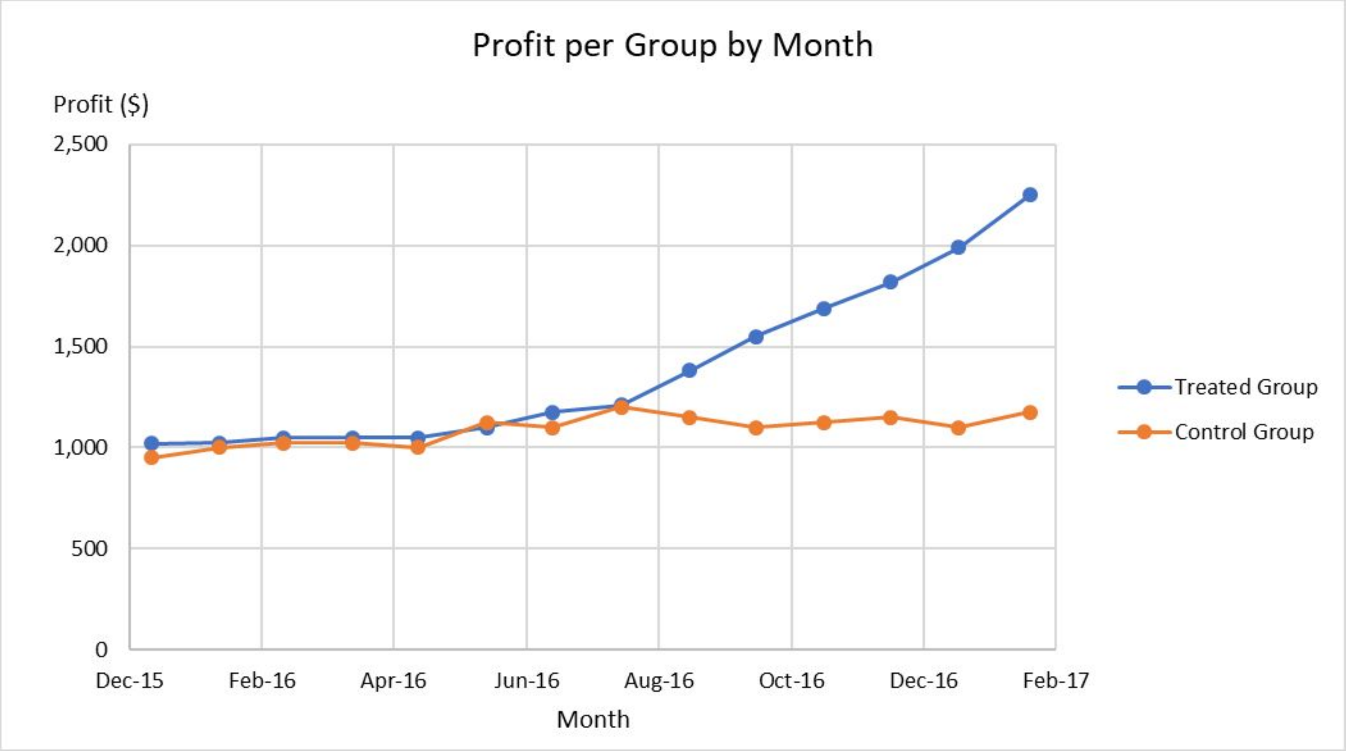
Example

Diagonal x-axis!



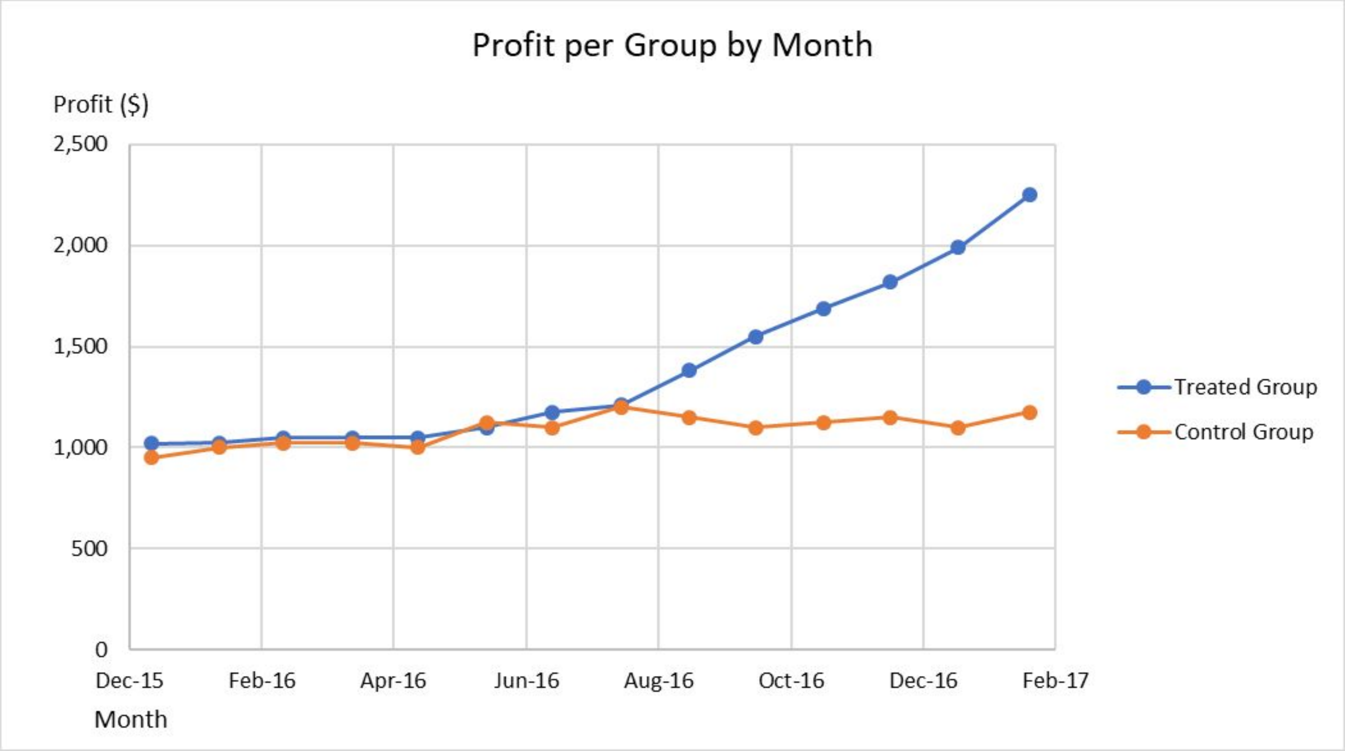
Example

X-Label left justified!



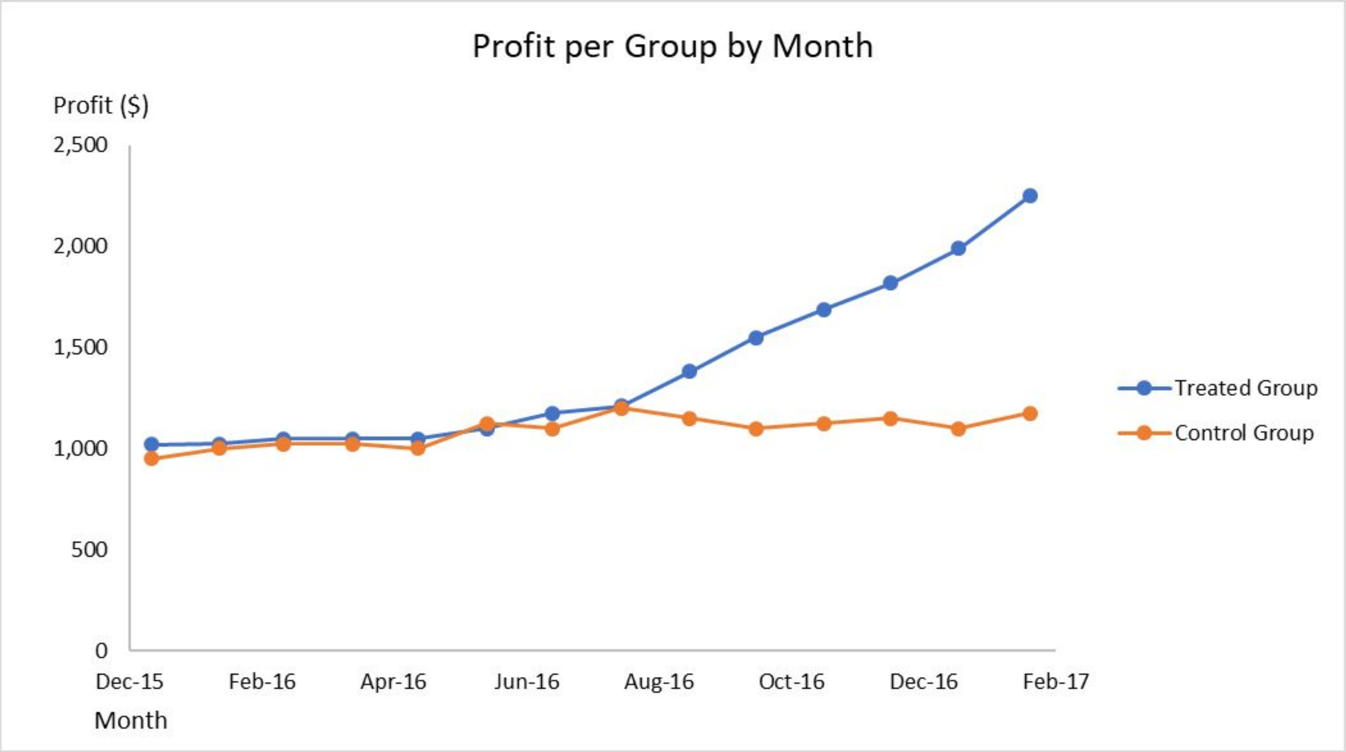
Example

Gridlines!



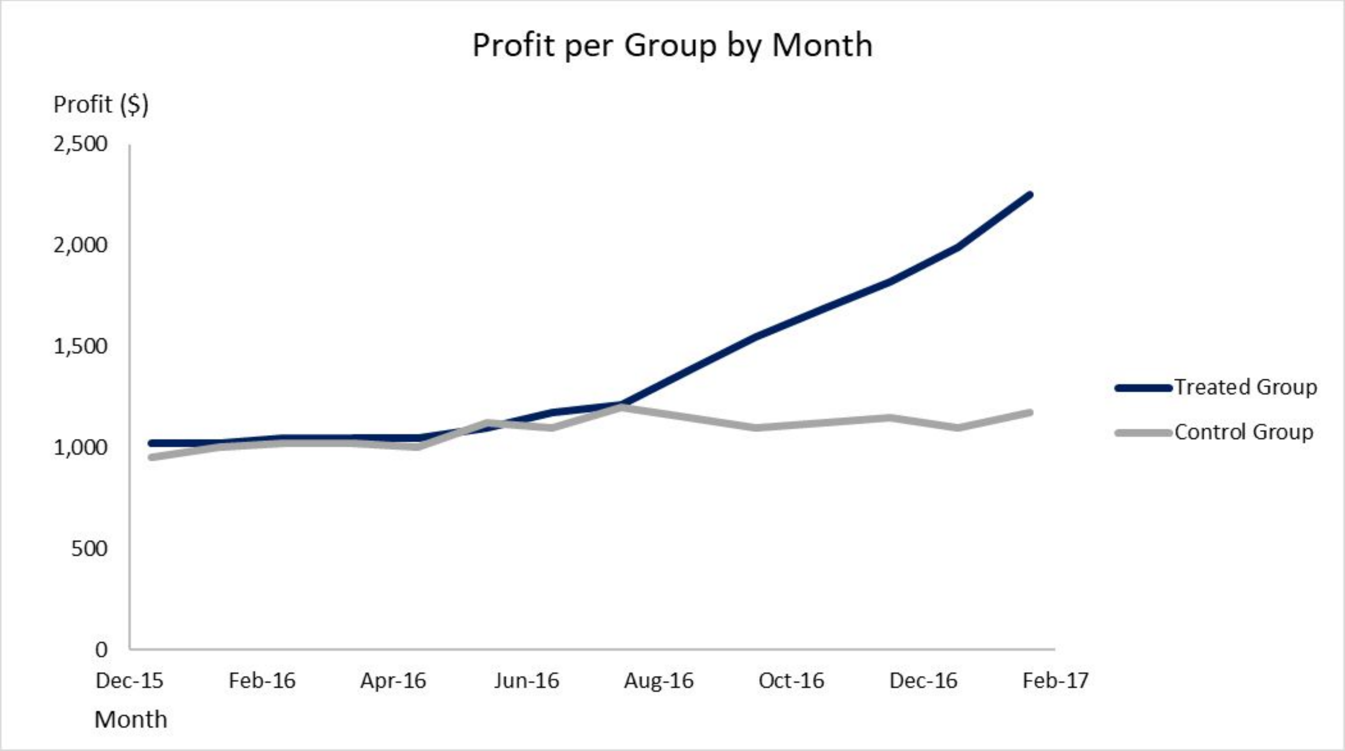
Example

Color and dots! (this is a *binary* time series)



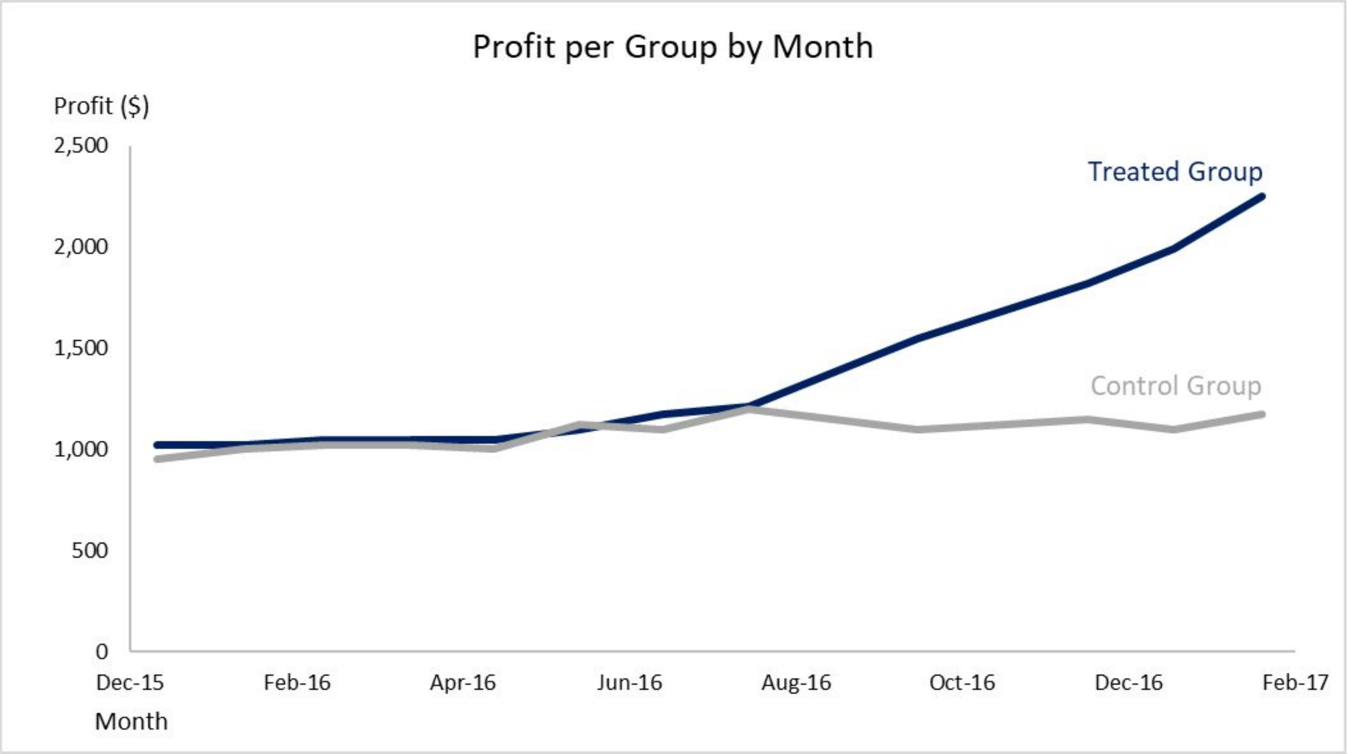
Example

Legend, proximity and space



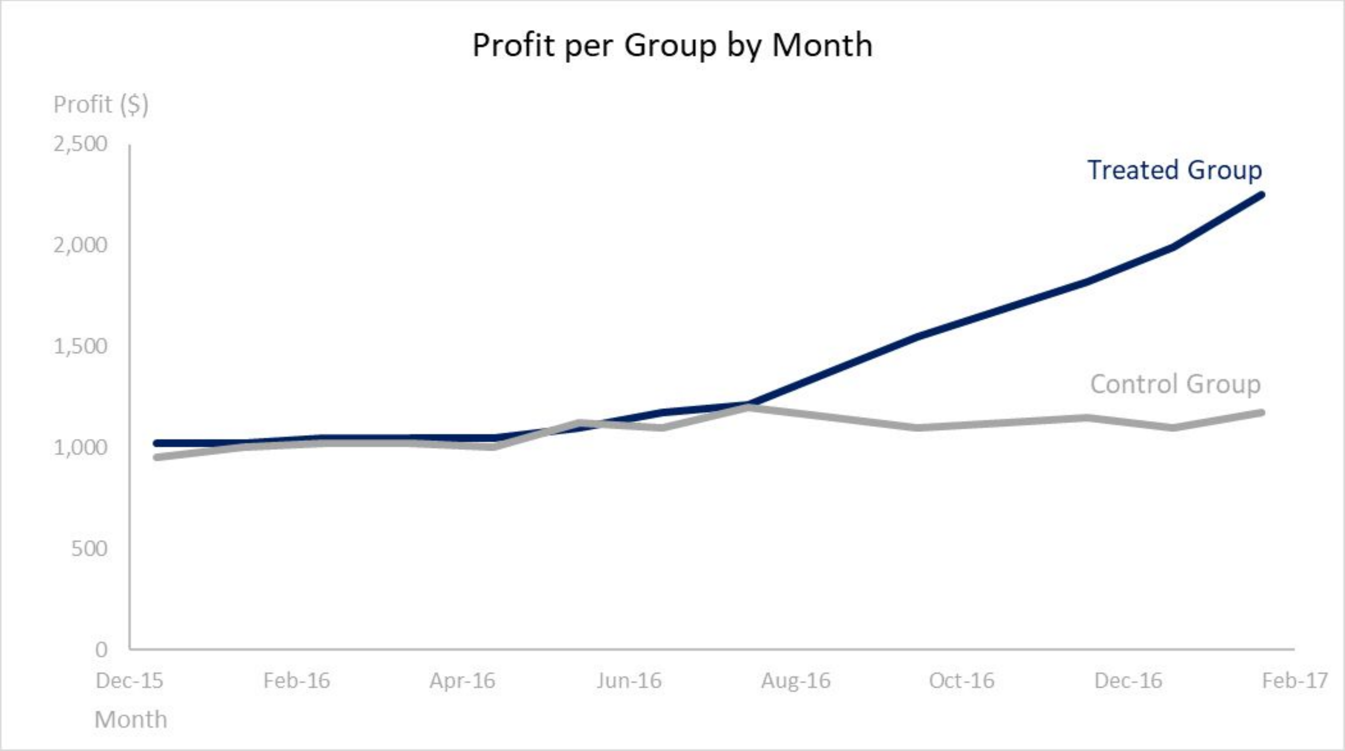
Example

Axis background



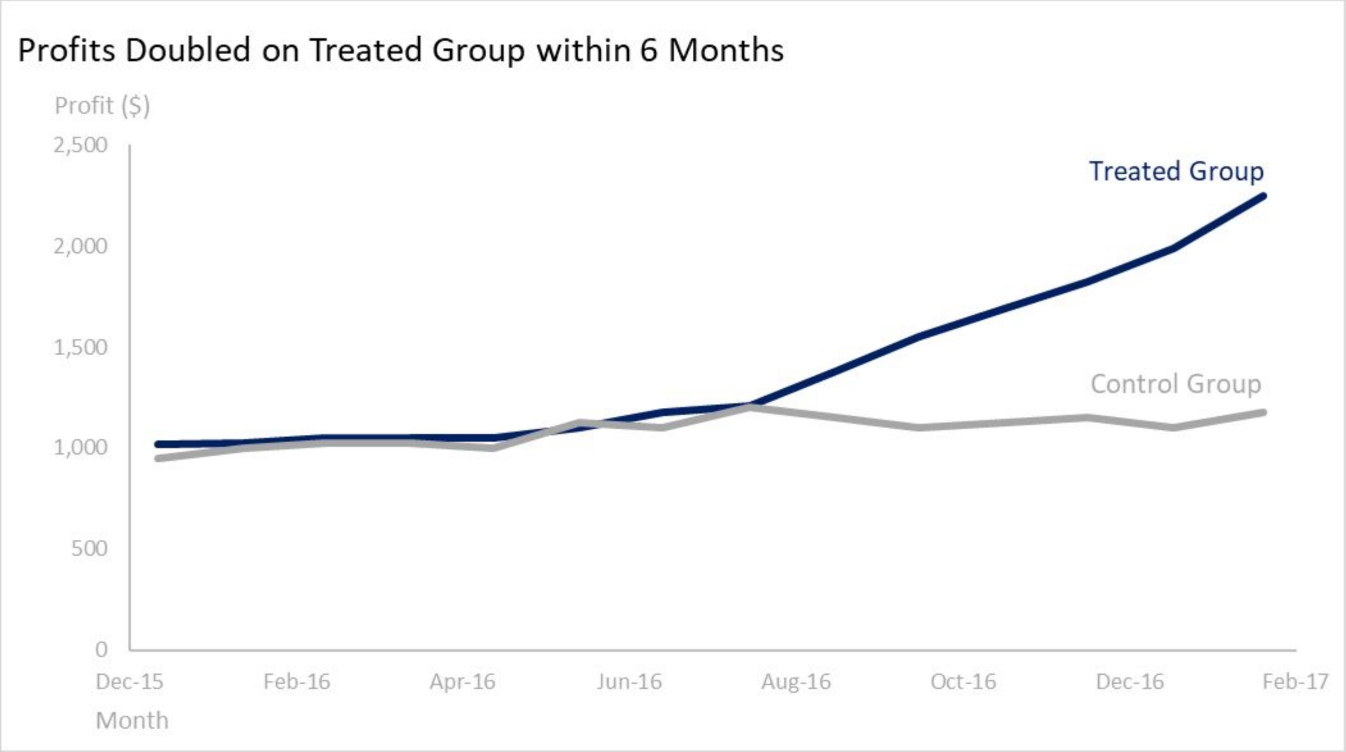
Example

Descriptive title and left!



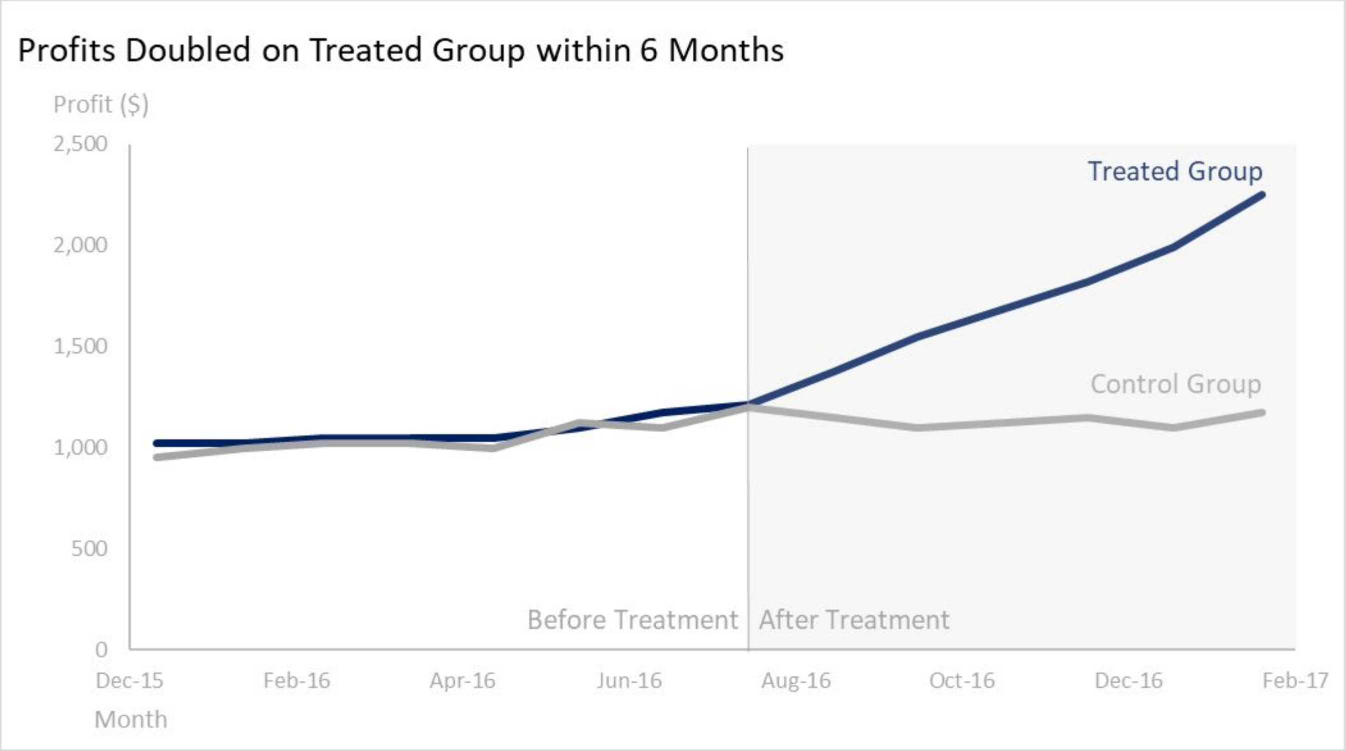
Example

Highlight!

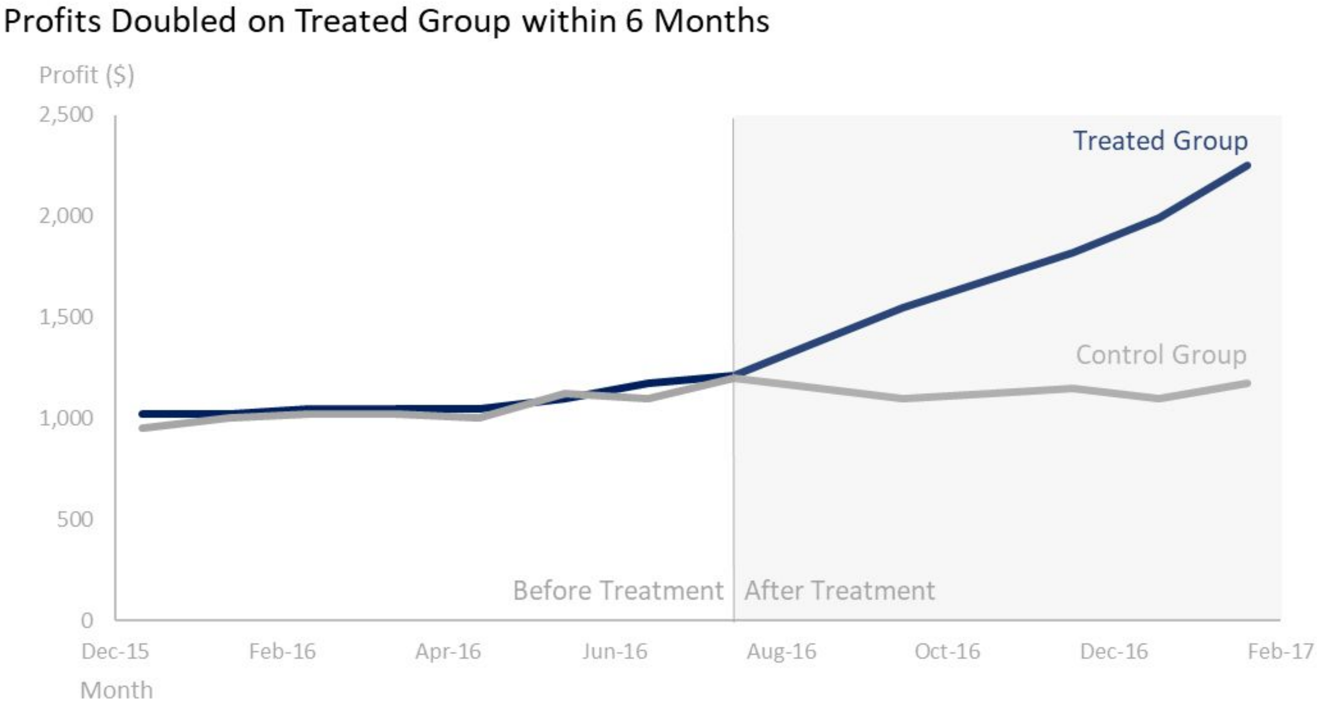


Example

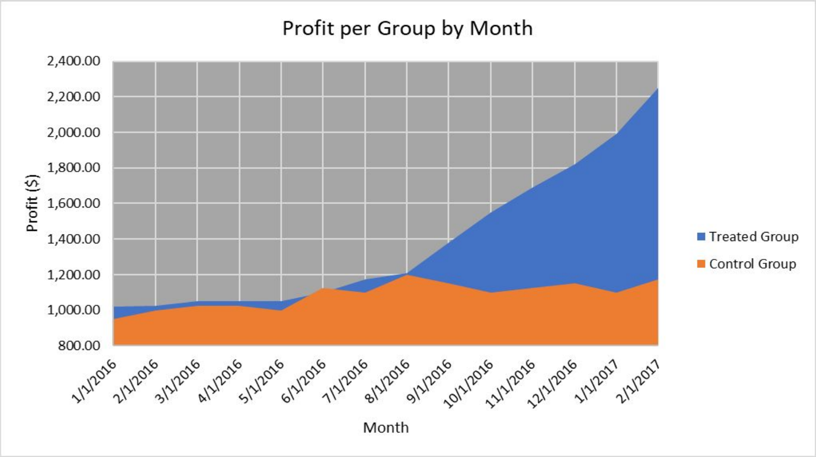
Box!

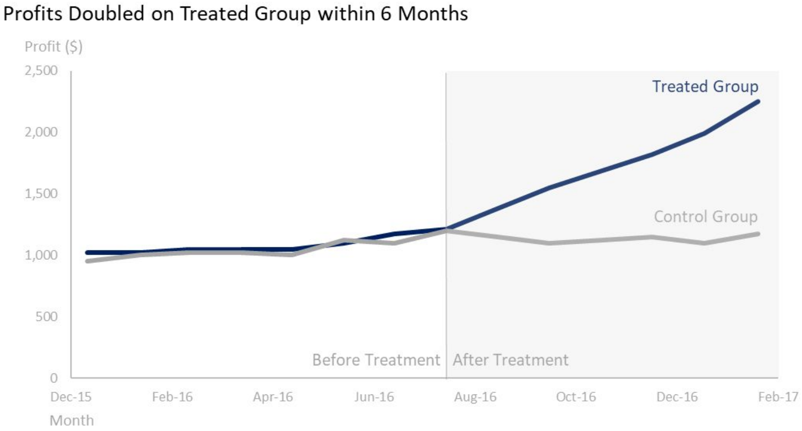

Profits Doubled on Treated Group within 6 Months
Profit ($)
2,500
2,000
1,500
1,000
500
0
Treated Group
Control Group
Before Treatment
After Treatment
Dec-15
Feb-16
Apr-16
Jun-16
Aug-16
Oct-16
Dec-16
Feb-17
Month

Example


Profits Doubled on Treated Group within 6 Months
Profit ($)
2,500
2,000
1,500
1,000
500
0
Treated Group
Control Group
Before Treatment
After Treatment
Dec-15
Feb-16
Apr-16
Jun-16
Aug-16
Oct-16
Dec-16
Feb-17
Month

Example


Profit per Group by Month
2,400.00
2,200.00
2,000.00
1,800.00
1,600.00
1,400.00
1,200.00
1,000.00
800.00
Profit ($)
1/1/2016
2/1/2016
3/1/2016
4/1/2016
5/1/2016
6/1/2016
7/1/2016
8/1/2016
9/1/2016
10/1/2016
11/1/2016
12/1/2016
1/1/2017
2/1/2017
Month
Treated Group
Control Group


Profits Doubled on Treated Group within 6 Months
Profit ($)
2,500
2,000
1,500
1,000
500
0
Treated Group
Control Group
Before Treatment
After Treatment
Dec-15
Feb-16
Apr-16
Jun-16
Aug-16
Oct-16
Dec-16
Feb-17
Month

Summary

- Keep design thinking in mind
  - Charts should be easy to understand
  - Visuals should be discoverable
- Use text, color, and highlighting techniques
- Avoid unnecessary ink
  - Avoid 3D
  - Pie charts with too many slices
- Don’t lie with data
- Don’t use defaults – spend some time making over your visualizations

“Above all else, show the data.”

~ Tufte, 1983