Lessons from

HOW TO LIE WITH STATISTICS

Book by Darrell Huff

7 January 2021

Muhammad Idrus Fachruddin
E-mail : muhammad.i.fachruddin [at] gdplabs.id

Timeless data literacy advice.

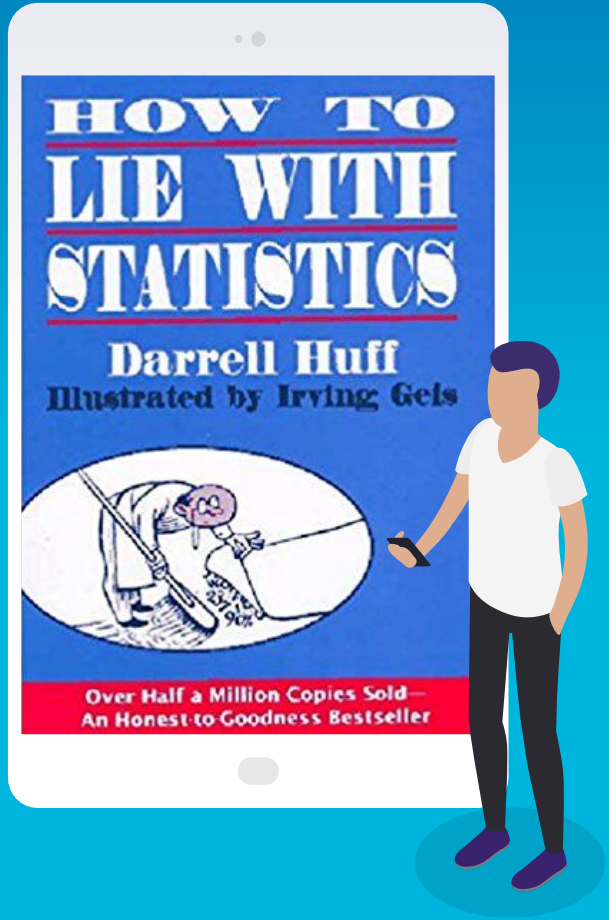
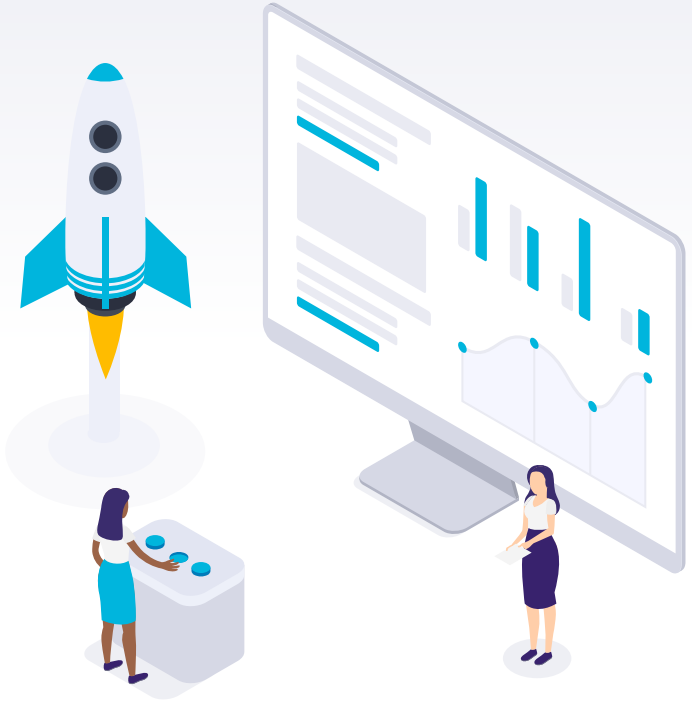


Table of contents :

1 The Sample with Built-In Bias



The Sample with Built-In Bias

Response Error

Caused by error when gathering response from respondent/observation

Sampling Procedure

Under-represented sample / bad sampling design

Example

The Literary Digest poll on 1932 election

The Sample with Built-In Bias

Response errors :

a. Inaccurate memory of respondent
   Most people don't know their own annual income to the nearest dollar
b. Exaggeration or minimization
   Some exaggerate their income out of vanity or minimize it out of fear of the taxman, tooth brushing frequency
c. Giving an answer expected to be pleasing to the one asking
   Japanese vs. Nazis
d. Non-response
   Many will choose not to respond to a questionnaire perceived to be personal. Many of those that do not respond to an income question will be those whose incomes are low

The Sample with Built-In Bias

Sampling procedure : Under-representation can come from several sources.

- Less likely to be able to easily locate the Yale men who were less successful
- Convenience samples. “All kinds of people can be found in a [railroad] station.
- Selecting subjects you are more comfortable talking to
- Difficulties in designing or collecting a stratified sample. Expensive and need extra effort.

The Sample with Built-In Bias

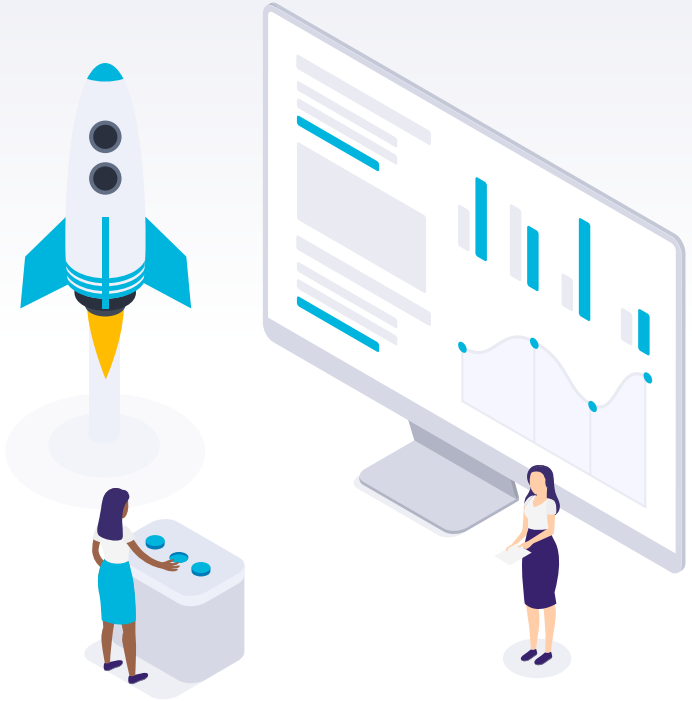
The Literary Digest : The source of the bias (or even its existence) may not be readily apparent

- As it had done in 1920, 1924, 1928 and 1932, *The Literary Digest* conducted a straw poll regarding the likely outcome of the 1936 presidential election. Before 1936, it had always correctly predicted the winner. It predicted Landon would beat Roosevelt.
- In November, Landon carried only Vermont and Maine; President F. D. Roosevelt carried the 46 other states. Roosevelt won significantly.
- The polling techniques used were to blame, even though they polled 10 million people and got a response from 2.4 million.They polled mostly their readers, who had more money than the typical American during the Great Depression. Higher income people were more likely to vote Republican.

A much lower number, such as 1,500 persons, is adequate in most cases so long as they are appropriately chosen.

2

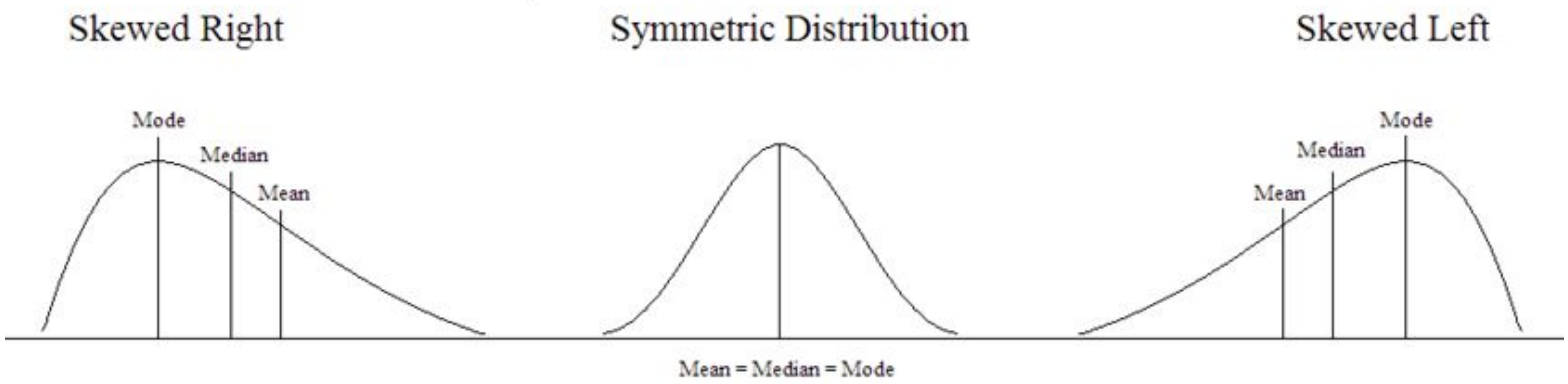
The Well Chosen Average



The Well Chosen Average

Central tendency (oftenly called : average) :

- **Mode**: the most frequent value.
- **Median**: the middle number in an ordered data set.
- **Mean**: the sum of all values divided by the total number of values.



Example

Jeff Bezos Moves to Jakarta

- Average monthly income of Jakartans is Rp. 5.000.000
- One day, Jeff Bezos moves and lives in Jakarta
- Now, average monthly income of Jakartans almost tripled, Rp. 14.000.000

Stock Profit Average:

- Median can be misleading when is used on data with negative value

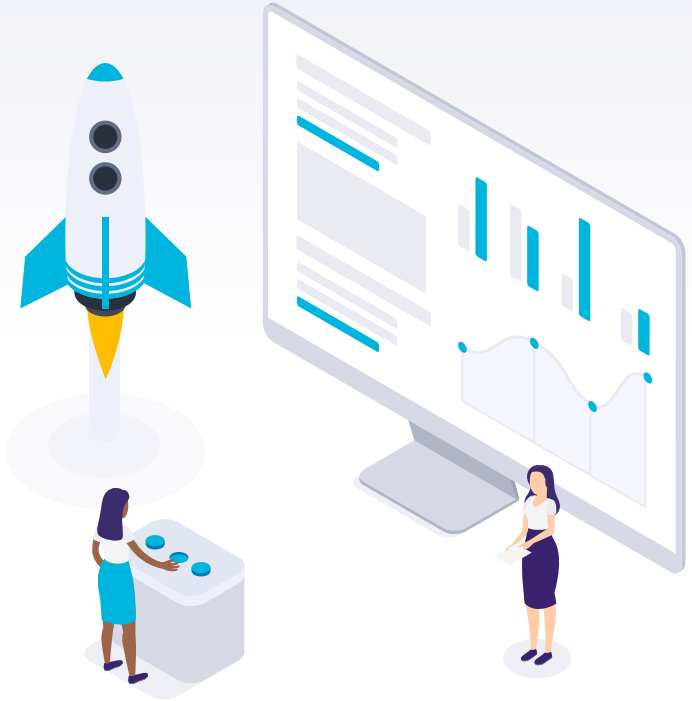
Mean is sensitive to outliers, while median is more robust to outliers and better for skewed data.

simulation

When someone say average, check which average they use.

3

Single Number Is Not Adequate

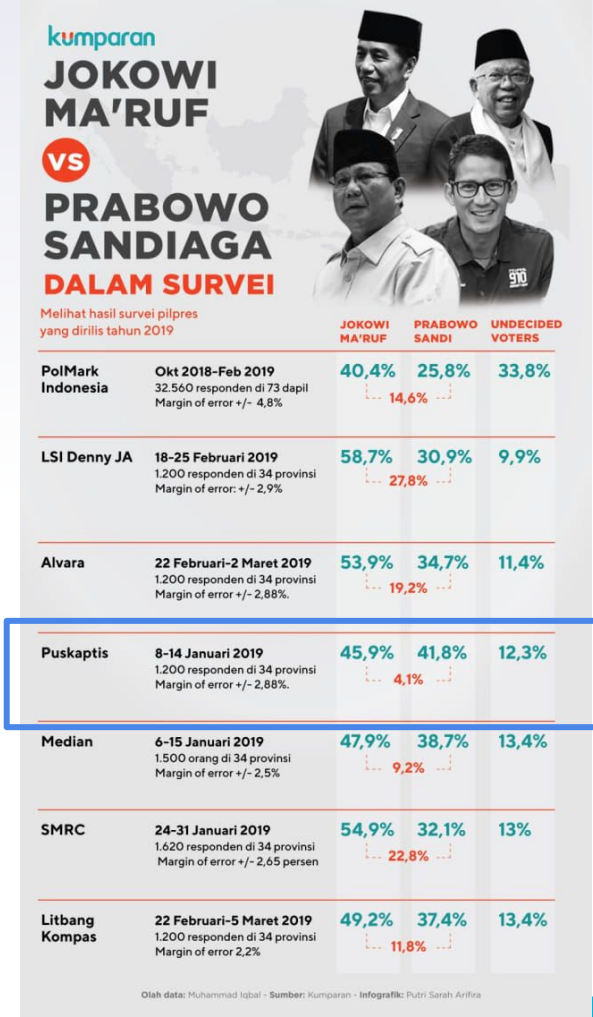


The Importance of Spread/Range/Variation

- A city has average temperature 27 celsius degree for a year. But it has -2 celsius degree in December and 35 celsius degree in June.
- Single statistic/number is not adequate to describe a dataset
- Always look mean with standard deviation or median with interquartile range

Look at Standard of Error of A Statistic Result

- E-books Preferred Over Paper By Men More Than By Women" sounds remarkable until you find out that of the actual polling results found that
  - **52% of men preferred e-books**
  - **49% for women**,
  - **The error** of the survey was **+/-5%**.
- If a survey shows electability two candidates fall within margin of error, no one can claim as the winner.
  Ex :
  - A survey has **MoE 3%**
  - **Candidate A get 49%** +- 3% -> 52% - 46%
  - **Candidate B get 51%** +- 3% -> 48% - 54
  - Candidate B **can not claim** they are leading

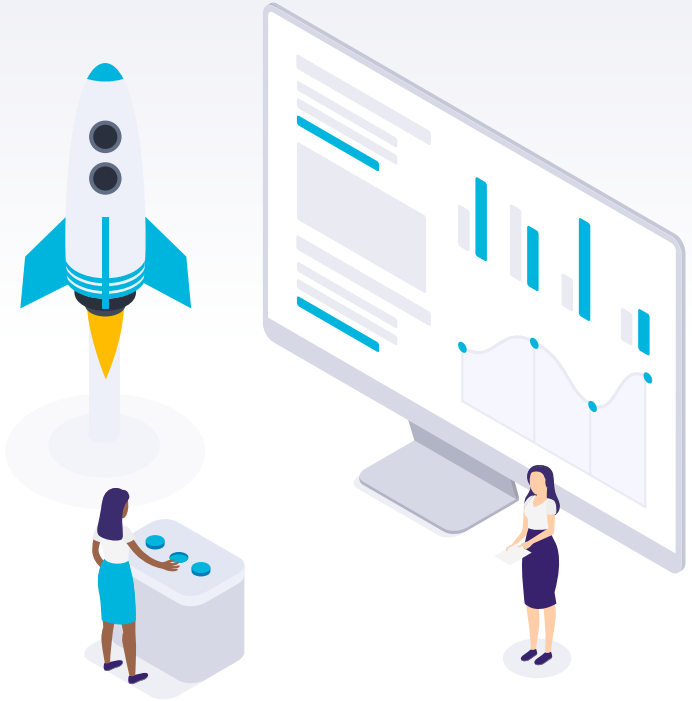


What is missing?

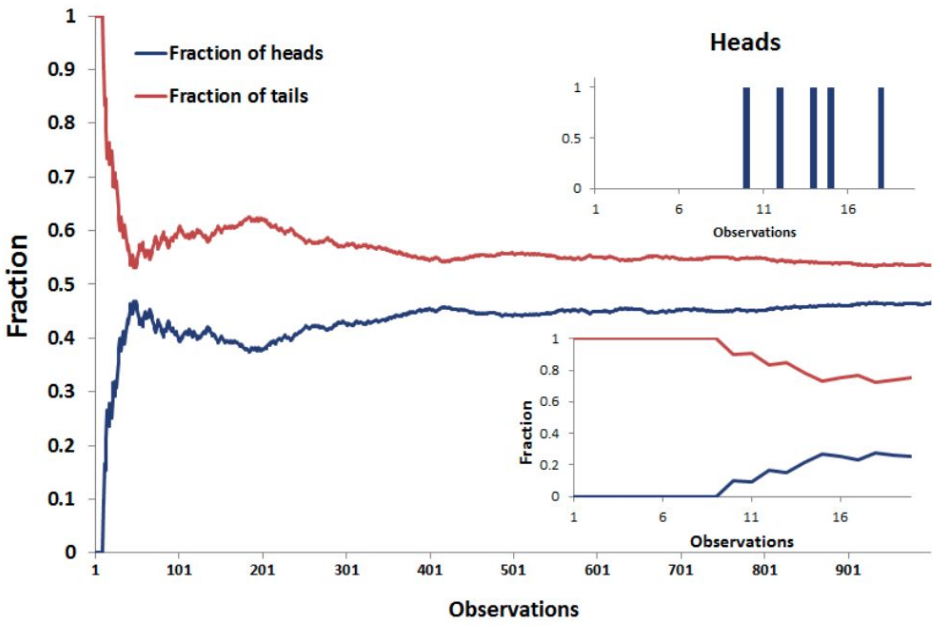
Do we have everything we need to know in order to fully understand the significance of the statistic that is being offered?

4

Small Sample Exploitation



Small Sample Exploitation

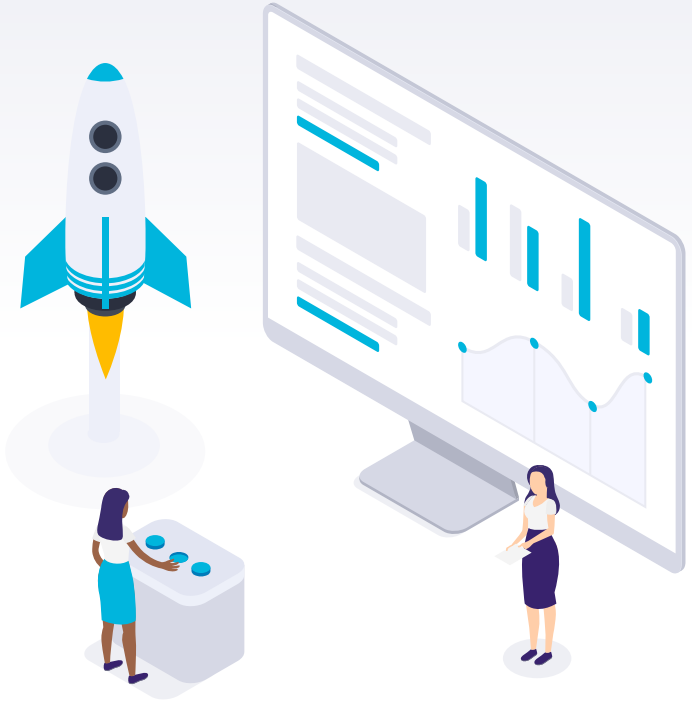


- If we toss a fair coin 10 times, we may get 8 head and 2 tail (80%:20%). But as the number of observation increase (toss 1000 times) the proportion of head & tail appearance will close to 0.5 (50%:50%)
- Small sample oftenly produces extreme value
- Statistics then used to amplify a very small difference between two phenomenon and try to prove one's superiority over other.
- 8/10 dentists choose toothpaste A

source

5

The Semi Attached Figure



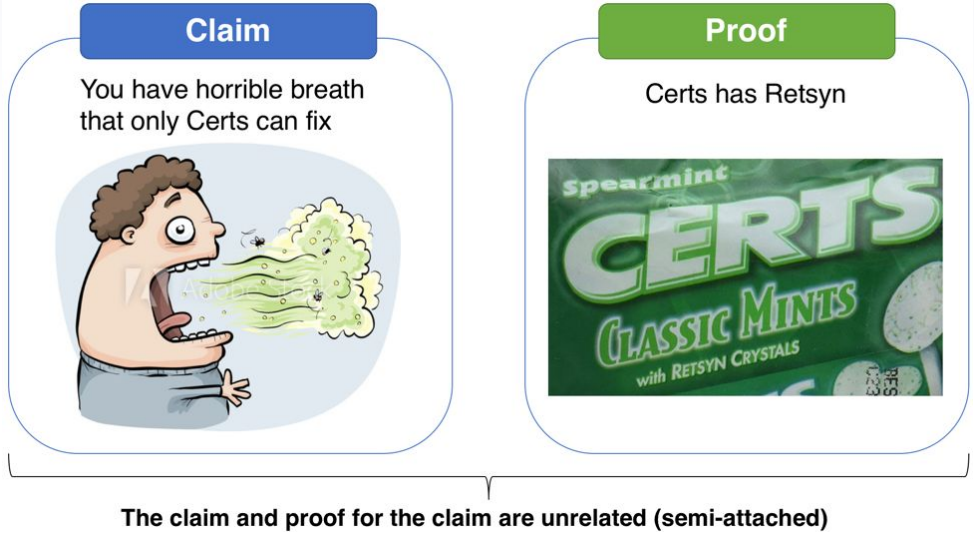
The Semi Attached Figure

A semi-attached figure occurs when proof is given for a claim, but when the reader looks at it closely, the proof and the claim are not related.

Example 1: Now, with Retsyn!

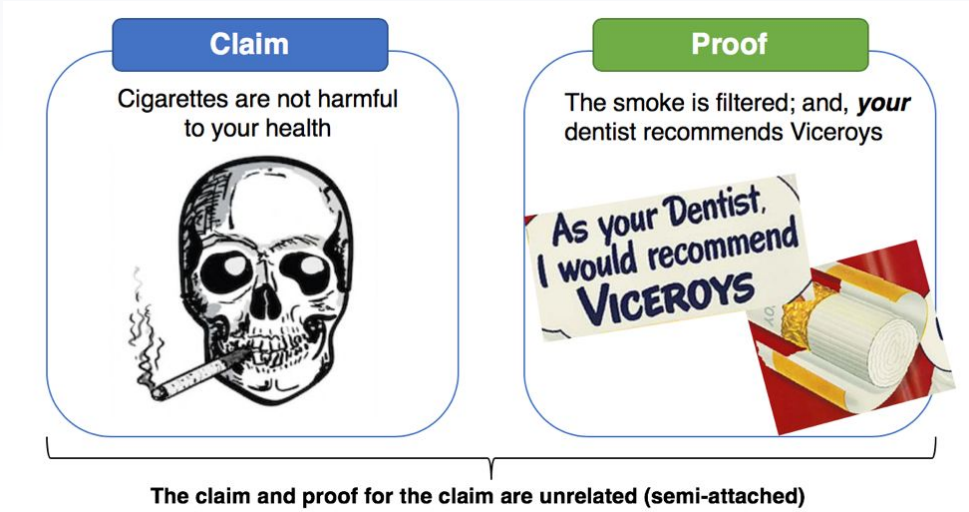
In Certs commercials, the narrator says *"Want fresh, clean breath? Get the only mint with Retsyn,"* or a similar slogan.

What exactly is Retsyn? According to an article by Slate, it's "...natural flavoring, partially hydrogenated cottonseed oil, and a nutritional supplement called copper gluconate, none of which will kill bacteria."



Source : https://www.secjuice.com/the-semi-attached-figure/

Example 2: These cigarettes are fine for your health



Source : https://www.secjuice.com/the-semi-attached-figure/

Example 3 : Indonesia Election and Covid-19 Cases

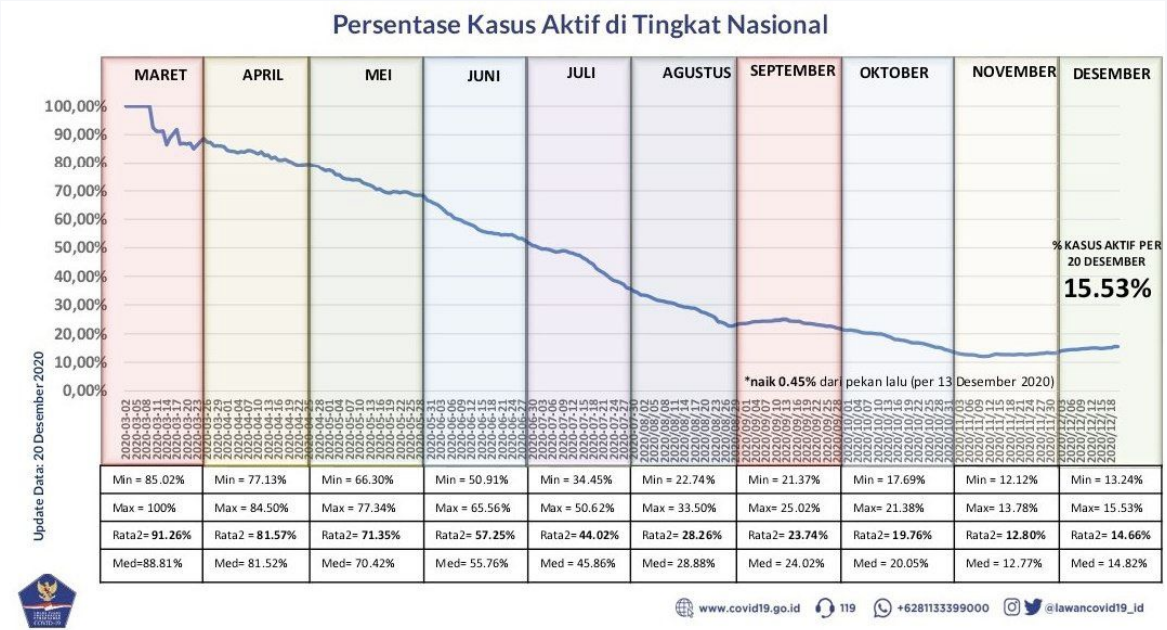
Claim :

Covid cases tend to decrease on region which will held election.

Proof :

- Only use exact number, ignore the testing rate
- Only show the number. No information about relationship (even correlation) but seems like related.

Example 4 : Data presented without context



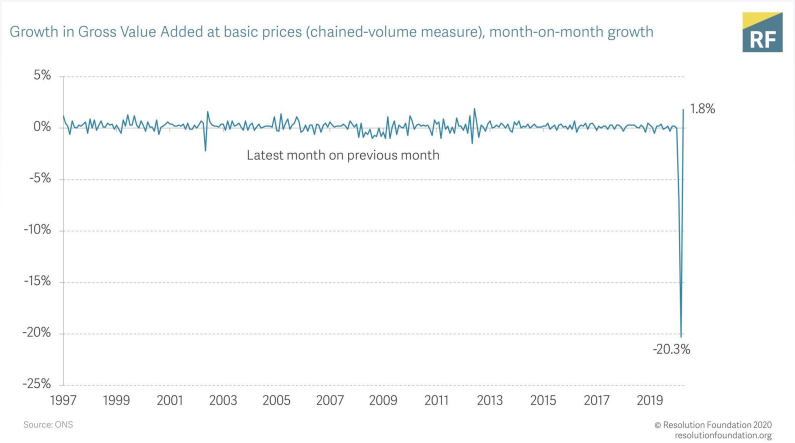
Claim :
% of active cases is decreasing (True).

Fact :

- The % of active case decrease but the base number also increase.
- It's possible we have only 1% active cases with the real number is 1 million cases.

Example 4 : Amplifying V-shape

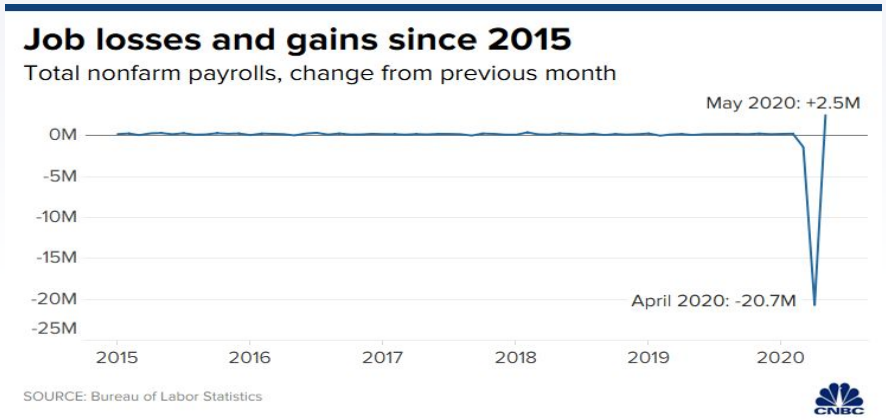
V-shape graph



Actual condition

Choose the metrics or number that can impress others, although it's contrast with actual condition

Example 4 : Amplifying V-shape

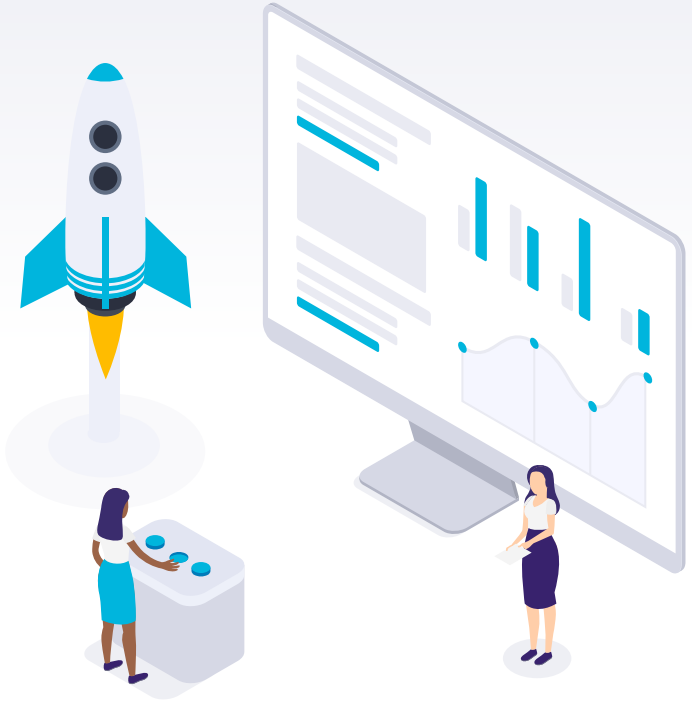


If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.

-Darrell Huff,

6

Graph Manipulation

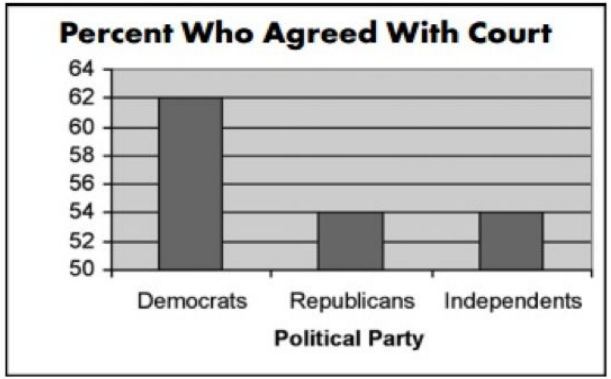


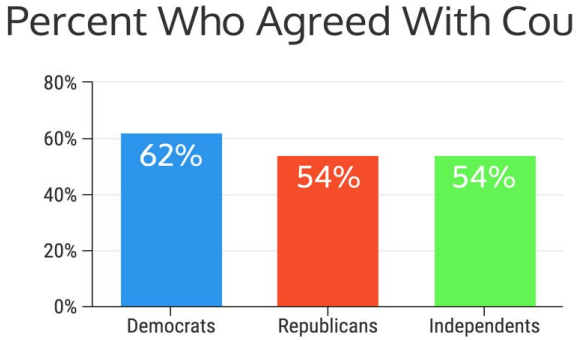
Graph Manipulation

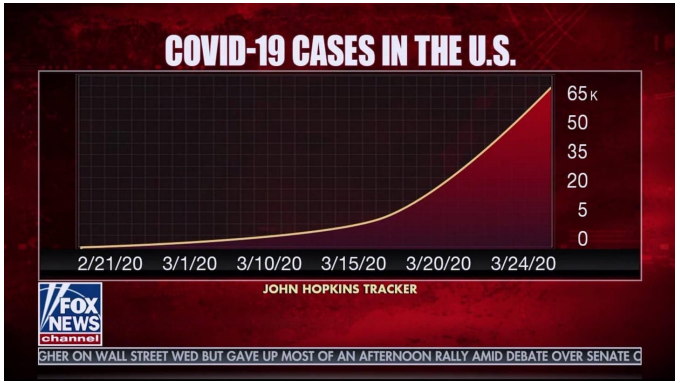
1. Omitting the baseline
2. Manipulating Y-axis
3. Cherry Picking data
4. Using the wrong graph
5. Going against conventions

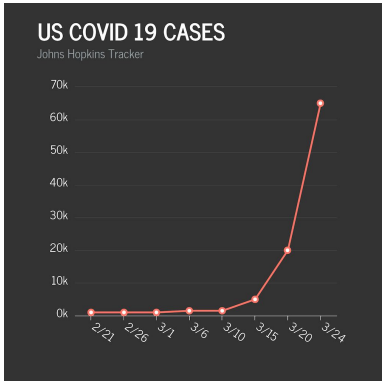
Source : vennage

Omitting the baseline



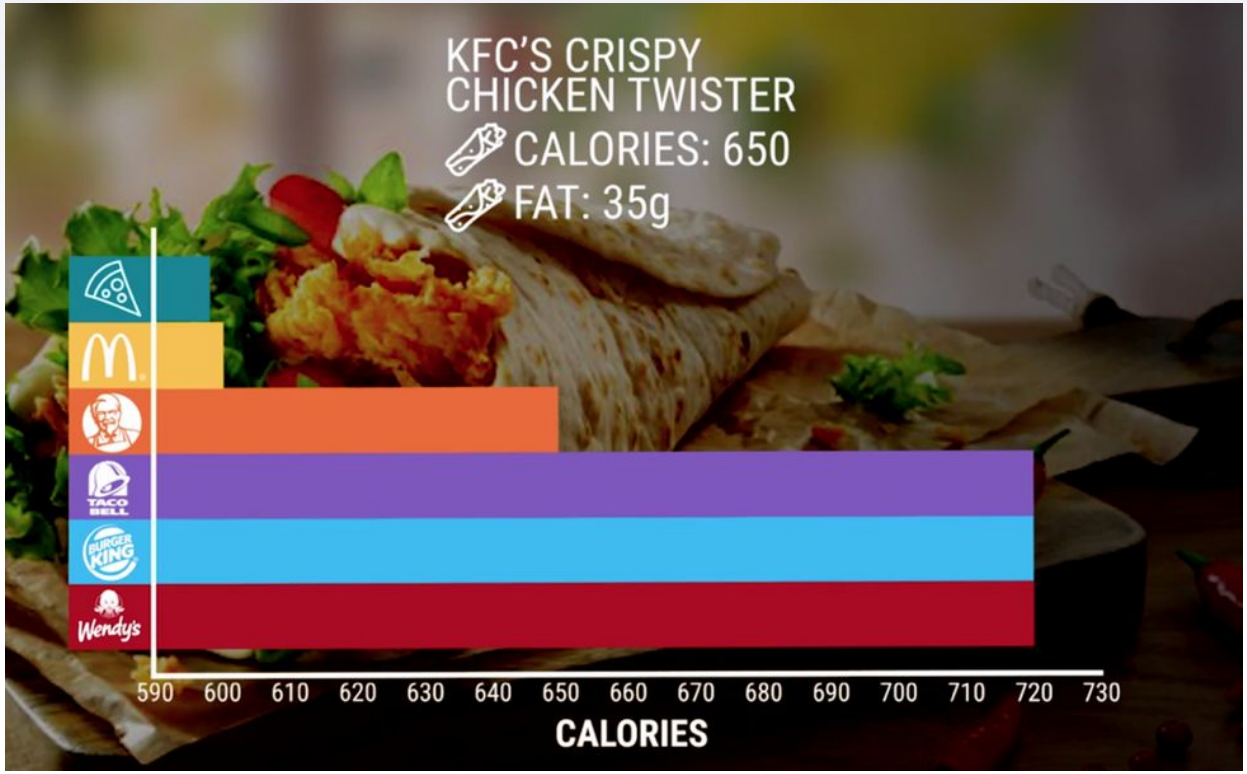






Source : [vennage]

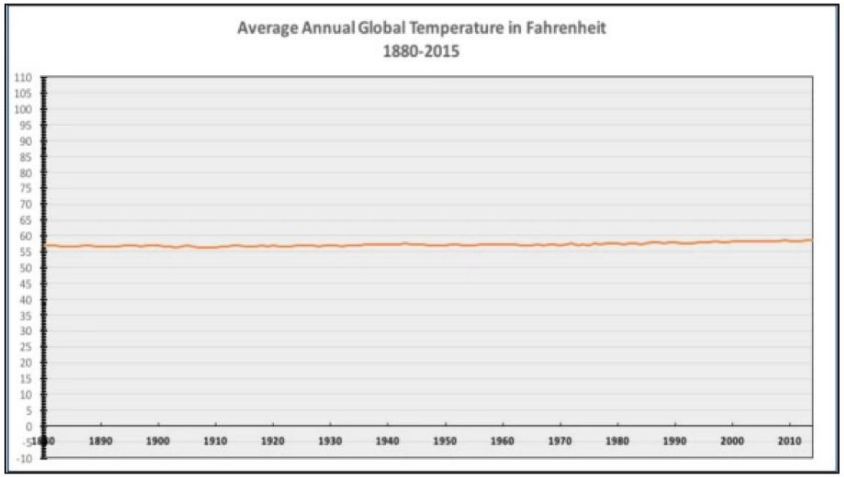
Omitting the baseline

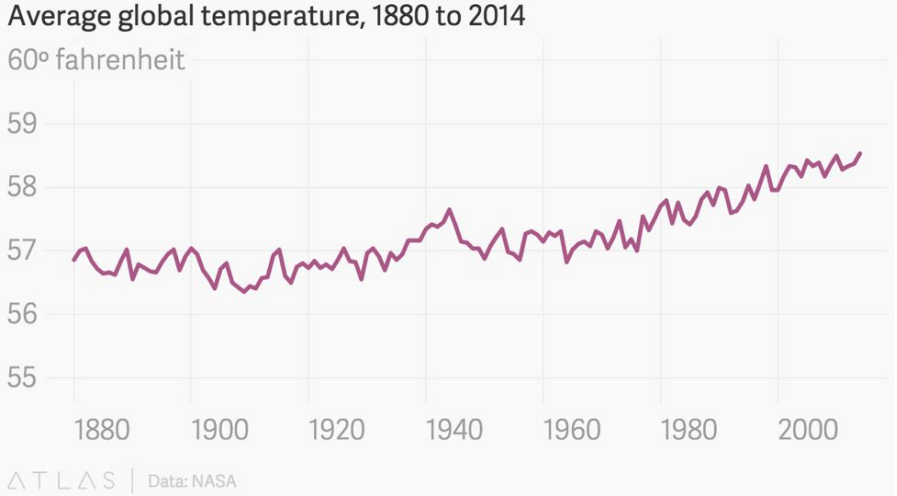


Source : [vennage]

Manipulating Y-axis

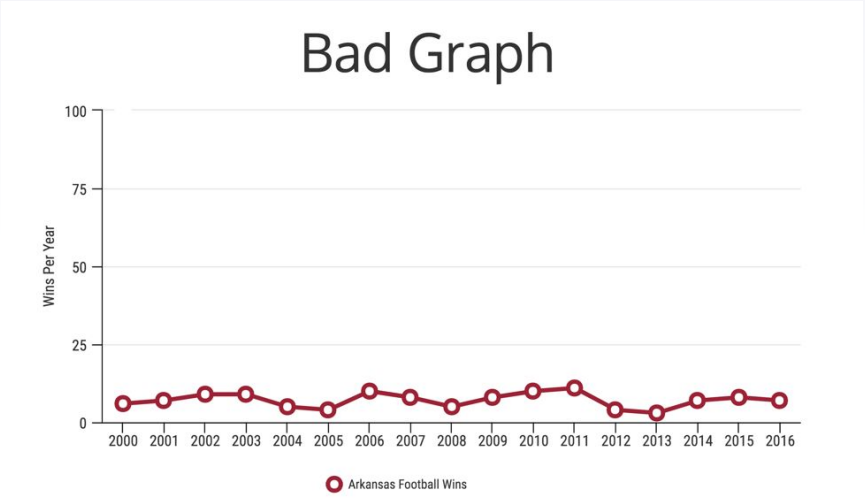
Bad graph corrected with another bad graph

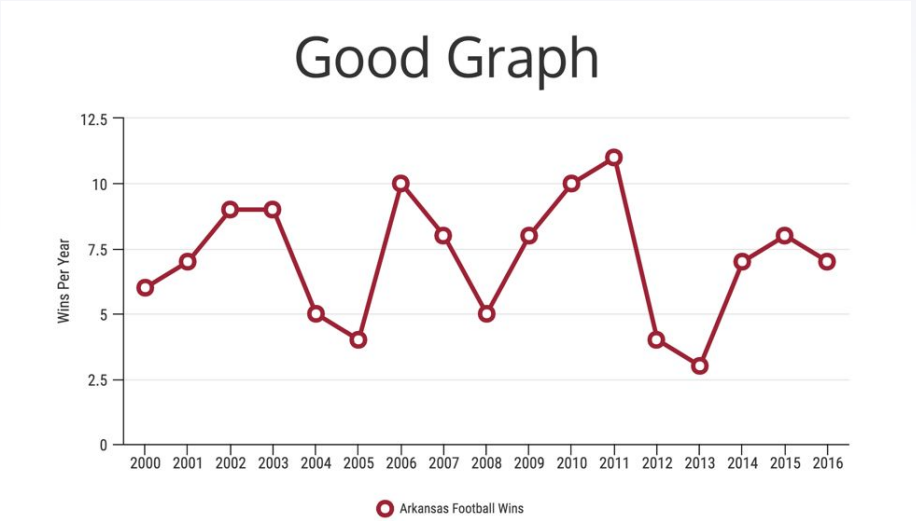




Source : [vennage]

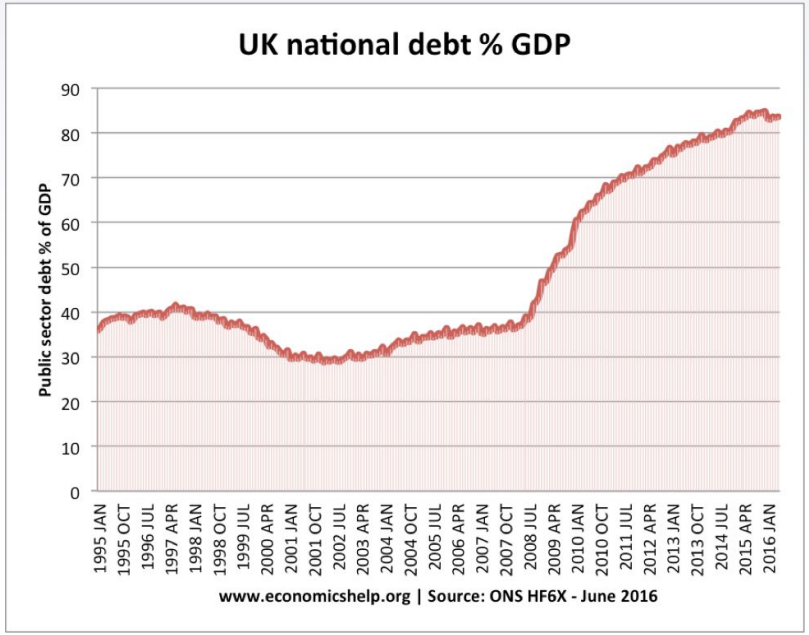
Manipulating Y-axis

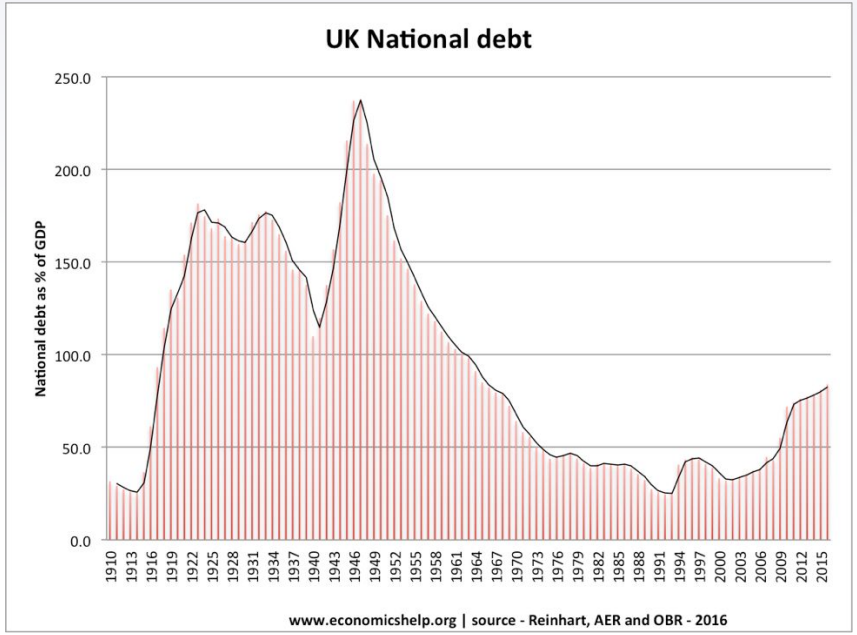




Source : vennage

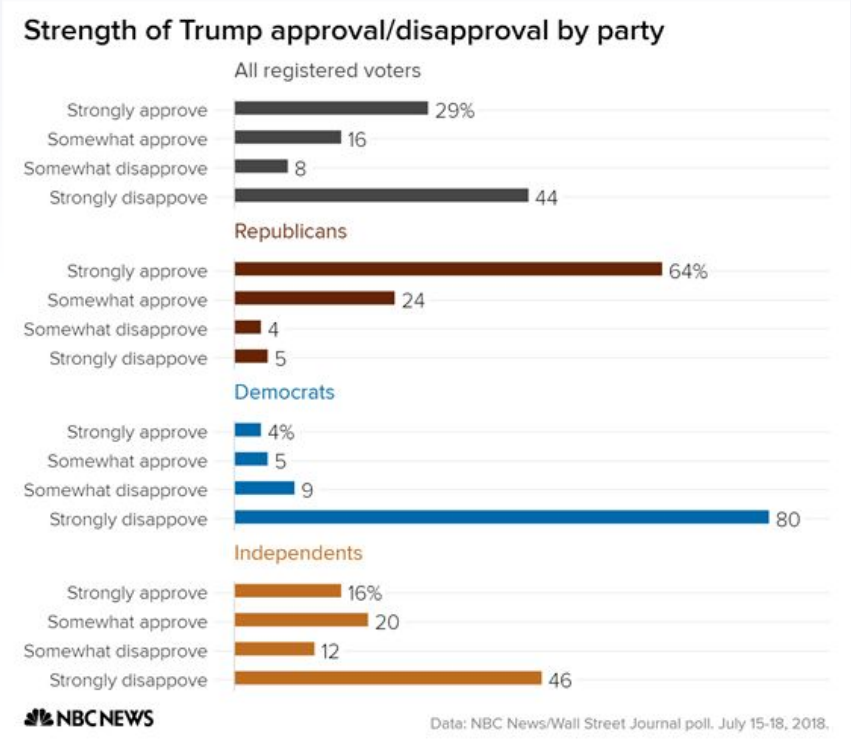
Cherry Picking data

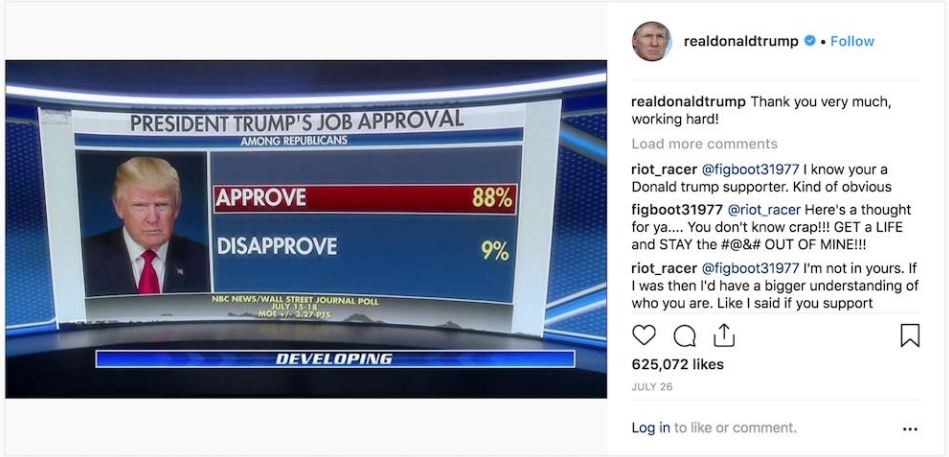




Source : vennage

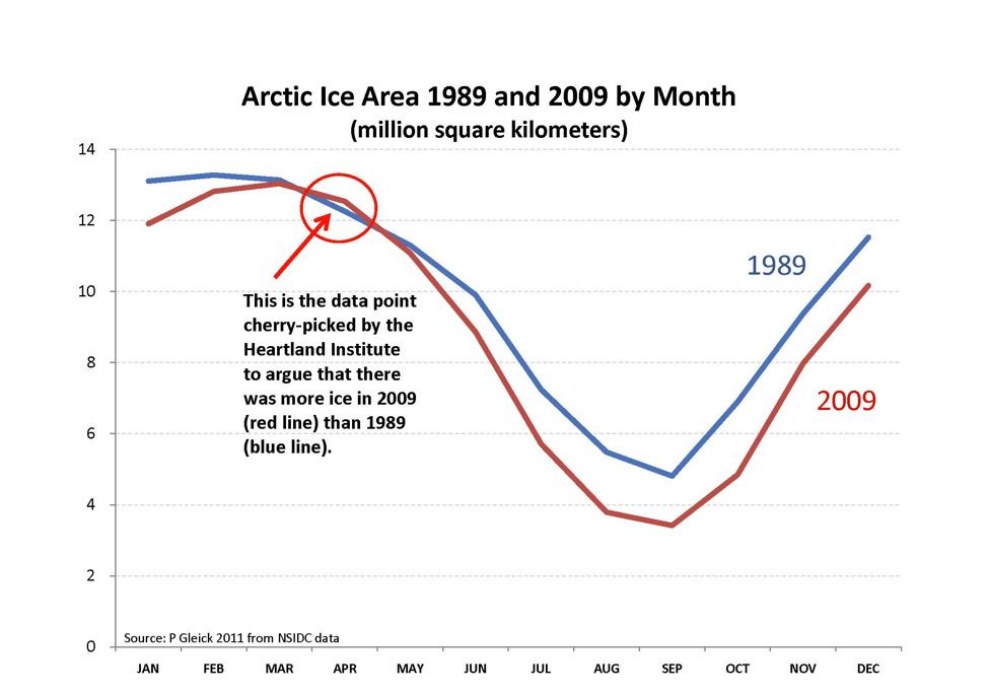
Cherry Picking data





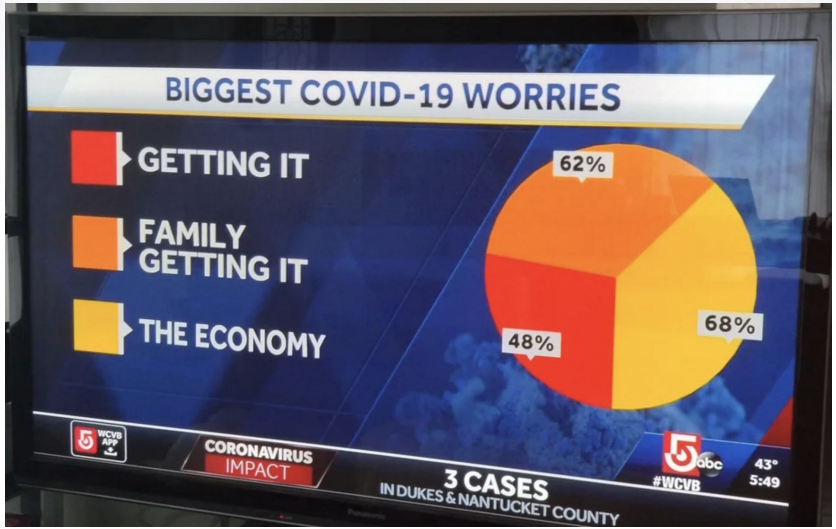
Source : vennage

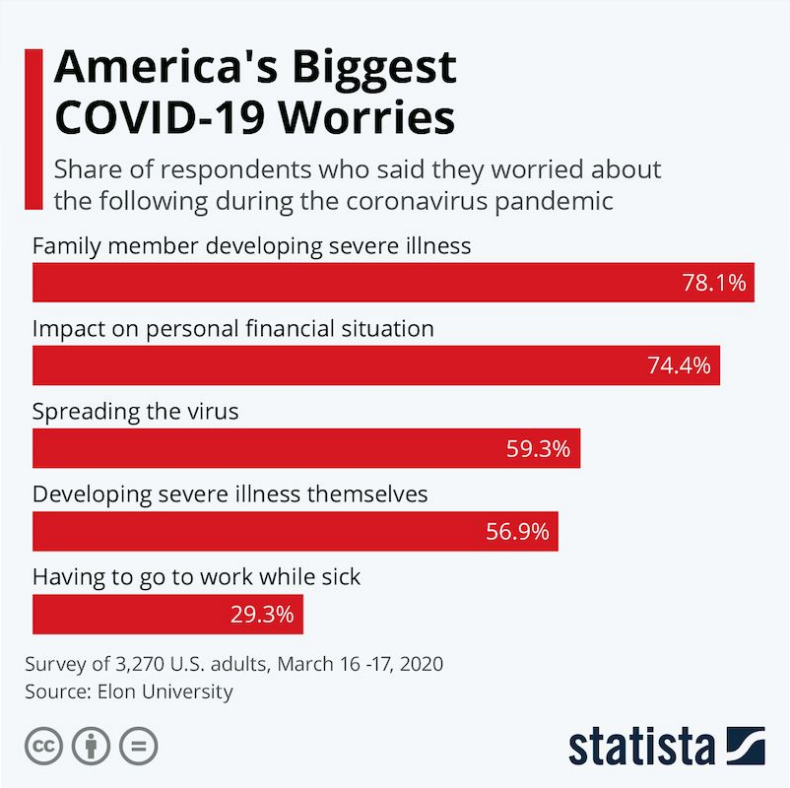
Cherry Picking data



Source : [vennage]

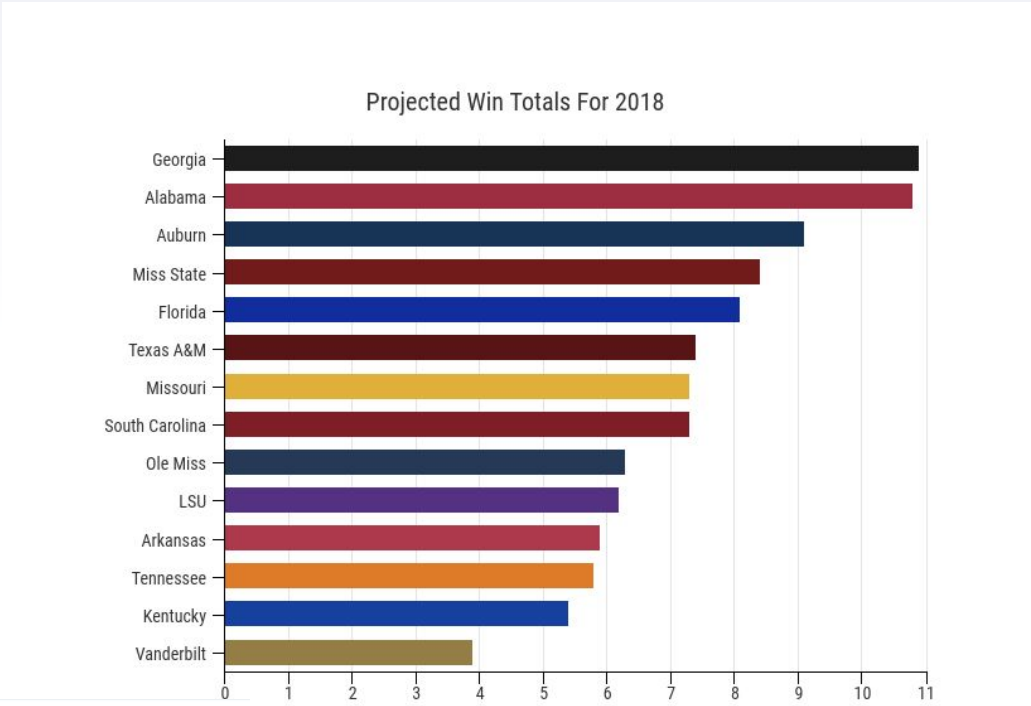
Using the wrong graph





Source : [vennage]

Using the wrong graph

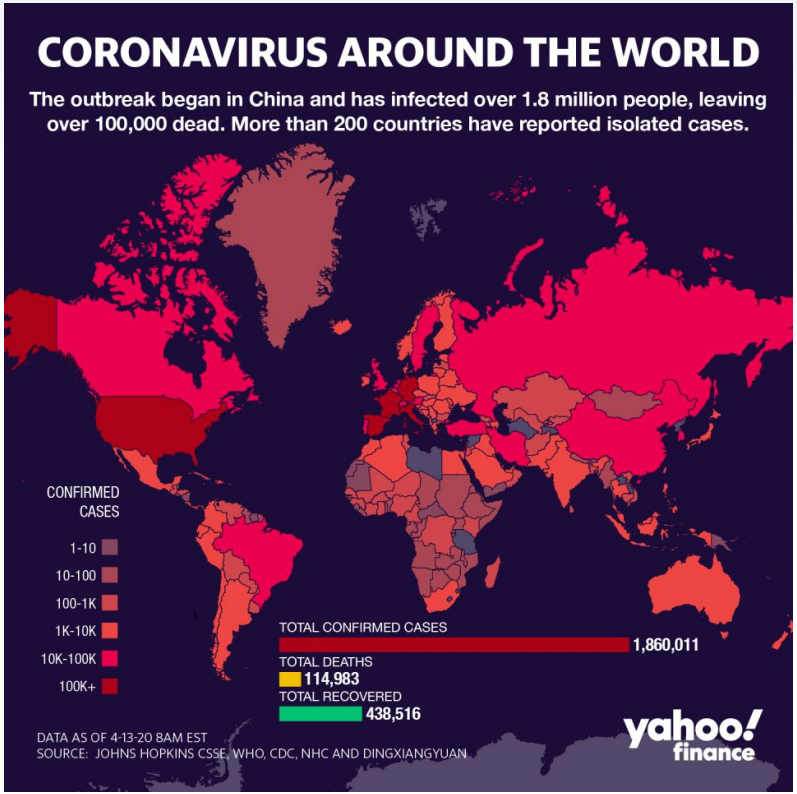


Source : [vennage]

Use the wrong graph

Source

Going Against Conventions



Color palette

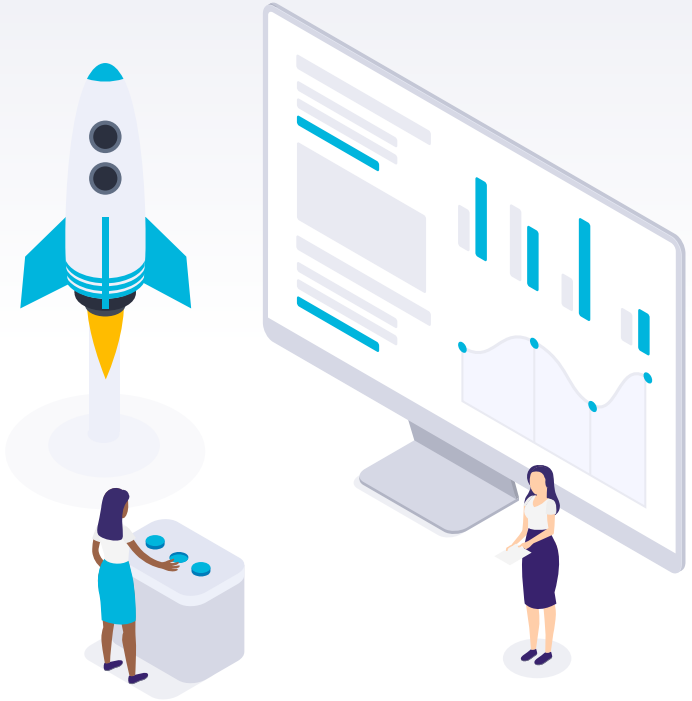
- Small number use lighter color, the higher the darker
- Color scale is confusing : start from dark and end also with dark. If it starts and ends with dark, use different color

Color legend

- Usually red is used for fatal/bad case (dead)

7

Correlation vs Causation



Correlation vs Causation

When two variables X and Y are correlated — meaning they increase together, decrease together, or one goes up as the other does down — there are four possible explanations:

A. X causes Y

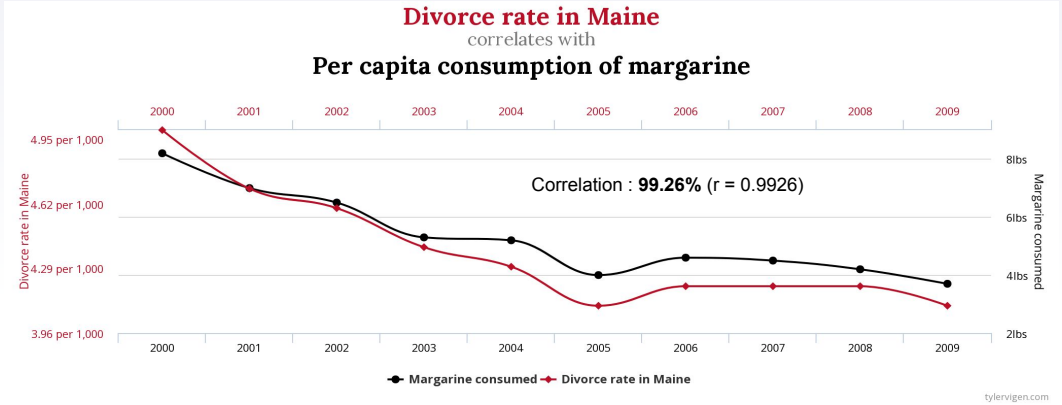
B. Y causes X

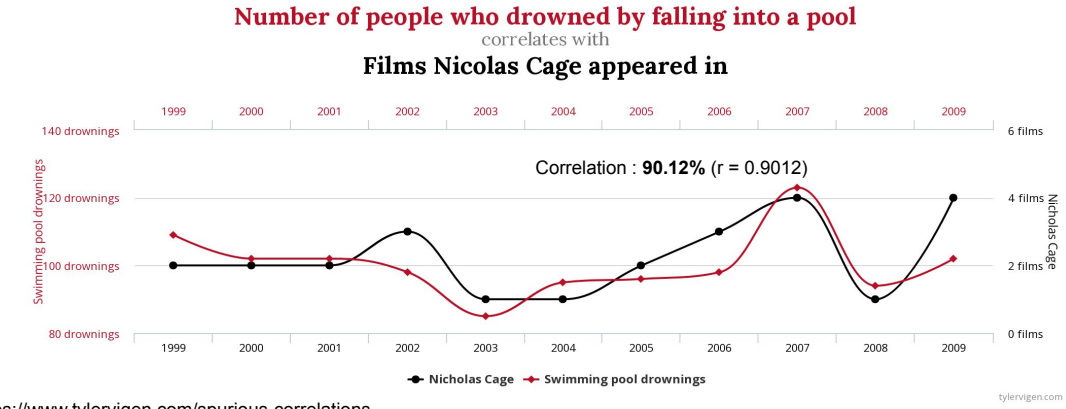
C. A 3rd variable, Z, affects both X and Y

D. X and Y are completely unrelated

We often immediately jump to — or are led to believe — A or B when C or D may be as likely.

Correlation doesn't imply causation



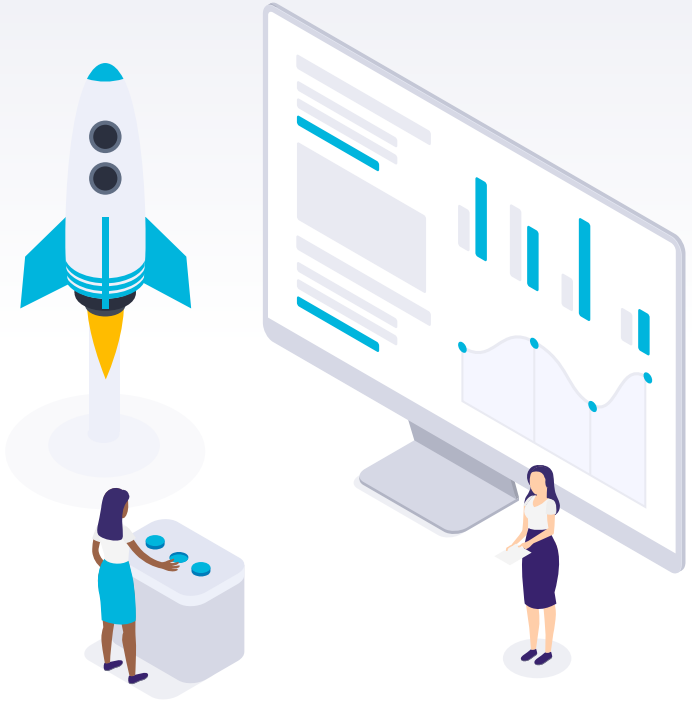


- Causation only can be obtained from randomized controlled trial / experimental study (ex : vaccine covid study, A/B testing, etc)
- When data are obtained from observational studies, we only can conclude X & Y are correlated/not-correlated. That's all.
- *Correlation is not causation but it's not not causation either*

https://www.tylervigen.com/spurious-correlations

8

How To Talk Back To A Statistic?



Who do you think are the people who are most likely to statisticulate, and for what purposes?

- *The media, to sensationalize.*
- *Politicians, to win your vote.*
- *Advertisers, to sell their product.*
- *Anyone trying to convince you of something, especially if they stand to profit from your becoming convinced.*

***Statisticulate** : is the process of misleading people using statistics.*

How to Talk Back to a Statistic?

Five questions help you avoid getting tricked by statistics.

1. Who says so?
   Is there likely to be bias in either the one analyzing the data or reporting the statistic? Is the cited authority really standing behind the statistic?
2. How does he know?
   Is there likely to be bias in the sample? Is it representative?
3. What's missing?
   Do we have everything we need to know in order to fully understand the significance of the statistic that is being offered?
4. Did somebody change the subject?
   Are definitions of all terms fully understood, and consistent for any comparisons? Is the data likely to be accurate, or was there opportunity and reason for the subjects to lie? Is correlation being represented as causation?
5. Does it make sense?
   Is it believable? Or are we being blinded by the seemingly sophisticated analysis and scientific-sounding statistic? Is it reasonable to extrapolate this far?

Critical things that are commonly missing when a statistic is reported in the media

1. Number of cases.
2. Confidence level or level of significance.
3. Variability.
4. What variety of average is being used.
5. Something to compare the figure to
6. Raw numbers to put percentages in perspective
7. Some indication of distribution to go with averages
8. Base for an index
9. The factor responsible for the change (Easter in a different month; a change in reporting procedures, lifespan, populations, diagnostic techniques.

THANKS!

Any questions?

You can find me at:

muhammad.i.fachruddin [at] gdplabs.id