Strategies for Effective Data Visualization

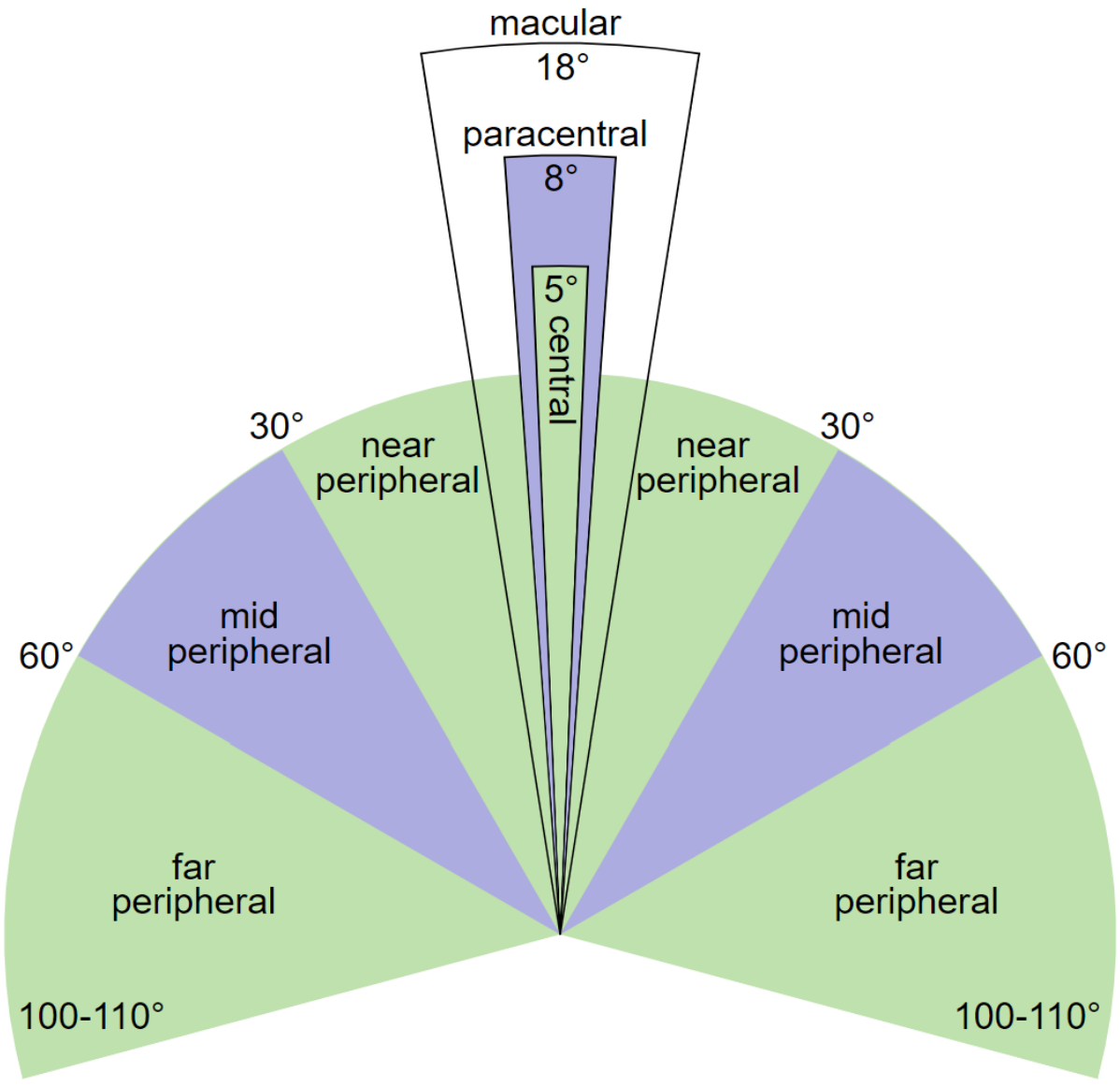
“Humans are pattern-seeking story-telling animals, and we are quite adept at telling stories about patterns, whether they exist or not.”

Michael Shermer

Plan of the Lecture

1. Human vision and visual attention
2. Advantages of visual representation of data
3. Doing it wrong: bad & misleading charts
4. Color
5. Plotting tutorial and exercises

Basics of Human Vision



- Central vision is narrow
- Peripheral vision is low res
- Our eyes jump all over an image and the brain creates an illusion of a broader visual field

Image credit: Zyxwv99, Wikipedia

Hierarchy of Visual Attention

- Our eyes jump from one part of an image that stands out to another
- Pre-attentive processing filters noise or irrelevant stimuli helping to maintain focus
- We are hard-wired to look at what differs from the 'average':
    1. Contrast
    2. Color
    3. Shape, size, orientation
    4. Patterns and connections

Pre-attentive Attributes

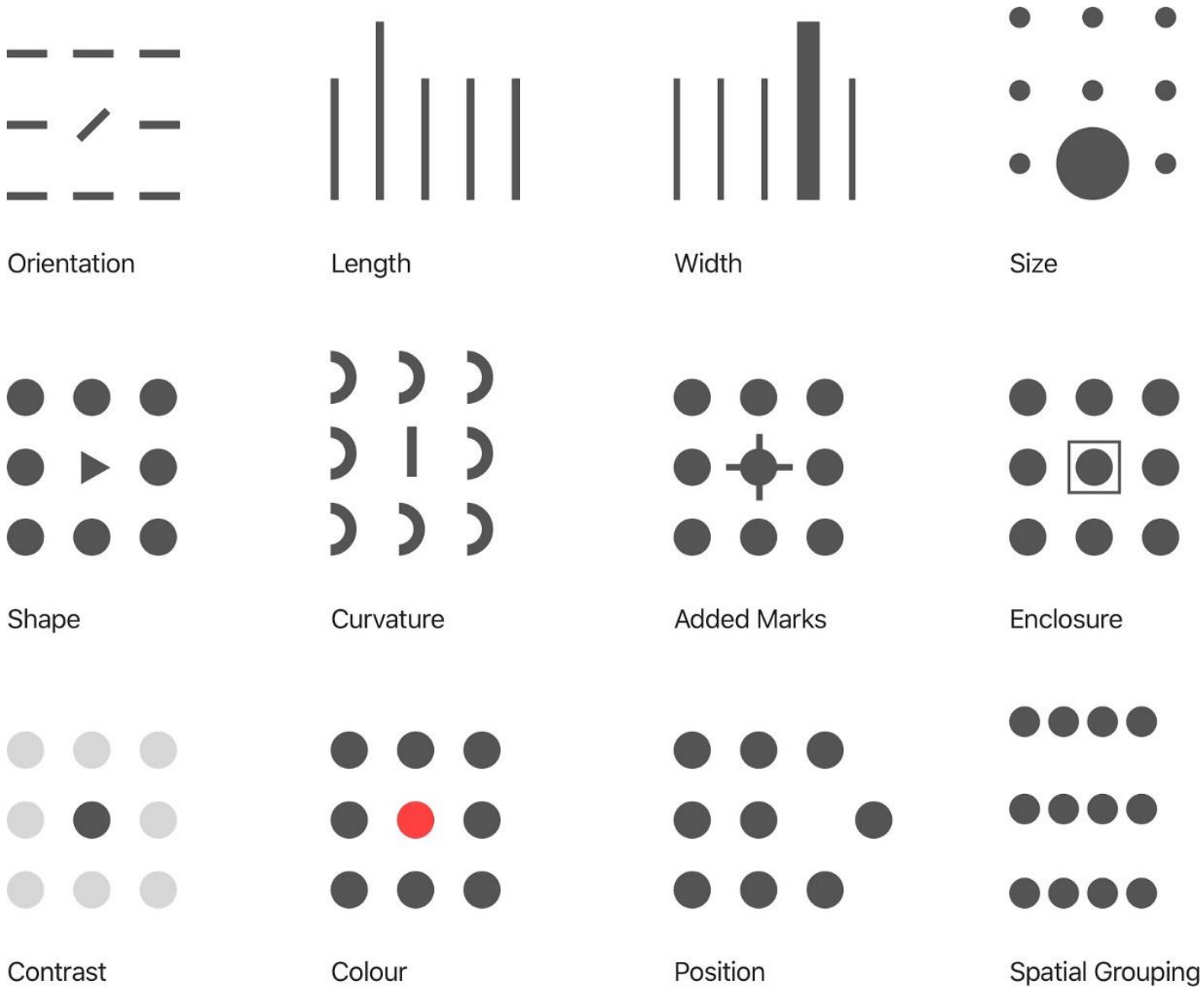
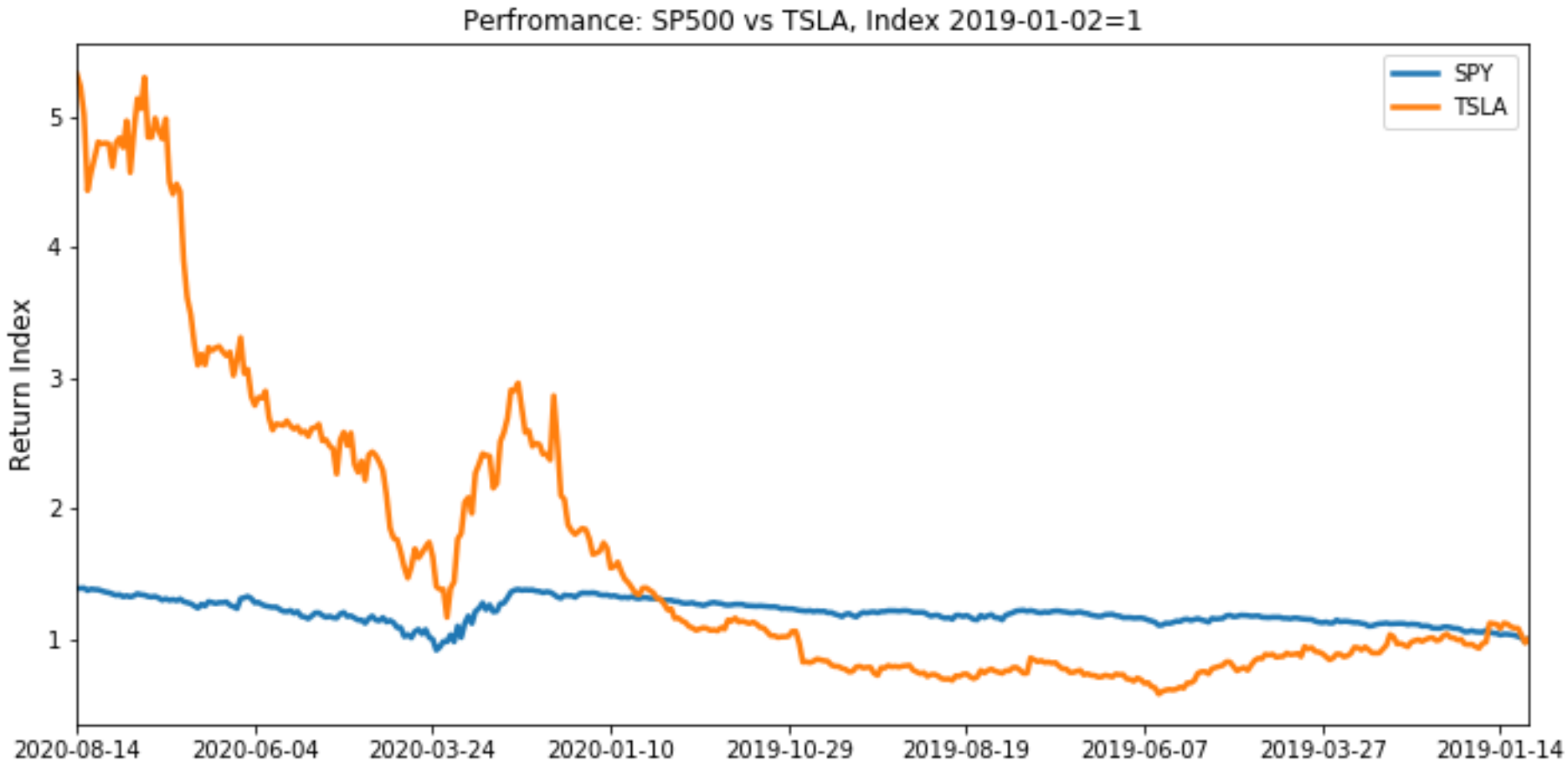


Image credit: Heldiney Pereira

Hierarchy of Visual Attention (continued)

- When more than a few distinct objects are present we view them as a single whole
- Spatially grouped objects are easier to distinguish: 10000000 vs 10,000,000
- We are pattern seekers, and our brain starts concocting a story even before we have a chance to process the whole picture
- Evolutionary baggage: warmer colors are associated with higher values or 'good' -- helpful to discern ripe fruits from the surrounding foliage, preference for symmetry
- Societal conventions: the time flows from left to right, text is read top-to-bottom, and left-to-right

Example: Cultural Conventions



- Let's have a look at the performance of TSLA vs SP500 index, wait…

Primary / Secondary Information

- As we try to make sense out of the slide, our brain has to process the following:
- the concept of a chart
- the concept of an index
- spatial representation of time flow
- **representation of dates**
- ... a million other things...
- the fact that Tesla has outperformed S&P500

Primary / Secondary Information

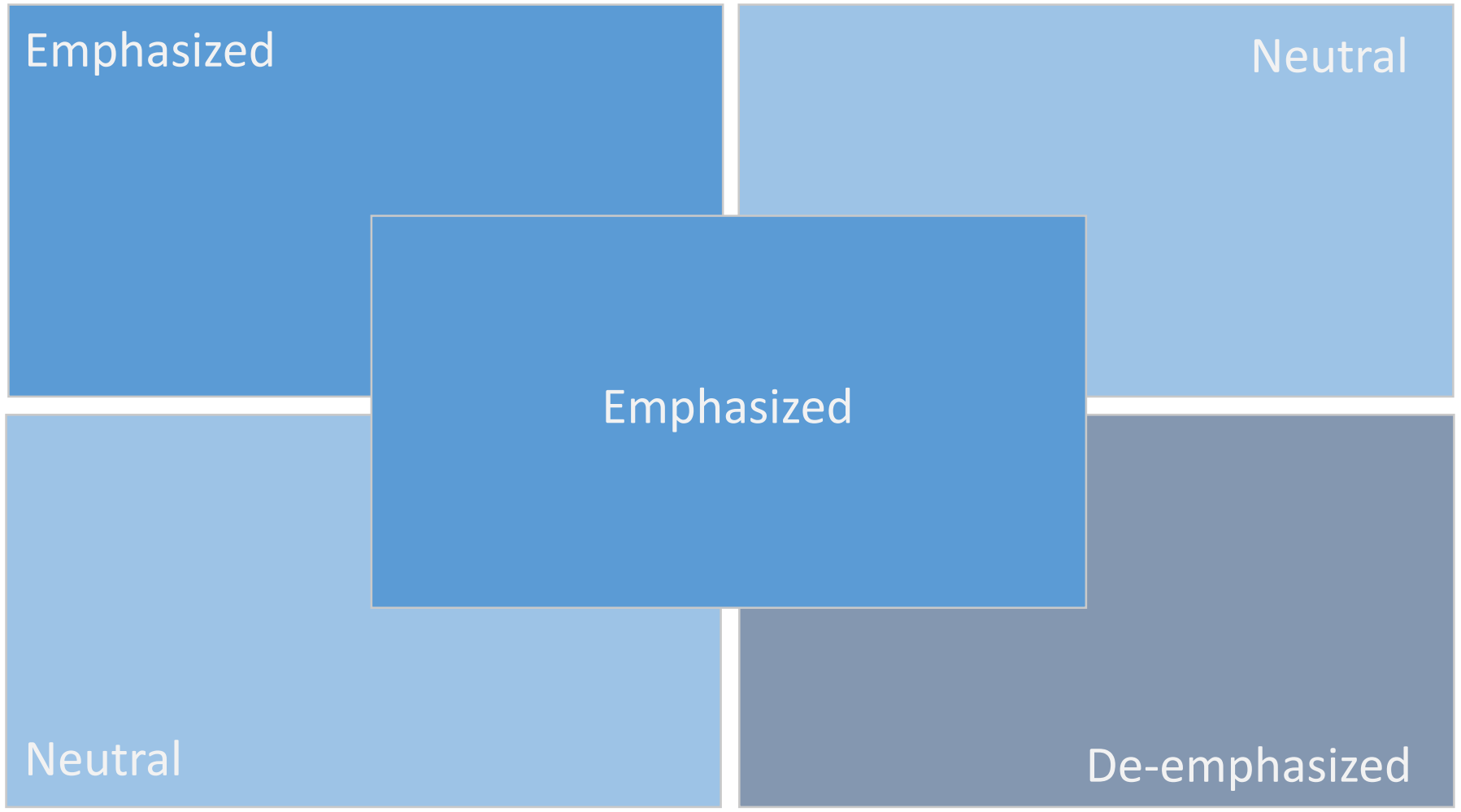
- The majority of us is accustomed to the time flowing from left to right
- But this is not a unique time convention – the Ayamara people refer to the past as being in *front* of them and the future as being *behind* them
- Although rather artificial, the reverse time example highlights several things to keep in mind when visualizing research findings:
    - there is what can heuristically be called **primary** and **secondary** information on every slide
    - the brain needs to process both while its resources are limited
- It is the presenter's task to be informed of the wiring of the audience's brains to foster concentration on the primary information. Which would maybe imply drawing charts differently for Aymara speakers.

Example: Color

	AAPL	CVX	DAL	GM	GS	JPM	LUV	MSFT	TSLA	XOM
daily	1.06 (0.02)	1.16 (0.02)	1.27 (0.04)	1.25 (0.03)	1.34 (0.02)	1.32 (0.02)	1.07 (0.03)	1.11 (0.02)	1.27 (0.06)	1.00 (0.02)
monthly	1.09 (0.16)	1.20 (0.11)	1.16 (0.22)	1.55 (0.16)	1.62 (0.14)	1.50 (0.12)	1.10 (0.19)	0.86 (0.12)	1.01 (0.38)	1.19 (0.10)
annually	0.87 (0.57)	1.33 (0.44)	-0.38 (1.22)	1.77 (0.52)	1.22 (0.61)	1.35 (0.43)	0.26 (1.01)	1.13 (0.34)	-0.30 (2.18)	0.74 (0.50)

- Our attention is immediately drawn to high-contrast cells, that contain more extreme values
- We intuitively expect the cells with colder hues to contain lower values

Example: Visual Layouts



Example: Spatial Grouping and Symmetry

Table 1: Descriptive Statistics: Monthly Returns Jan 2019 – Jul 2020

		Tech		Banks		Auto	
Asset	SPY	AAPL	MSFT	JPM	GS	TSLA	GM
Panel A: Descriptive statistics							
Mean	1.73	5.80	4.00	0.63	1.60	10.13	-0.60
Median	2.21	7.64	3.12	3.74	2.97	6.76	-0.85
Std. dev.	5.78	8.40	5.14	8.71	10.40	22.62	11.44
Skewness	-0.81	-1.06	0.01	-1.02	-0.58	0.44	-0.68
Kurtosis	1.36	0.51	-0.46	1.55	0.79	-0.55	1.73
Panel B: Correlation matrix							
SPY	1.0	0.83	0.71	0.85	0.94	0.50	0.85
AAPL	0.83	1.0	0.71	0.68	0.73	0.64	0.61
MSFT	0.71	0.71	1.0	0.56	0.67	0.51	0.47
JPM	0.85	0.68	0.56	1.0	0.84	0.30	0.8
GS	0.94	0.73	0.67	0.84	1.0	0.46	0.89
TSLA	0.50	0.64	0.51	0.30	0.46	1.0	0.28
GM	0.85	0.61	0.47	0.80	0.89	0.28	1.0

Example: Spatial Grouping and Symmetry

Table 2: Descriptive Statistics: Monthly Returns Jan 2019 – Jul 2020

		Tech		Banks		Auto	
Asset	SPY	AAPL	MSFT	JPM	GS	TSLA	GM
Panel A: Descriptive statistics							
Mean	1.73	5.80	4.00	0.63	1.60	10.13	−0.60
Median	2.21	7.64	3.12	3.74	2.97	6.76	−0.85
Std. dev.	5.78	8.40	5.14	8.71	10.40	22.62	11.44
Skewness	−0.81	−1.06	0.01	−1.02	−0.58	0.44	−0.68
Kurtosis	1.36	0.51	−0.46	1.55	0.79	−0.55	1.73
Panel B: Correlation matrix							
SPY							
AAPL	0.83						
MSFT	0.71	0.71					
JPM	0.85	0.68	0.56				
GS	0.94	0.73	0.67	0.84			
TSLA	0.50	0.64	0.51	0.30	0.46		
GM	0.85	0.61	0.47	0.80	0.89	0.28	

- Type II errors (false negatives) are evolutionary costly
- Thus it is advantageous to see something even if there is nothing
- In fact we have limited control over what we are looking at

Why visualization is important

- Humans receive disproportionally large share of information from vision in comparison to other senses
- Visualization is information/knowledge compression, we are bad at comprehending large collections of objects like raw numbers
- Other compression methods like descriptive statistics can miss important features of the dataset at hand, especially if the data is not iid.
- It really helps you to tell *your* story, control the narrative and emphasize what *you* think is important

Example: Sales Data

	Region A		Region B		Region C		Region D	
Store ID	Revenue	Profit	Revenue	Profit	Revenue	Profit	Revenue	Profit
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
2	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

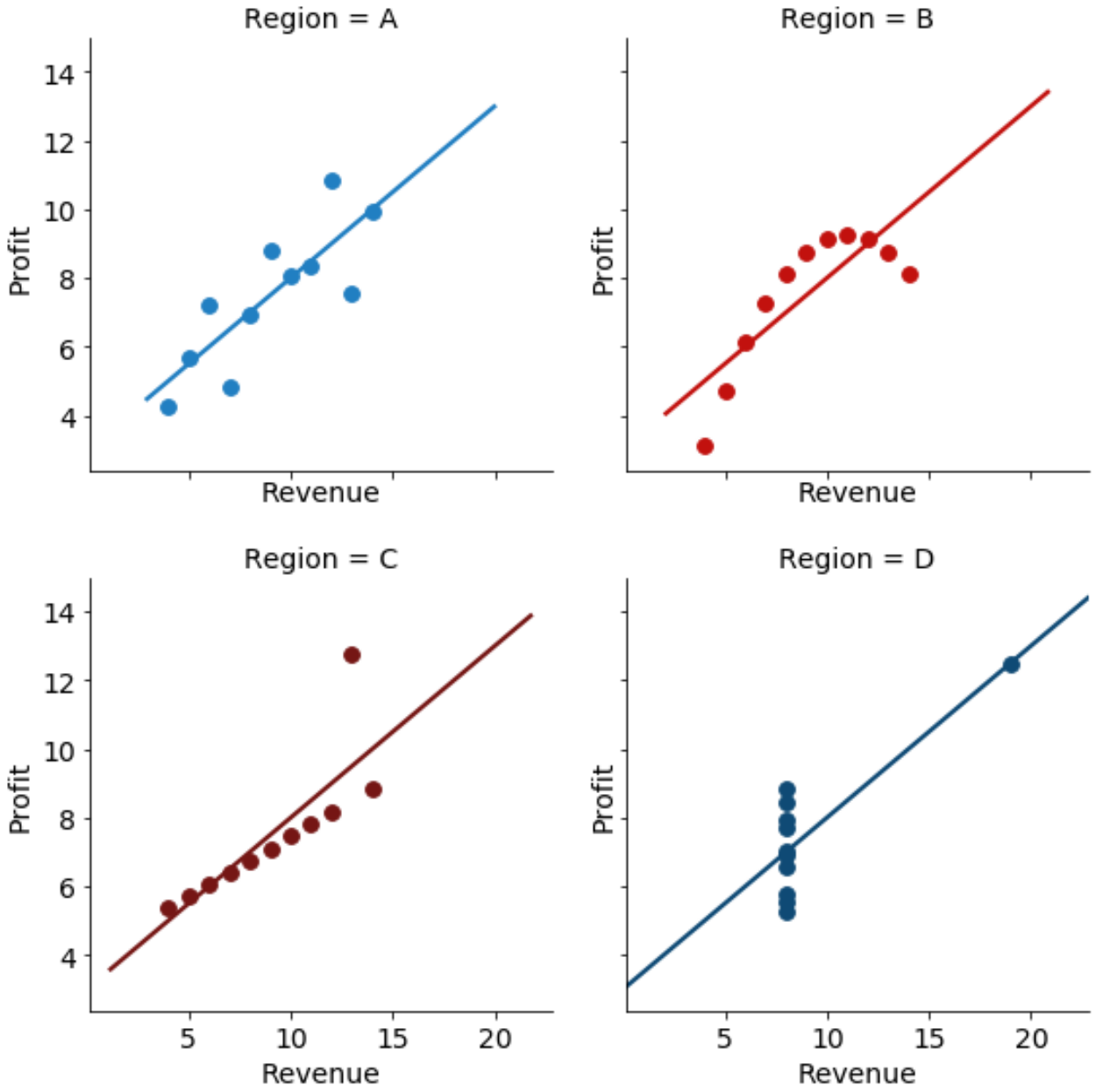
- Assume you are asked to evaluate profitability over four regions with 11 stores in each

Example: Sales Data

	Region A		Region B		Region C		Region D	
Store ID	Revenue	Profit	Revenue	Profit	Revenue	Profit	Revenue	Profit
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Variance	11	4.125	11	4.125	11	4.125	11	4.125
Regression Line	P = 3 + 0.5R		P = 3 + 0.5R		P = 3 + 0.5R		P = 3 + 0.5R	

- Well, that wasn't helpful

Example: Sales Data



“The goal of a good data visualization is to help your reader in transition from 'WFT?' to 'Aha!'”

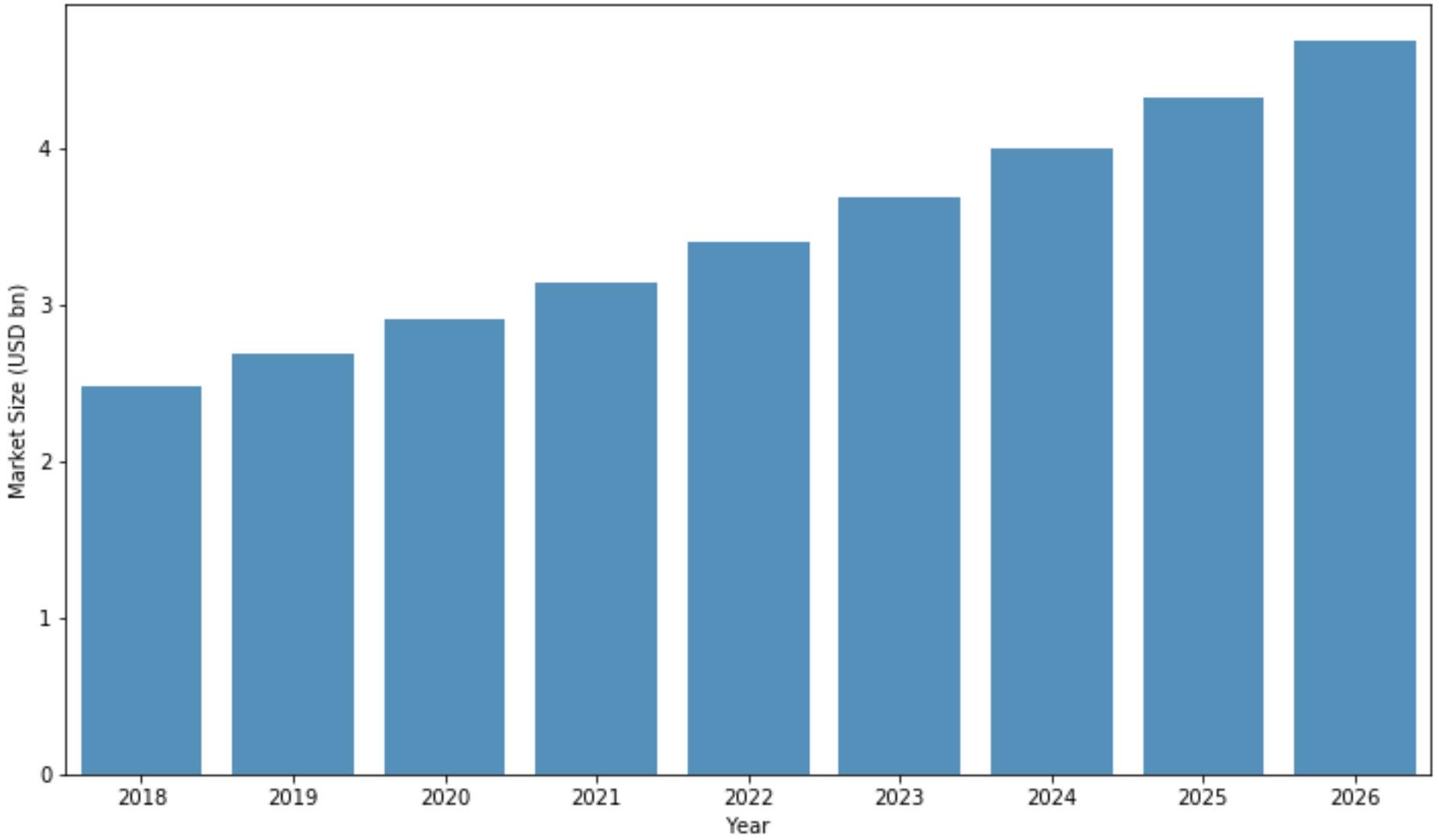
Karl Marx

- Let’s start from ‘wtf’

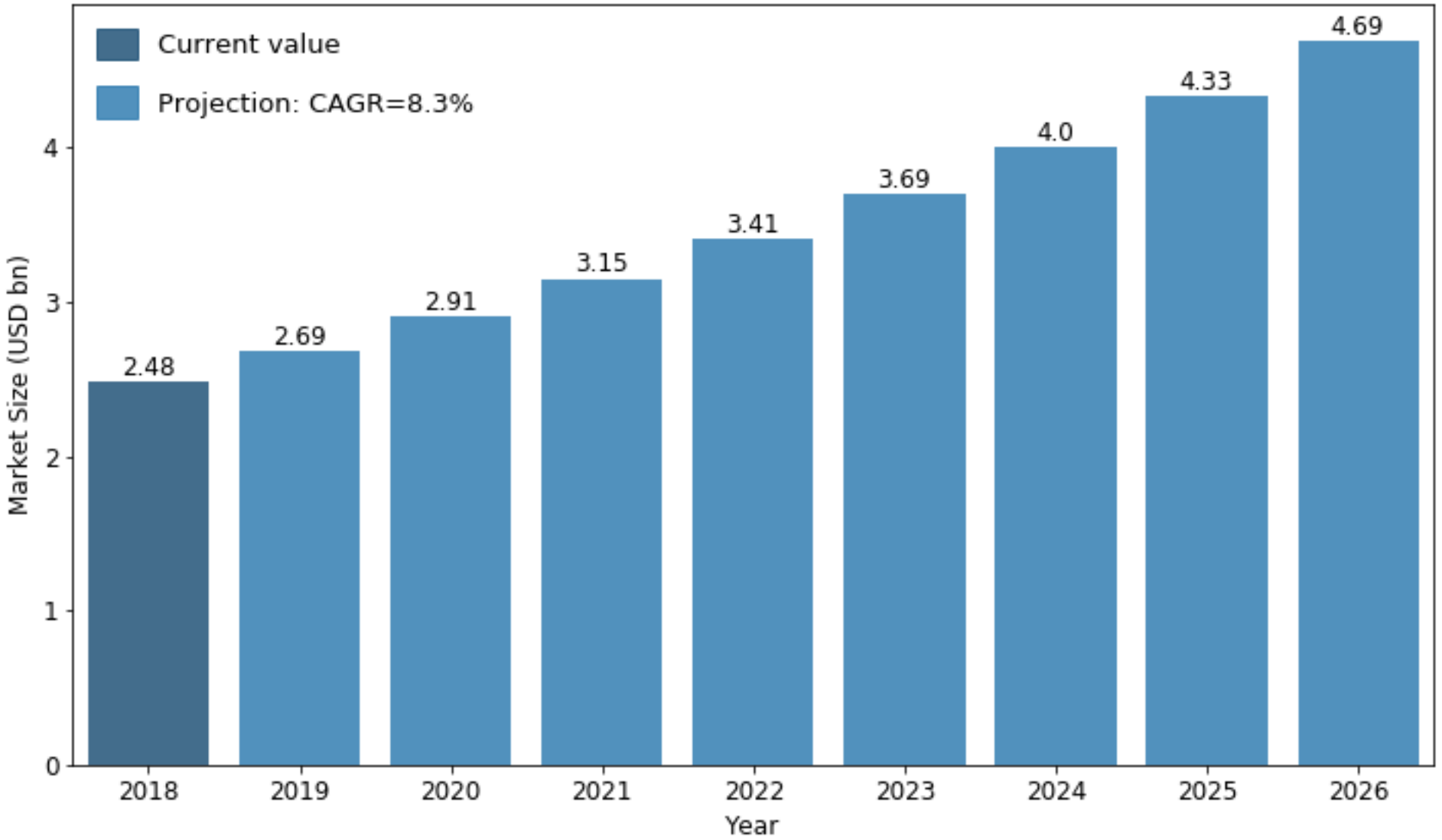
Abuse of Visualization: Pie Charts

- Look at this utter nonsense (viz.wtf has plenty of those)

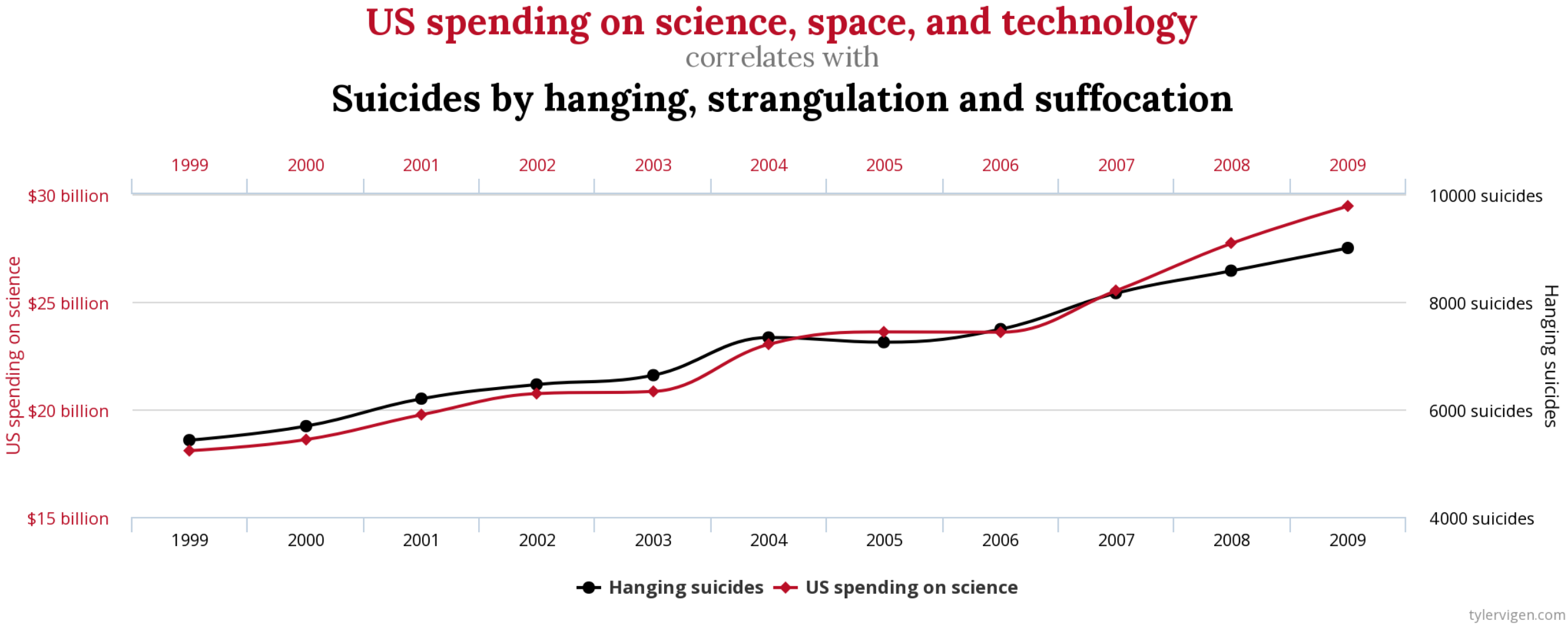
Abuse of Visualization: Bar Charts > Pie Charts



Abuse of Visualization: Bar Charts > Pie Charts



Abuse of Visualization: Axes of Evil

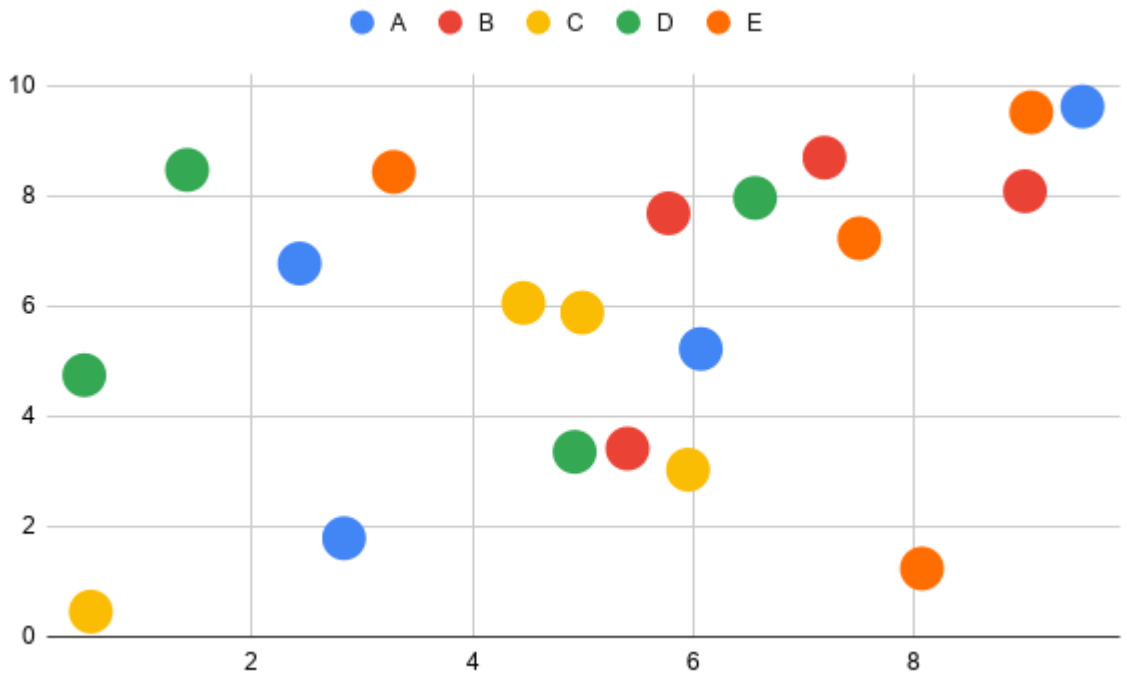


- An entire gallery
- There are books on the topic

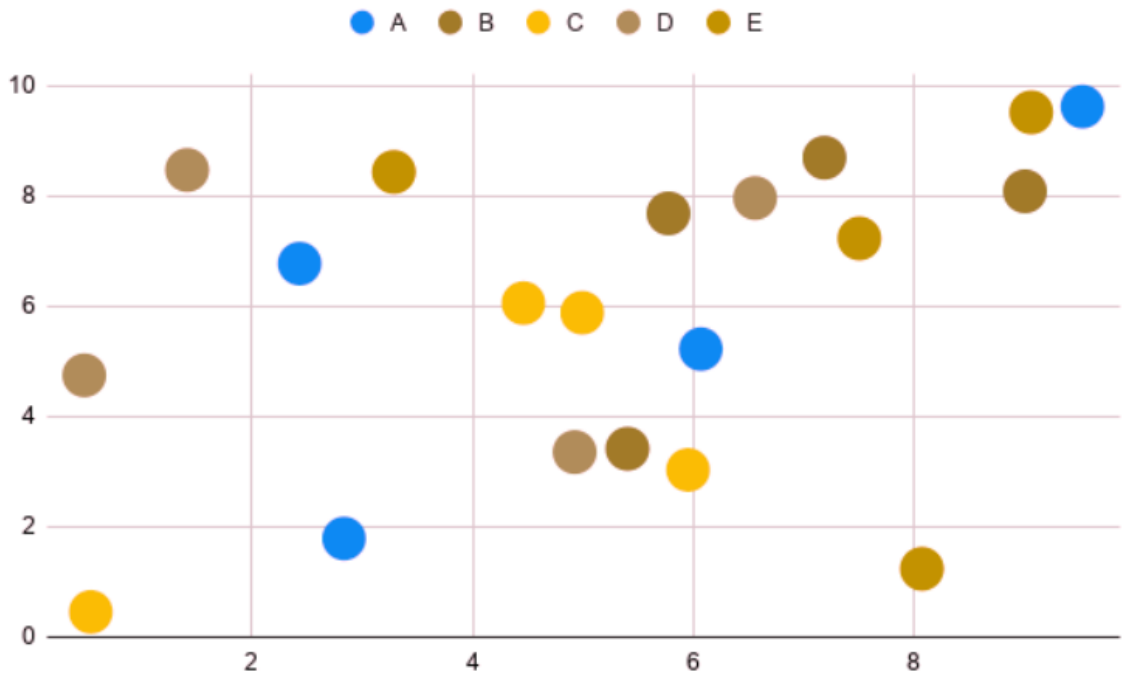
Color: Palettes and Color Harmonies

- Color psychology is a good starting point
- There are several tools to build your own palettes which follow the color harmony rules, and check if the images are color-blind-friendly:
  - https://color.adobe.com/create/color-wheel
  - https://paletton.com/
- Color blindness simulator allows to display images as seen by people with different types of color vision deficiency.
- See this article on creating color blind-friendly visualizations. tl;dr version – avoid combinations of red and green

Color Blindness



Original

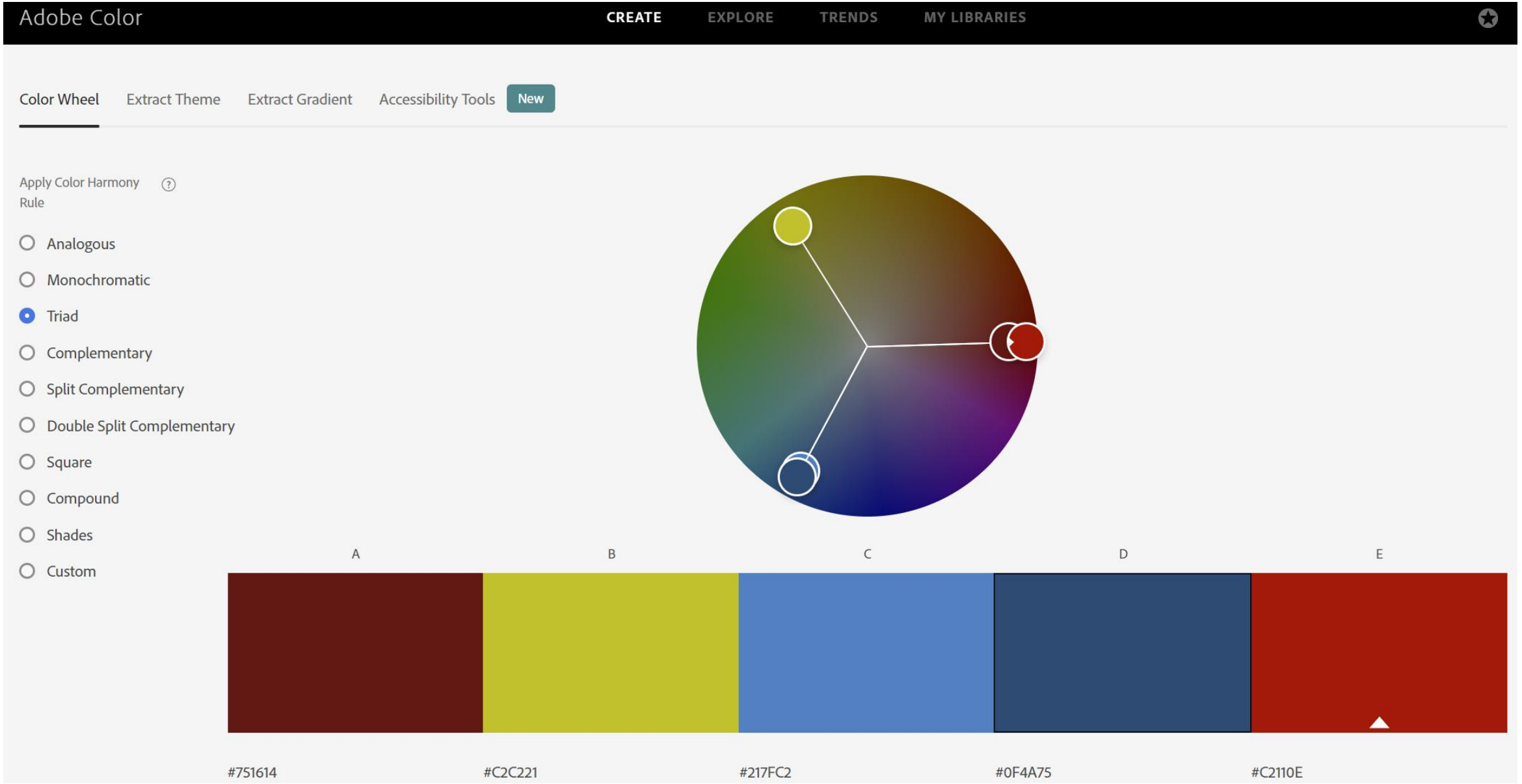


Green-blind

- The most common types of color blindness affect ability to discriminate red and green colors

Image credit: Chaulio Ferreira

Color Wheel: Custom Palettes



Notes and Hints

- Currently we are bombarded with data and visuals. Use 'active viewing' approach -- when you see, for instance, a chart in a scientific or in media, pause for a moment and ponder: 'what is good or bad about this chart?' or 'what would I do differently?'.
- A similar technique can be applied to reading and writing.
- Maintain a gallery of the charts you make: some rare operations are quite forgettable, so you will save time by referring to your own work instead of sifting through stackoverflow posts.
- The less is better: if you can avoid cluttering a single chart and do several charts instead, do it!

Links

- David McCandless on data visualization
- National Geographic article on visualization
- A short video on chart manipulation