Data Visualization, Good and Bad

STAT 103

Data Visualization

1. In a world where data is so important, we all want to create clear and effective charts. But the truth is, data visualization isn’t something that is usually taught in school or during on-the-job training. Instead, we often pick it up as we go, which means we sometimes make choices that can be misleading.

2. Similarly, incorrect data visualization has been frequently used in traditional media (especially TV and newspaper) and social media to mislead the audience, whether intentionally or unintentionally.

3. In this chapter, we will learn about data visualization techniques, and call out bullshit on various misleading data visualizations that are published in the media throughout the years.

General rule of thumb:

1. When visualizing categorical data, we typically want to compare the frequency or proportion of these categories. We typically use pie-charts to show the proportion of categories within a whole, bar-charts to compare quantities across different categories, and stacked bar charts to compare the composition of categories across different groups.

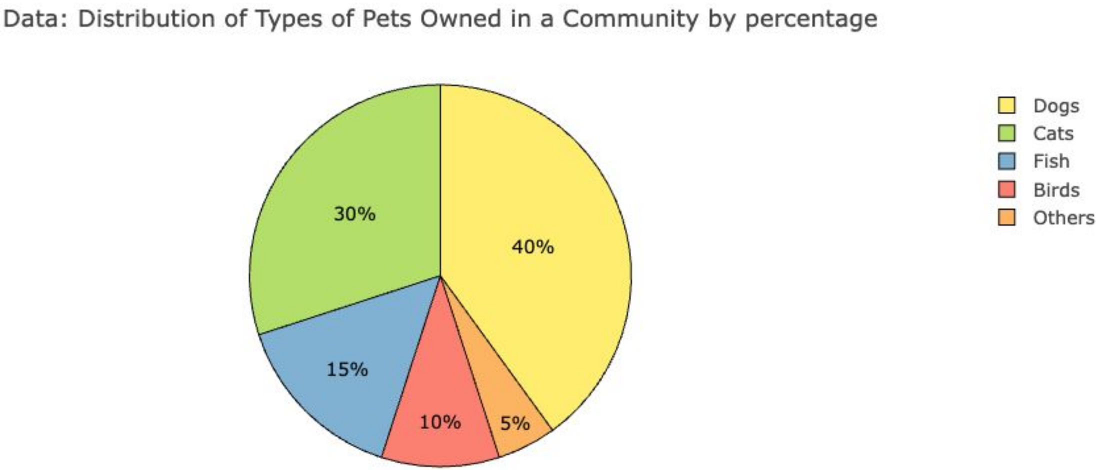
2. When visualizing numerical data, we typically want to show distributions, trends, or relationships. Line charts are ideal for showing trends over time or continuous data, histograms are great for showing the distribution of a single continuous numerical variable, and scatterplots are useful for showing the relationship between two numerical variables.

Pie chart

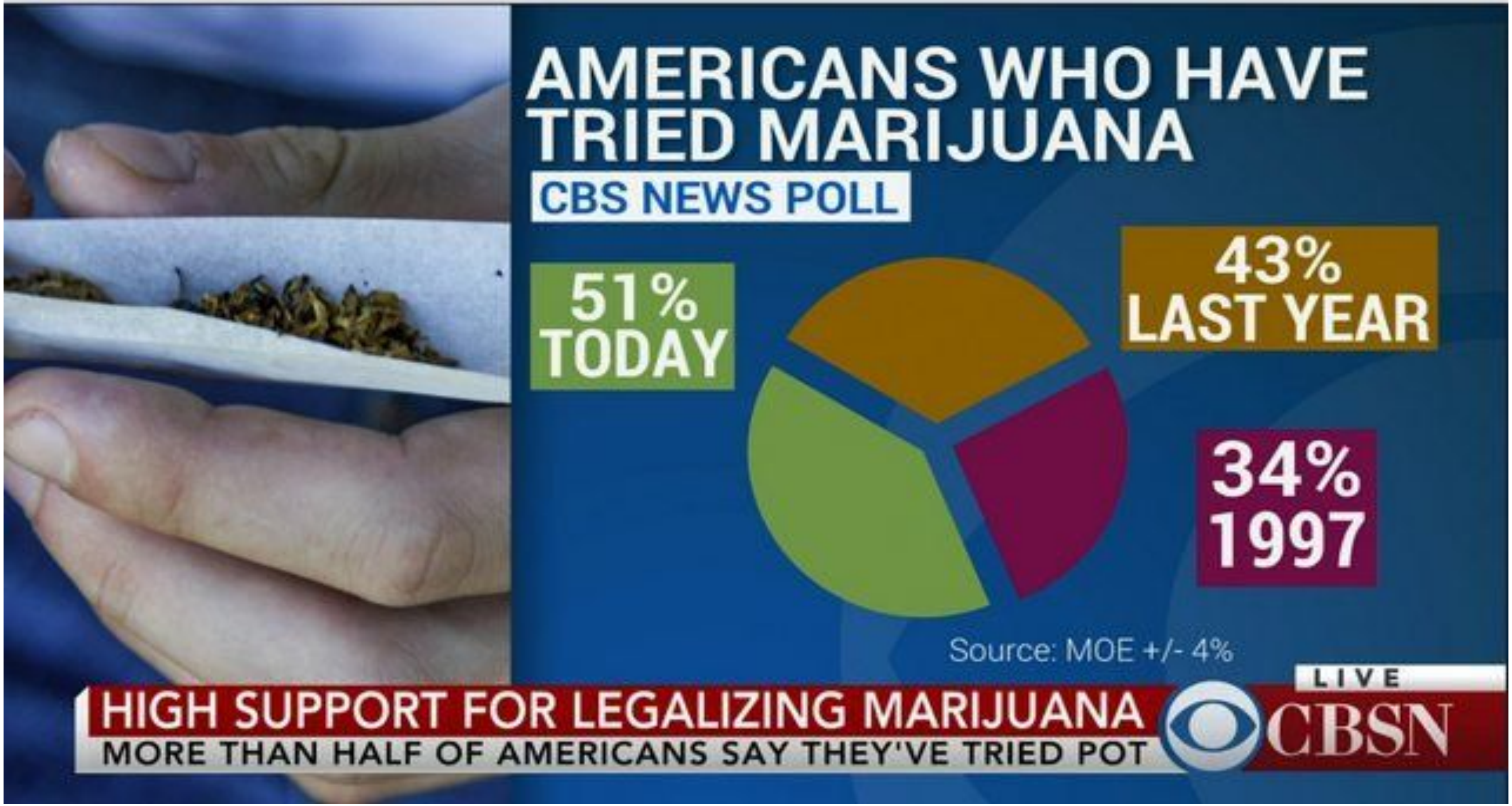
1. A pie chart is the most common technique to represent categorical data. Each slice of a pie represents a category's contribution to the whole, with the size proportional to the quantity it represents. The slices of the pie add up to a 100%.

2. Pie charts work best with a small number of categories and are most useful for simple, straightforward data visualizations.

Example: Let's visualize the distribution of pet types in a community by percentage, where the data shows 40% dogs, 30% cats, 15% fishes, 10% birds, and 5% others.



What is misleading about this news report?



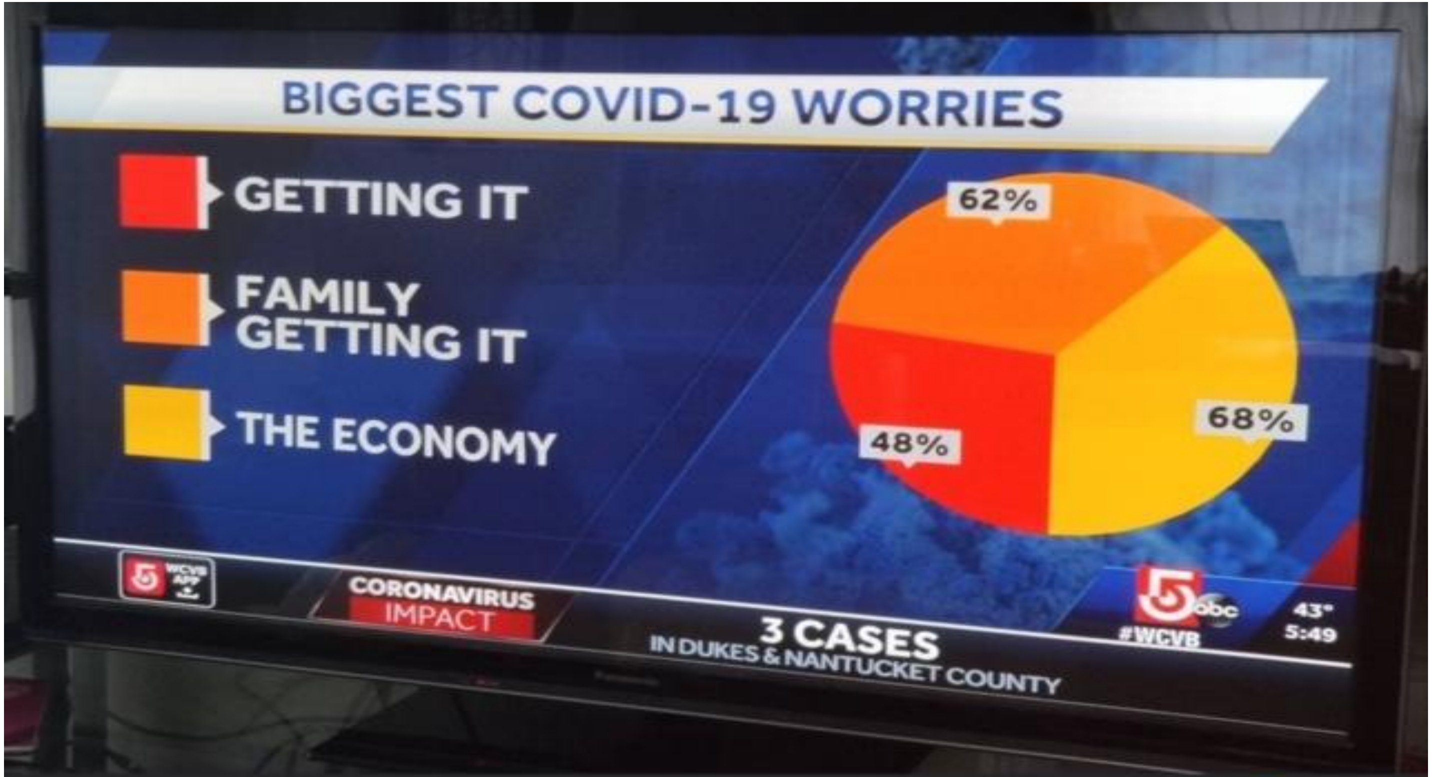
HIGH SUPPORT FOR LEGALIZING MARIJUANA

MORE THAN HALF OF AMERICANS SAY THEY'VE TRIED POT

1. Percentage of Americans who smoke marijuana in different years is a trend over time. **Pie-chart isn’t used to represent a quantitative data. Line graph is an ideal graphical representation.**
2. A line graph from 1997 to present would better represent the trend in Americans who have smoked marijuana.
3. All the individual slices of pie do not add up to 100%
4. Misleading data visualization to convince the audiences that the percentage of Americans who have smoked marijuana has drastically increased compared to 1997.

Conclusion: Bullshit!

What is misleading about this news report?



Posted in r/funny by u/veganator

reddit

1. Biggest COVID-19 worries is a categorical data, and the author is trying to show the proportion of categories within a whole. **So a pie-chart is the correct graphical representation.**

2. The individual slices of pie do not add up to a 100%. This is the most common way to mislead the audience.

3. Even if we assume that individual slices added up to 100%, red, orange, and yellow are too similar colors. It's always better to use three easily distinguishable colors.

4. There's no mention of the **sample size and source.** Did you sample 10 people in your studio or 1000 random Americans? What am I looking at?

Conclusion: Bullshit!

Bar chart

1. A bar chart is a graphical representation of data where the length or height of each bar is proportional to the value it represents.

2. Bar charts are commonly used to compare quantities across different categories.

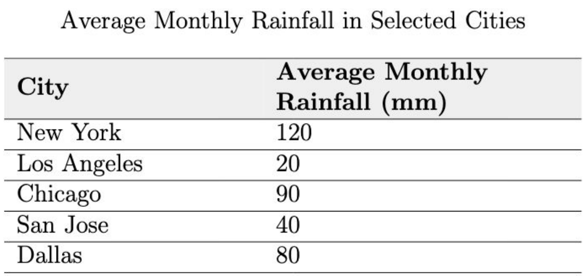
3. To create a bar chart, data is first collected and organized into categories, with each category assigned a corresponding value. These values are then represented by bars, either horizontally or vertically, with the axis providing a scale for measurement.

4. Bar charts should be used when comparing different groups or tracking changes over time, especially when changes are larger.

5. However, it is important to avoid using too many categories, which can make the chart cluttered and hard to read.

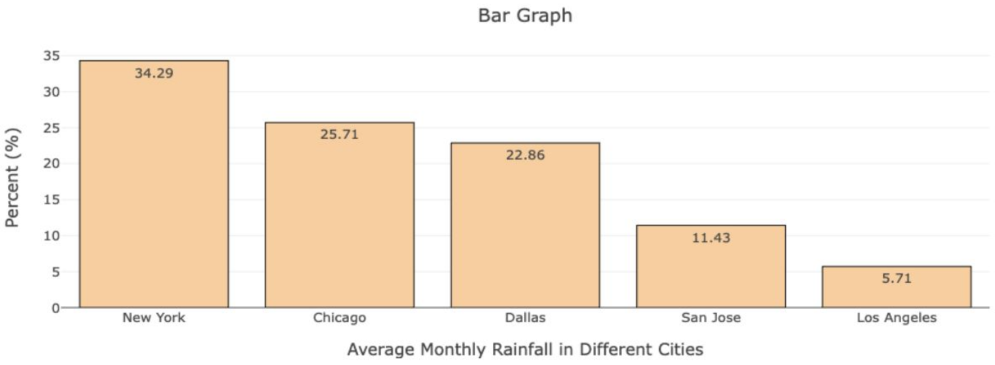
Example of a barchart

A barchart can be used to display the average monthly rainfall in different cities for the data below.



Average Monthly Rainfall in Selected Cities

City	Average Monthly Rainfall (mm)
New York	120
Los Angeles	20
Chicago	90
San Jose	40
Dallas	80

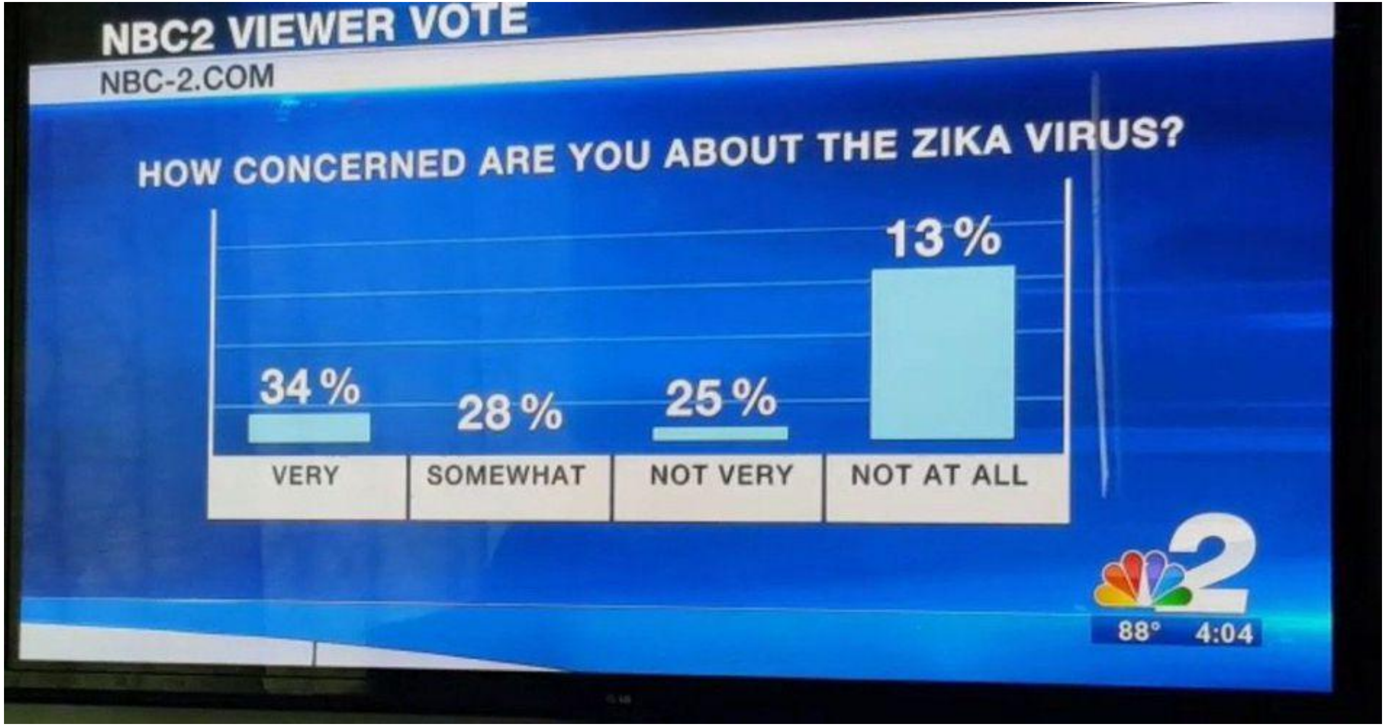


Most common mistakes in bar charts

Bar charts are one of the most commonly used graphical representations. As a result, it should be no surprise that a lot of misleading bar charts are found in the media to mislead the audience.

Now we have seen what a good bar chart should look like, now we will study the common mistakes to avoid in bar chart and how media can use these tactics to mislead the audience.

Common mistake #1.

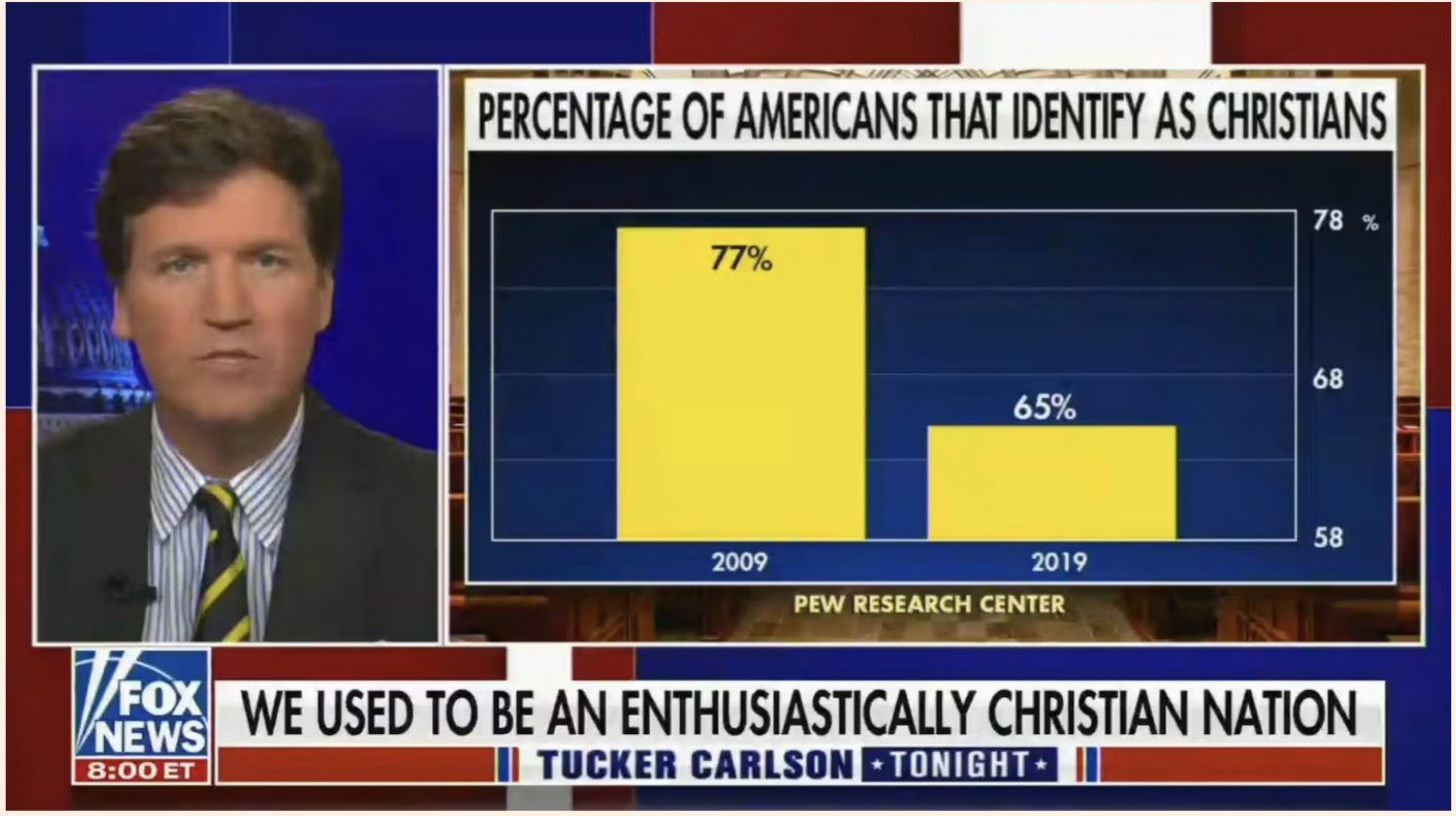


1. No calibration of vertical axis (y-axis). A good bar chart always has the vertical axis starting from 0.

2. No consistency in calibration of percentages on the vertical axis. 13% has the highest bar where in reality it should be the lowest. 28% doesn’t even have a bar to it.

3. Clearly a misleading bar chart to forcefully show that NBC2 viewers are “not at all” concerned about the zika virus.

Common mistake #2.



WE USED TO BE AN ENTHUSIASTICALLY CHRISTIAN NATION

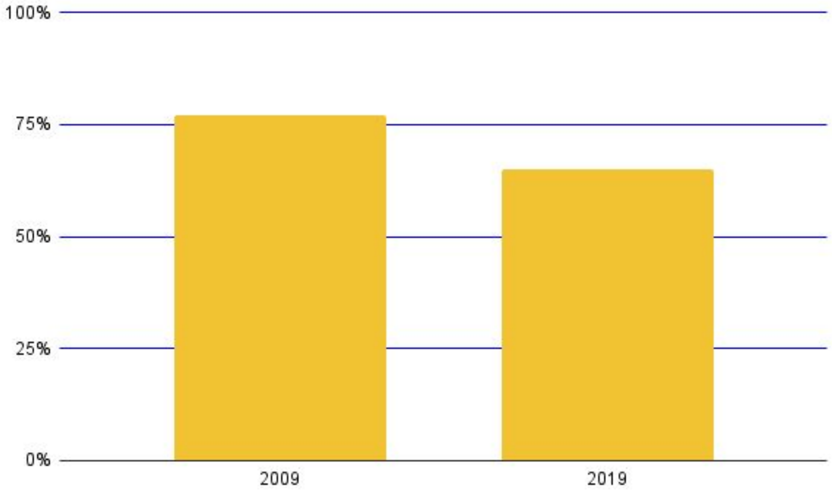
TUCKER CARLSON ⋆TONIGHT⋆

1. A good bar chart always has the vertical axis starting from 0 so that we have an even comparison. Not starting the vertical axis from 0 is the most common ways to mislead the audience and make a flawed bar chart.

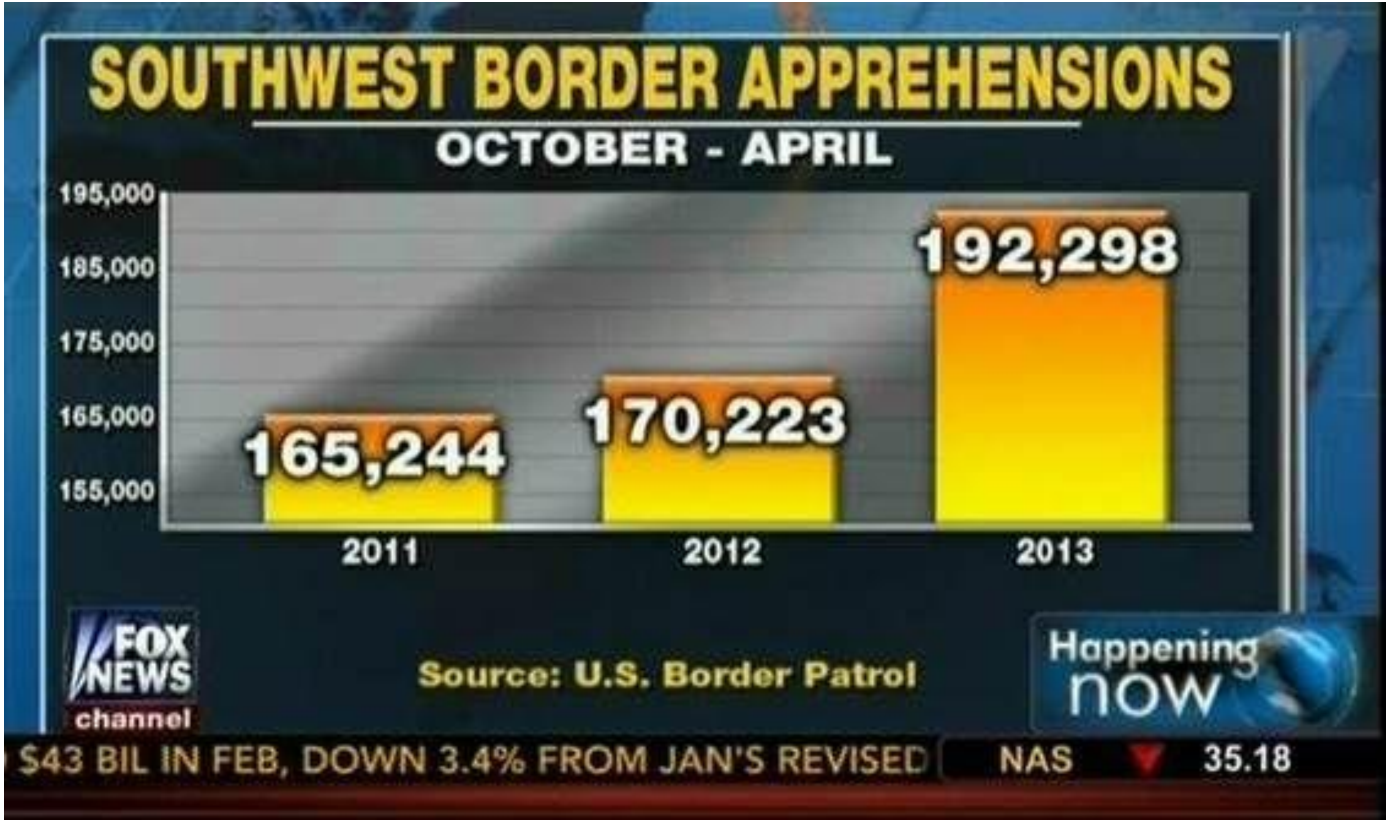
2. Here the y-axis starts at 58% instead of 0%, exaggerating the difference between the two bars. The difference between 77% and 65% is significant, but the chart makes it look far more drastic than it actually is.

3. This is how the bar chart would look like if

   y-axis started at 0.

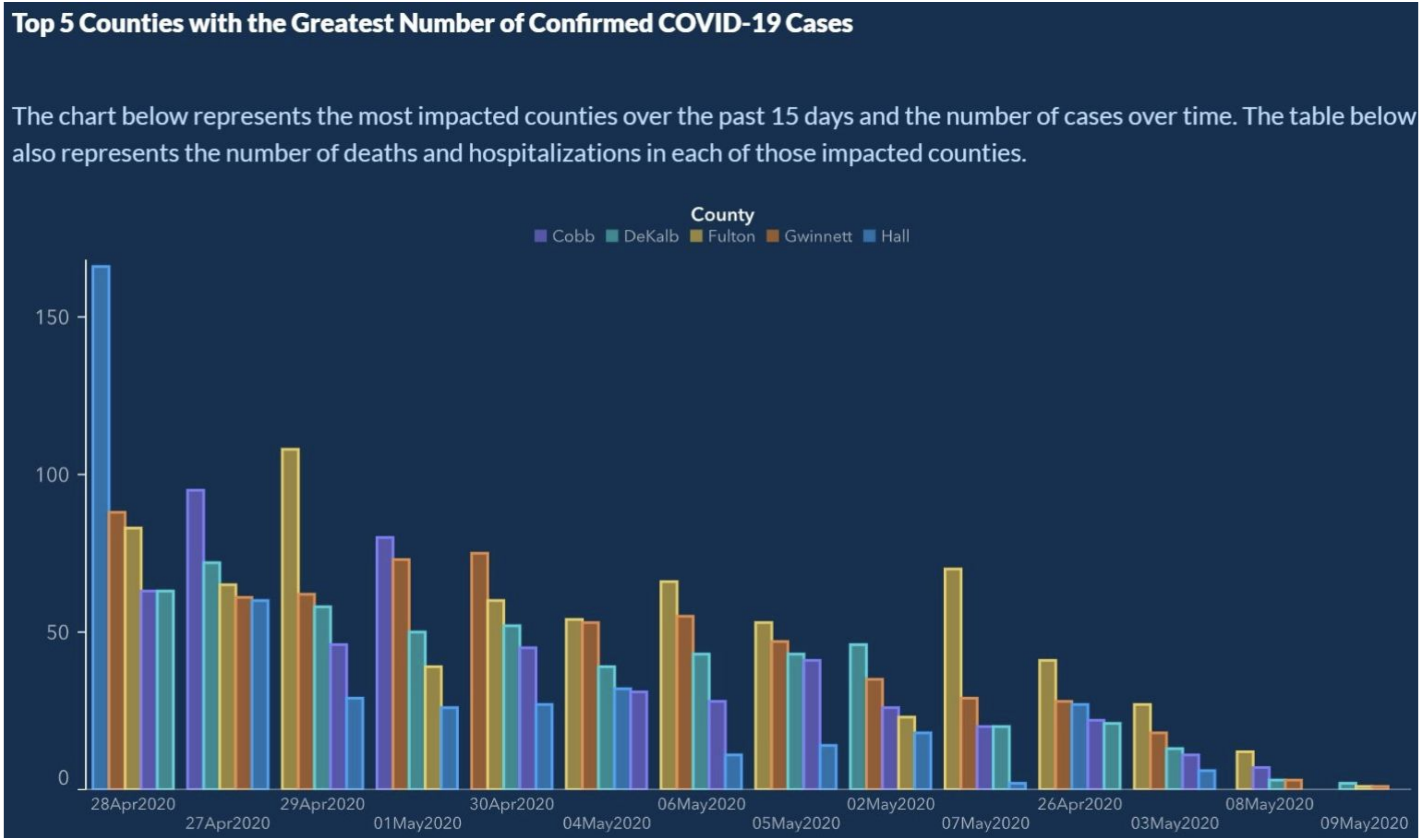


What’s the problem here?



Here the y-axis starts at 155,000 instead of 0. This makes the change look drastic than it actually is.

Common Mistake #3.



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

- In this graph we can see that, x-axis is not on a chronological order. This is one of the most commonly used technique to make a deceptive graph.

- Both axes should have consistent, uniform increments. For example, if the vertical axis starts with increments of 5 (0, 5, 10, 15, 20), it shouldn't suddenly jump to larger values like 70, 80, 90, 100.

- Always be on the lookout for whether the axes are properly calibrated or not.

Line Chart

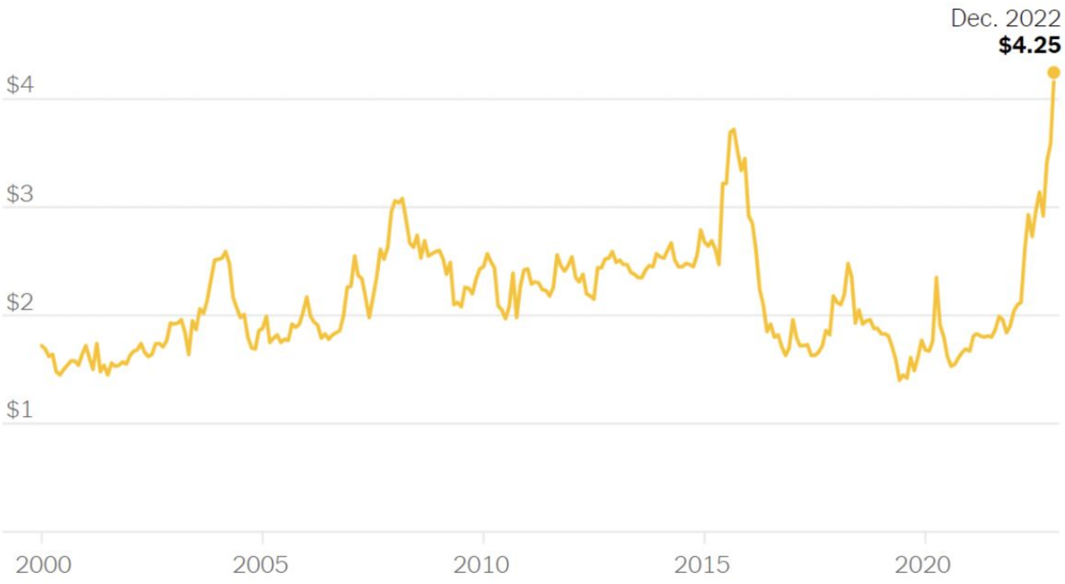
Line charts are commonly used to display trends or changes over time, making them ideal for visualizing continuous data. To create a line chart:

- Data points are first collected and plotted on a graph, typically with time on the x-axis and the variable of interest on the y-axis.
- These points are then connected by lines, which help to illustrate the trend or pattern over the given period.

The typical errors seen in bar charts, such as incorrect axis calibration and inconsistent increments on both axes, also apply to line charts.

!!! It's generally acceptable not to start a line chart at 0, but it depends on the context. If starting the y-axis at a value other than 0 provides a clearer view of the data's trends or variations, it can be appropriate. However, it's important to be cautious, as not starting at 0 can sometimes exaggerate differences and potentially mislead the viewer. !!!

Example of a line chart

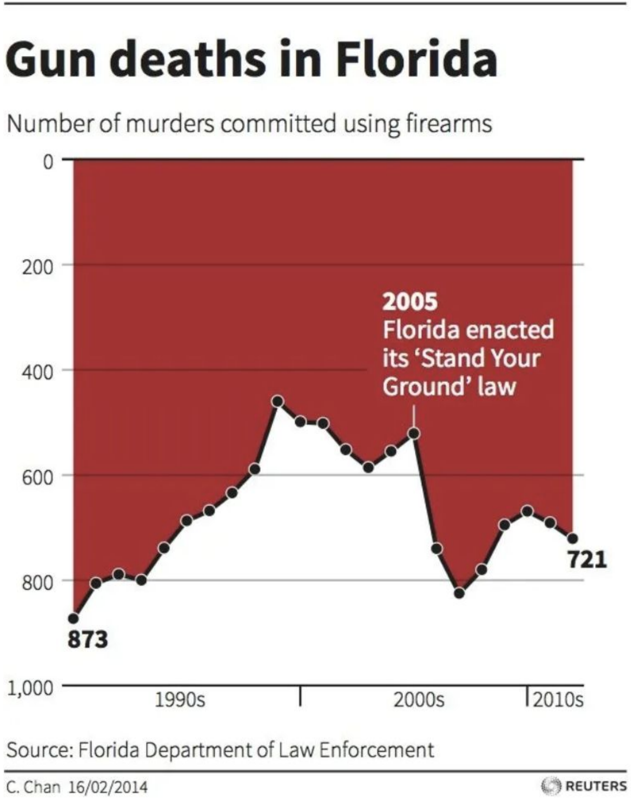


Monthly average price of a dozen eggs

Note: Prices are in December 2022 dollars. • Source: Bureau of Labor Statistics

Now we have drilled down the basics of data visualization, let's look at misleading data visualizations that have been seen over the years at various places and call out bullshit.

What’s wrong with this?

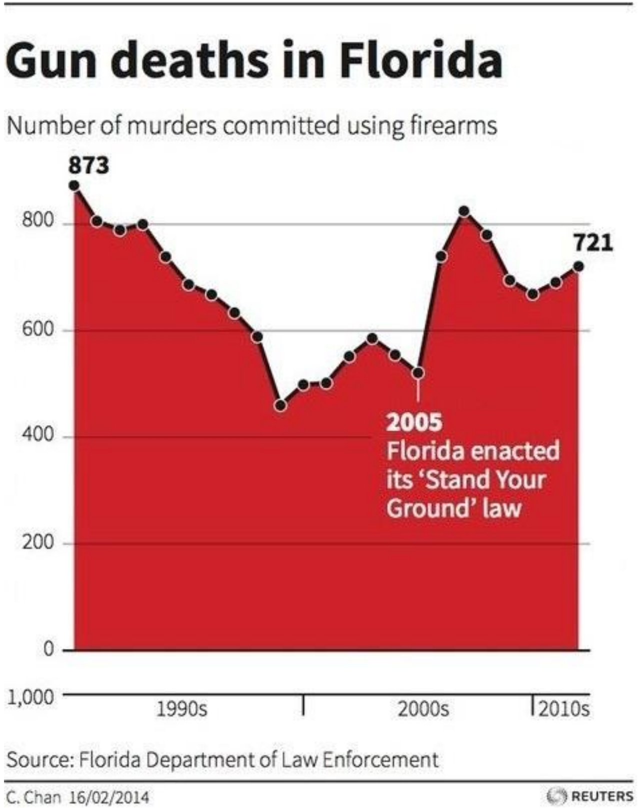


- It looks like the gun death in Florida was on a rise in 1990s and once ‘ Stand Your Ground law’ was enacted in 2005, then the firearm murder dropped drastically

- However, the vertical axis inverted!
  It doesn’t start at 0.

- What we usually perceive as an increasing trend is actually a decreasing trend, and vice-versa.

- This is what a corrected graph would look like.



Gun deaths in Florida

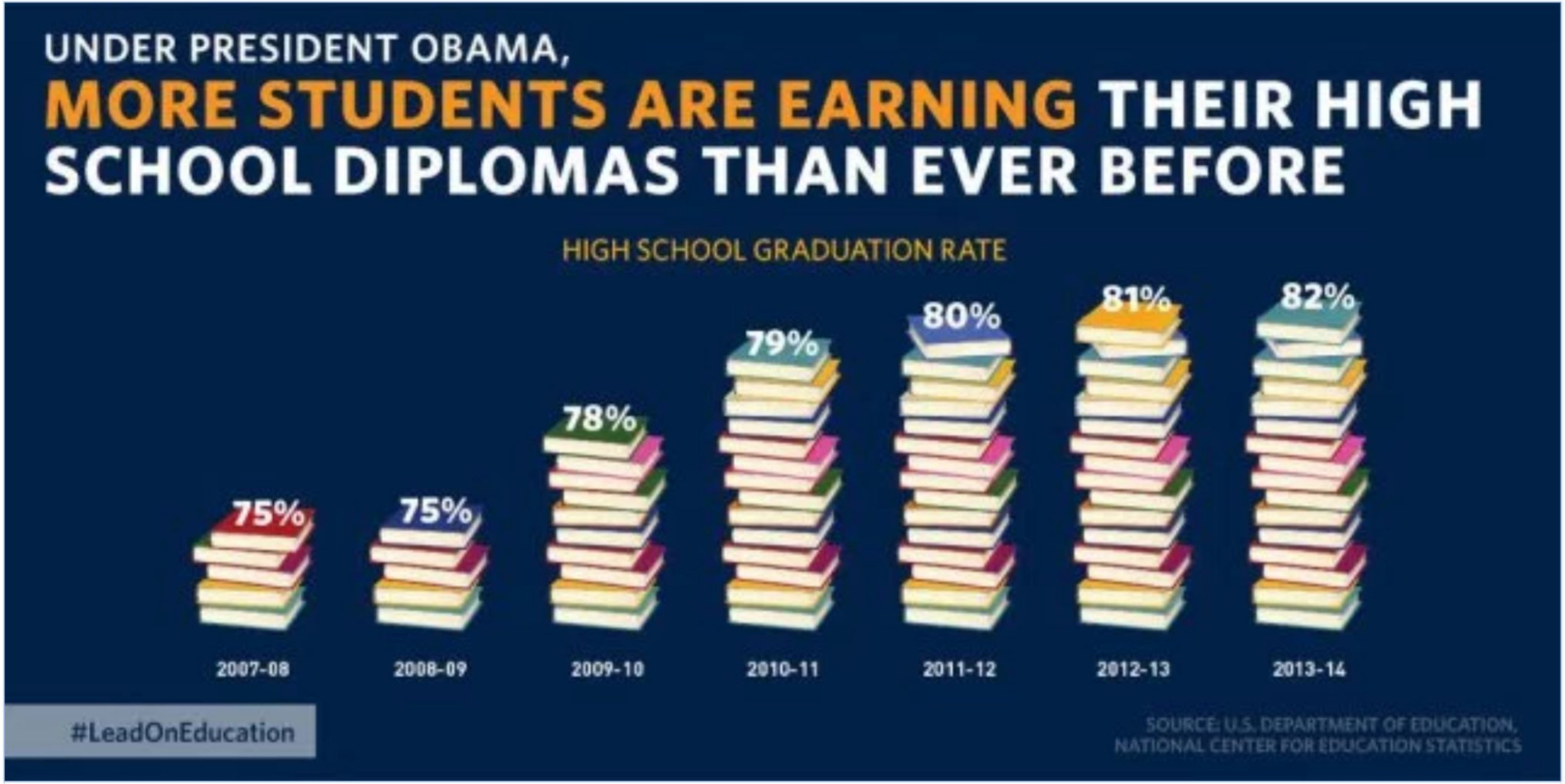
Number of murders committed using firearms

Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

Let’s look at some pictograms. What’s the verdict?



UNDER PRESIDENT OBAMA,

MORE STUDENTS ARE EARNING THEIR HIGH SCHOOL DIPLOMAS THAN EVER BEFORE

HIGH SCHOOL GRADUATION RATE

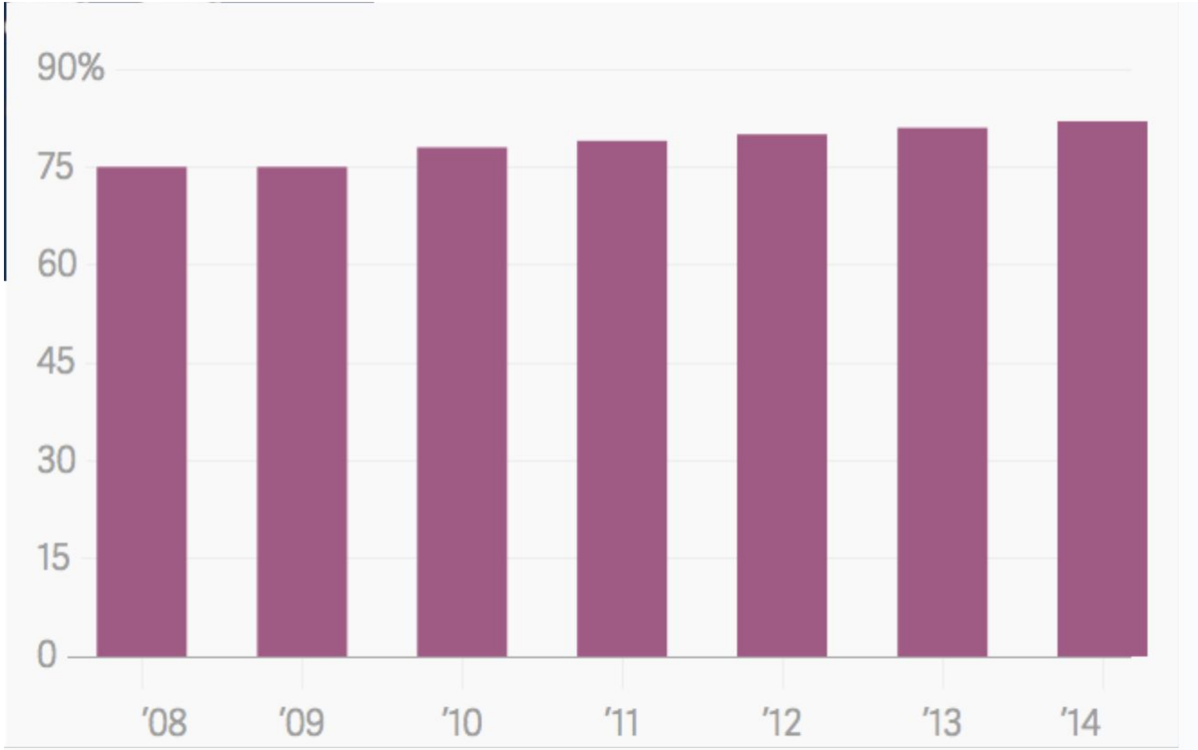
SOURCE: U.S. DEPARTMENT OF EDUCATION, NATIONAL CENTER FOR EDUCATION STATISTICS

1. High school graduation rate by year a numerical data, and the **line graph is the ideal graphical representation.** In this case, if you are treating seven different years as seven different categories, barchart is okay.

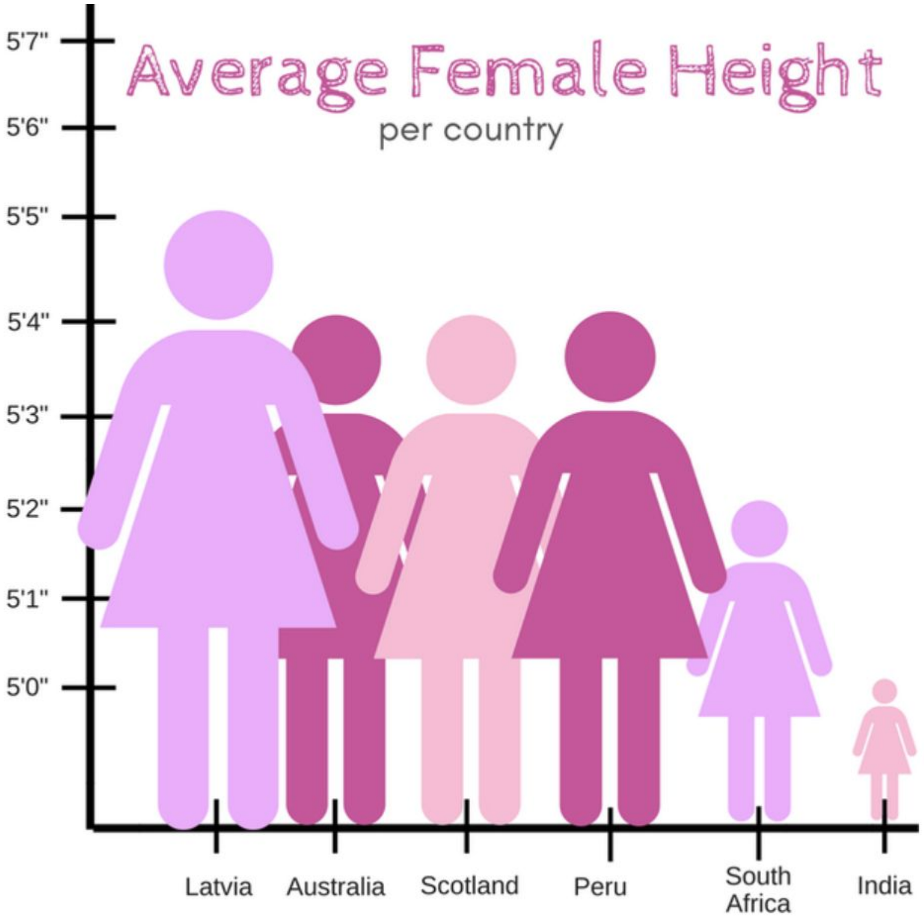
2. Another classic mistake of not starting the vertical axis from 0. Vertical axis isn't starting from 0, hence the audiences are being misled to think there is a significant improvement of high school graduation rate under Obama whereas it was just an increase of 7% in seven years.

3. The use of books **isn't necessary** as well. A simple bar chart would be the correct representation of the data which can be seen in the following slide.

There is a drastic change in how the audience will perceive this correct way of representing the data to the misleading way of representing the data seen before. Reminder again, it is extremely easy to mislead the audience with bad data visualizations.



What’s the verdict on this graph?



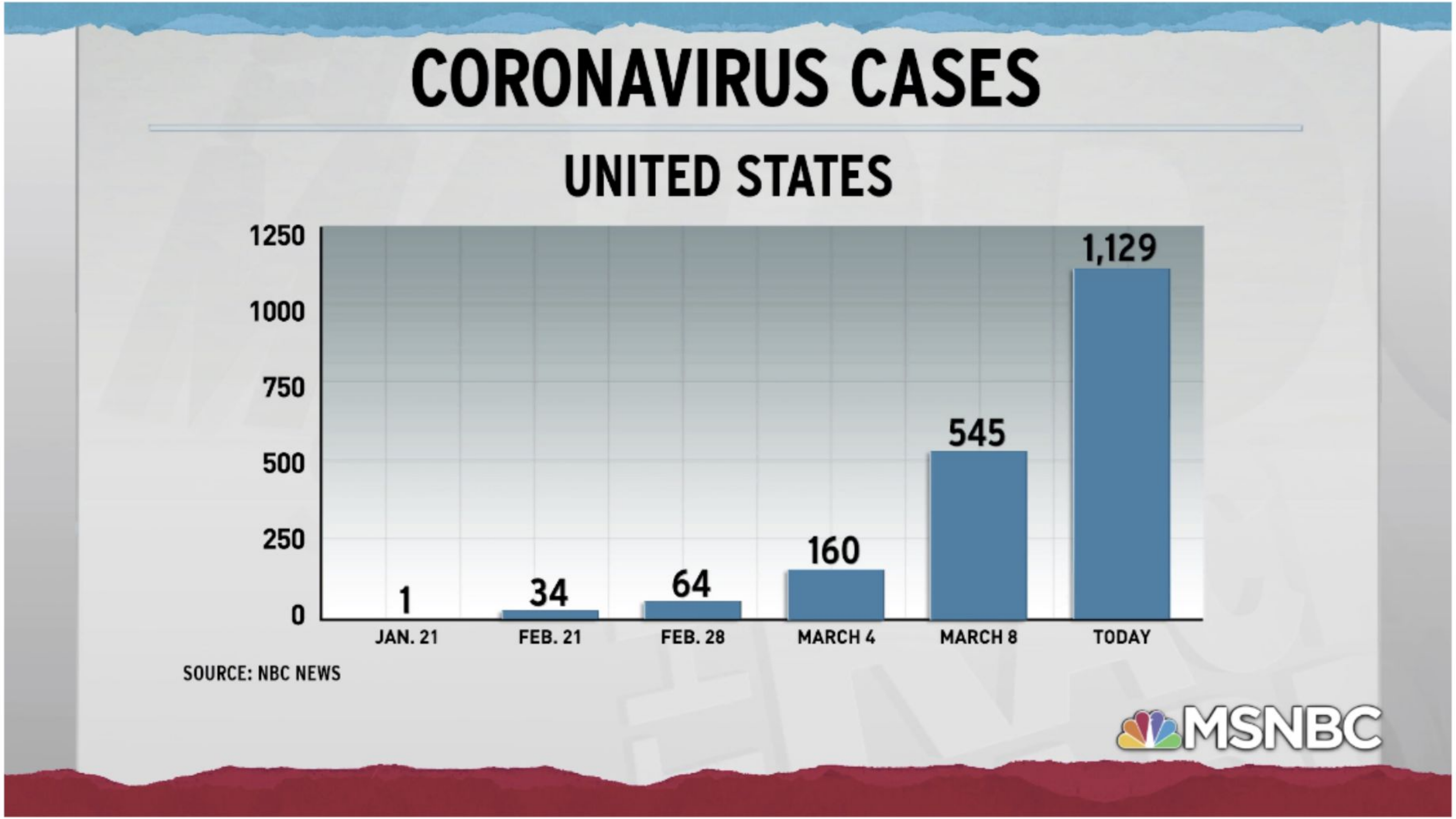
While using pictograms, if the pictures are not scaled uniformly, this creates a perceptually misleading comparison. The area of the pictogram is interpreted instead of only its height or width. This is the classic method that is used to mislead the audience using pictograms

Two classic mistakes which leads to it being a misleading graph:

1. The vertical scale does not start at 0.

2. The visualization is not drawn up to scale. The scaling used to create pictures in the pictogram are not consistent.

This is from the March 12 broadcast of "The Rachel Maddow Show." What's the verdict?

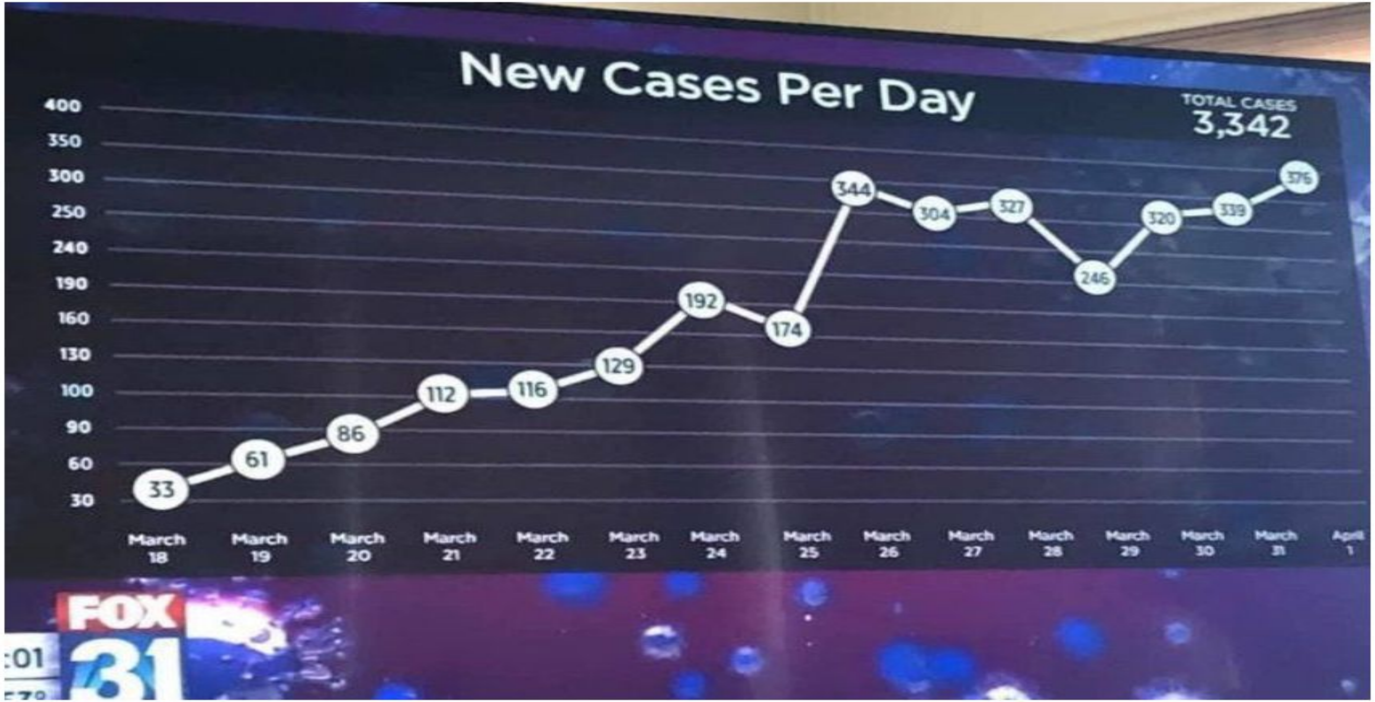


1. The bar chart looks fine at the first glance, but it is a misleading way of data visualization.
2. COVID-19 cases by day is a numerical data, and so line graph is the ideal graphical representation to highlights trends over time. In this case, you should be using a line graph instead of a bar chart.
3. Even at a barchart, there is an uneven spacing on the horizontal axis (x-axis). It is jumping from January 21 to February 21st to February 28 and an increment of 4 days. NEVER do this! There should always be an even spacing between the dates on for correct data visualization.

Hence, the verdict is that the data visualization is not as misleading as the ones we've seen before, but it is still a bad graph nonetheless.

Exercise:What’s the verdict on this graph?

This graph was posted on Twitter with the caption “You can teach an entire semester of how to lie with statistics with the y-axis of this chart”: Can you spot the problem?



Always start with: Is the graphical representation technique a good one to represent the kind of data you're dealing with?

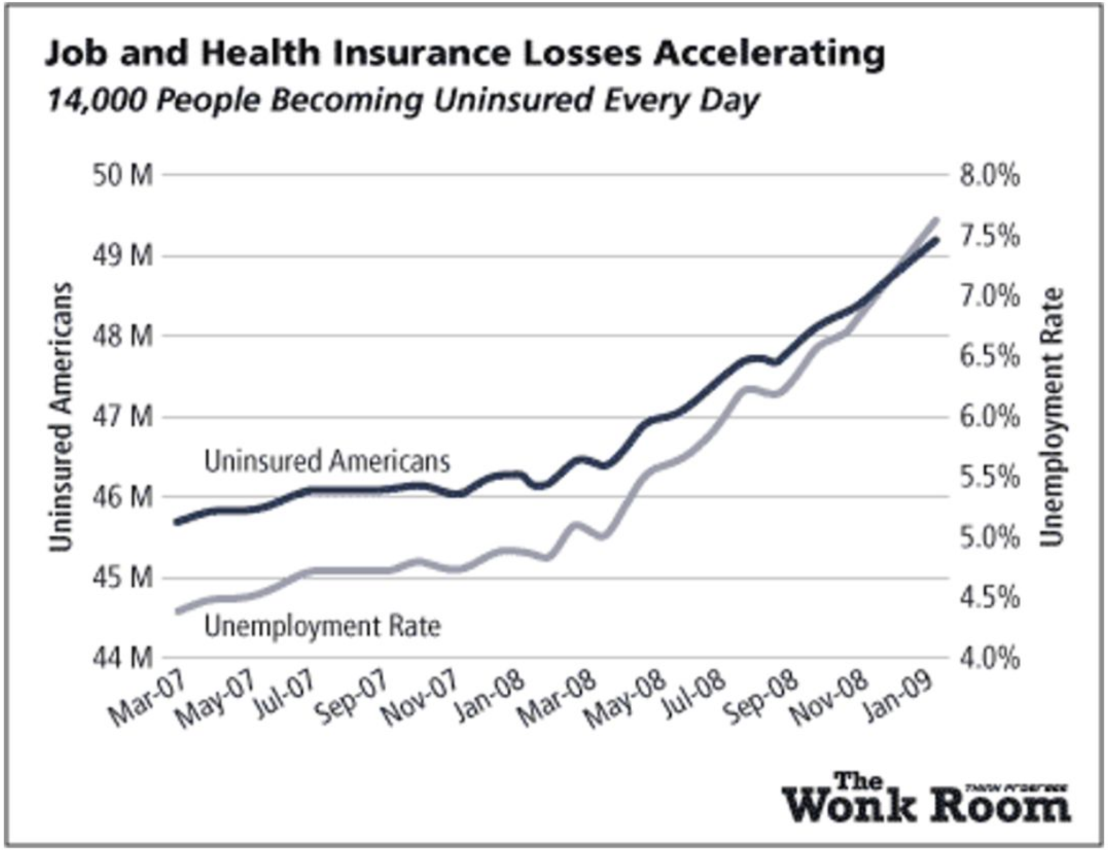
COVID-19 Cases per day is numerical data. So, line graph is the ideal graphical representation to highlights trends over time. Look carefully at the y-axis and the uneven graph intervals. It gives the graph a flattened look.

Hence, some graphs are intended to deliberately mislead. Don't be surprised if you can't immediately see what's wrong with the graph: like everything, it comes it a lot of practice.

Now we have looked at some data visualizations which are clearly misleading and are easy to spot, let’s look at some not so clear ones.

Calling Bullshit - Part 2:

What’s the verdict on this one?



Job and Health Insurance Losses Accelerating

14,000 People Becoming Uninsured Every Day

- The graph is trying to compare the amount of uninsured Americans to the unemployment rate over time. Job and health insurance losses is numerical data, so line graph is the ideal graphical representation to highlight trends over time.

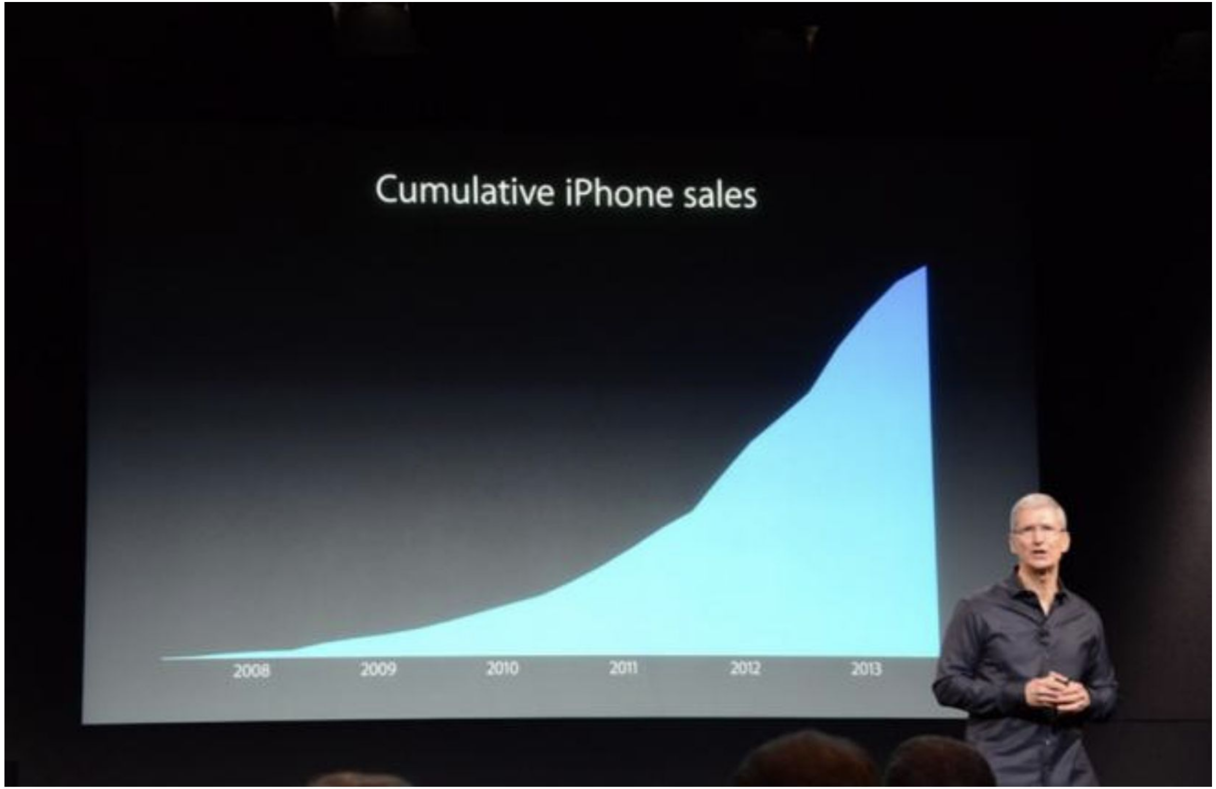
- Recap: if you can justify that there are no data observations below a certain value and one just zoomed into the trend by not starting the vertical axis at 0, it is acceptable to not start the line graph at 0. However, one should be extremely careful in doing so that the visualization doesn't mislead the audience. In this case, we do not know whether the author was honest in just zooming on the trend or intentionally misleading the audience.

- Similarly, in this graph, there are two y-axis on the same graph. This graph is doing an apples-to-oranges comparison by comparing amount of uninsured Americans to unemployment rate over time which have two totally different observations.

- A better data visualization would be to create two different line graphs: one for uninsured Americans and the other one for unemployment rate over time.

- Final verdict: Graph not as misleading as the ones we saw earlier, but a bad data visualization nonetheless.

What’s the verdict on this one? What does Tim Cook doesn’t want you to see?

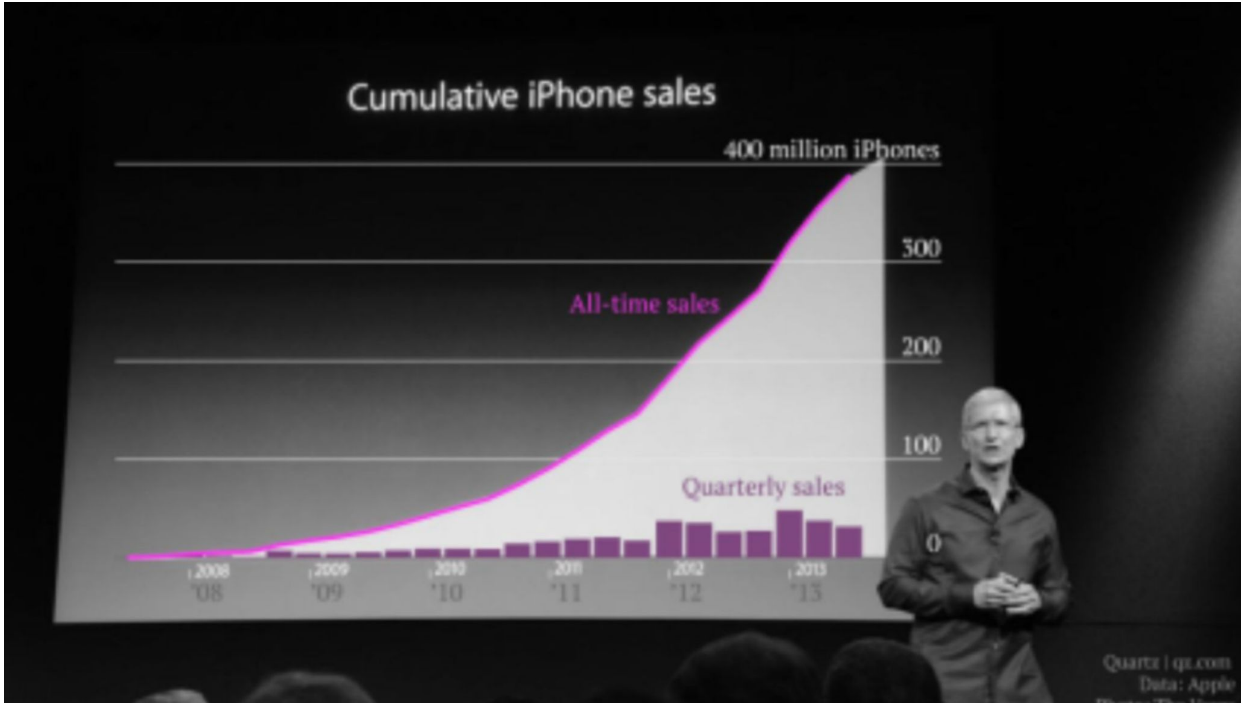


- IPhone sales by year is a numerical data, so line graph is the ideal graphical representation to highlight trend over time.

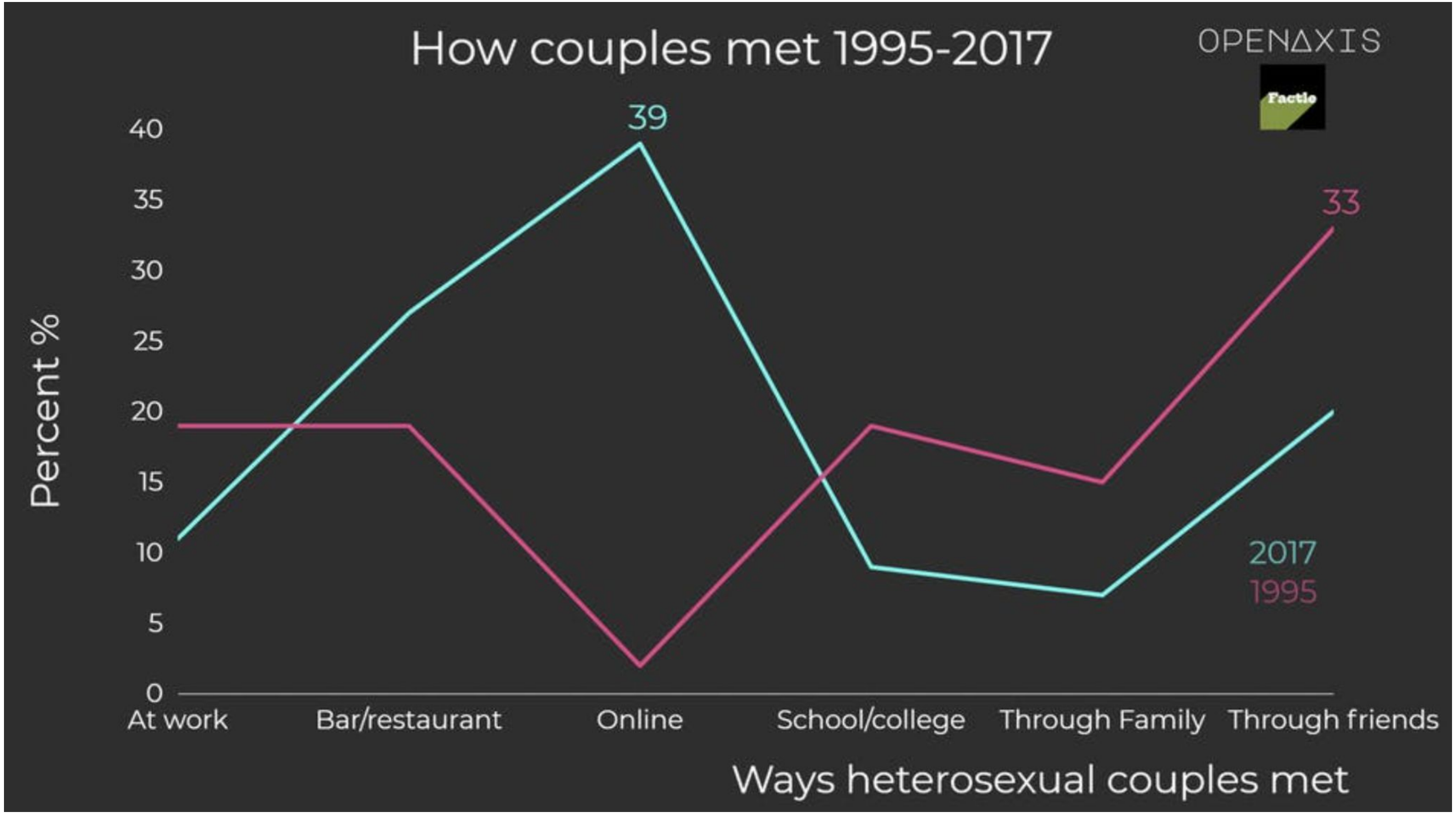
- This graph is entirely misleading because the graph is plotting cumulative sales over time. If you plot cumulative sales, your graph can only go up. Ofcourse, iPhones sales are increasing every year, so the cumulative sales will only be drastically up every year.

- Statisticians immediately looked at the quarterly reports of iPhone sales and found out the iPhones sales were decreasing in the two consecutive quarters which is shown on the next slide.

So to hide the fact that quarterly sales were decreasing on the last two quarters, Apple plotted a cumulative sales graph. Extremely misleading!!



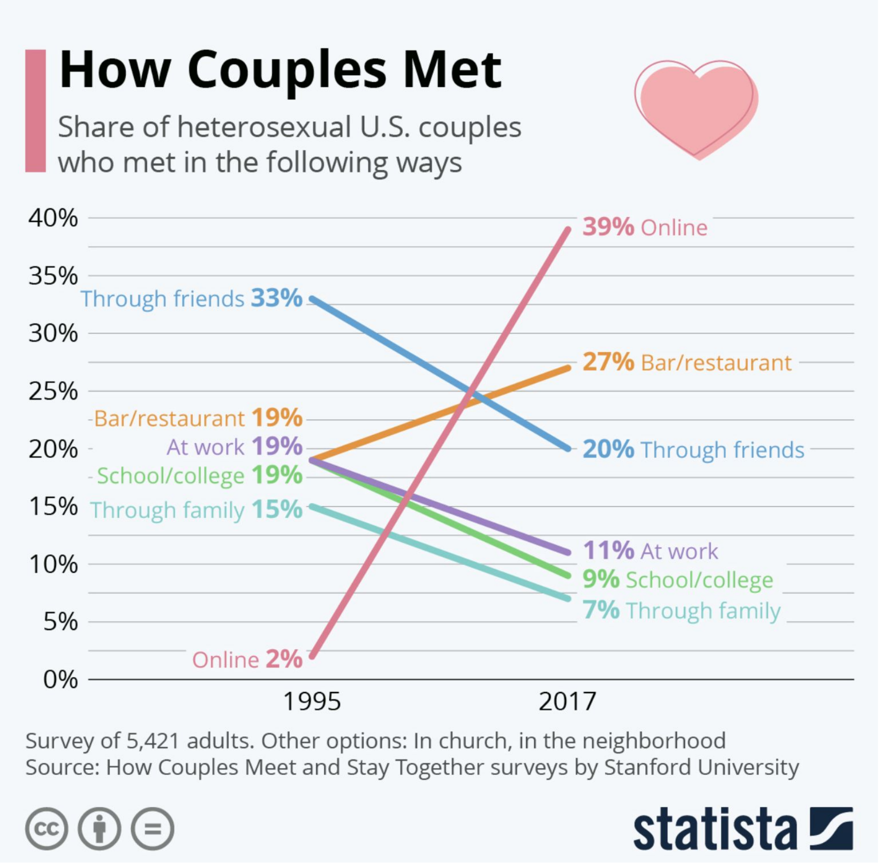
What do you think about this one?



1. Let’s start with: Is the graphical representation technique a good one to represent the kind of data you’re dealing with? We DO NOT use a line chart to represent a categorical data.
   Side by side barchart would have been the ideal and appropriate graphical representation.

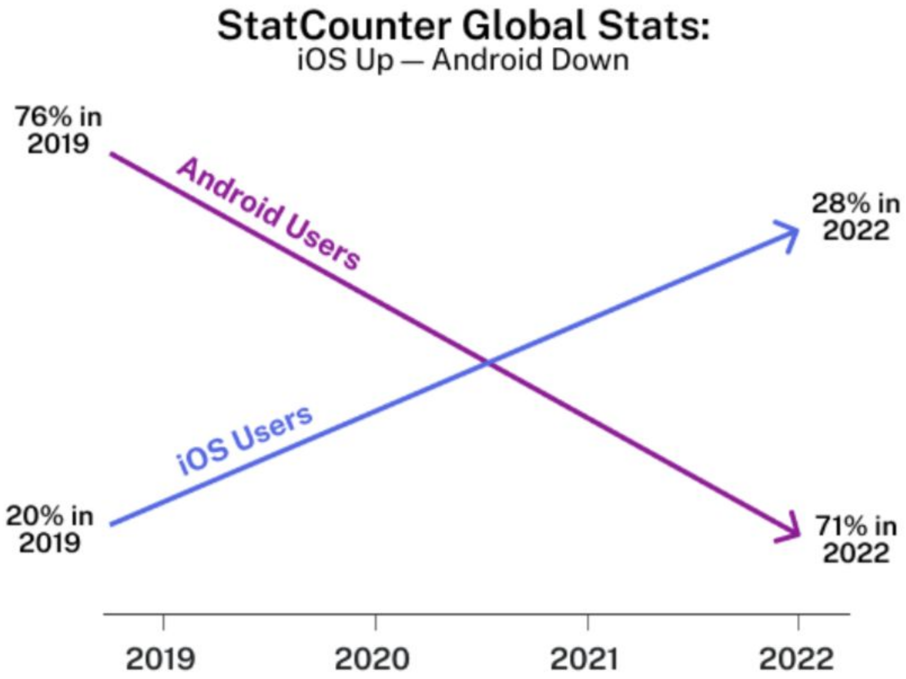
2. Connecting the dots between the categories tries to show the trend, but it makes the data look like a time series. This is not a time series data.

3. This is a proper way to represent the data if you are hell bent on using lines.



Survey of 5,421 adults. Other options: In church, in the neighborhood
Source: How Couples Meet and Stay Together surveys by Stanford University

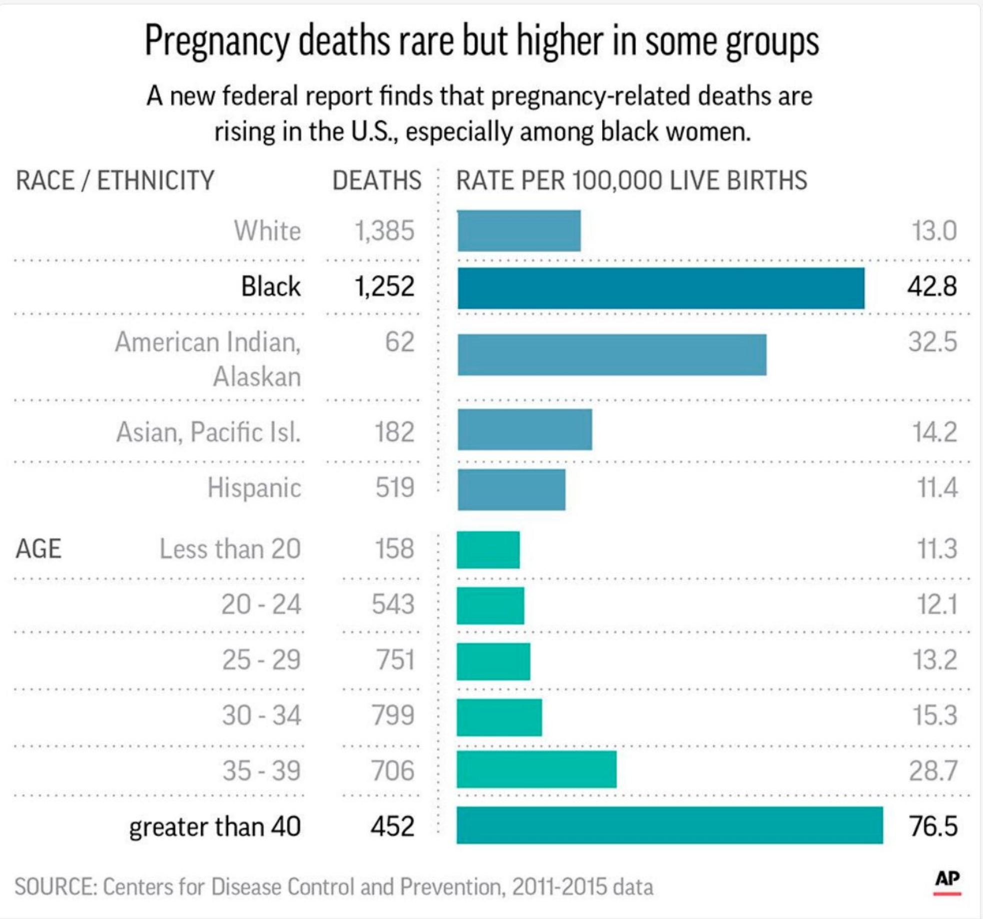
Mini class activity: Discuss this among yourselves.



1. Without a y-axis, any line is placed arbitrarily on the graph because the distance between two points cannot be calculated.

2. Also, the differences between groups can be incorrectly compared. For example, the ending number of Android Users was 71% in 2022, but that point is below the number of iOS Users reported in 2022, which was 28%

The following graphical display provides a visualization of pregnancy-related deaths in the U.S., especially among black women.



Pregnancy deaths rare but higher in some groups

A new federal report finds that pregnancy-related deaths are rising in the U.S., especially among black women.

RACE / ETHNICITY		DEATHS	RATE PER 100,000 LIVE BIRTHS
	White	1,385	13.0
	Black	1,252	42.8
	American Indian, Alaskan	62	32.5
	Asian, Pacific Isl.	182	14.2
	Hispanic	519	11.4
AGE	Less than 20	158	11.3
	20 - 24	543	12.1
	25 - 29	751	13.2
	30 - 34	799	15.3
	35 - 39	706	28.7
	greater than 40	452	76.5

SOURCE: Centers for Disease Control and Prevention, 2011-2015 data

1. Is this graph clear in communicating patterns/trends?

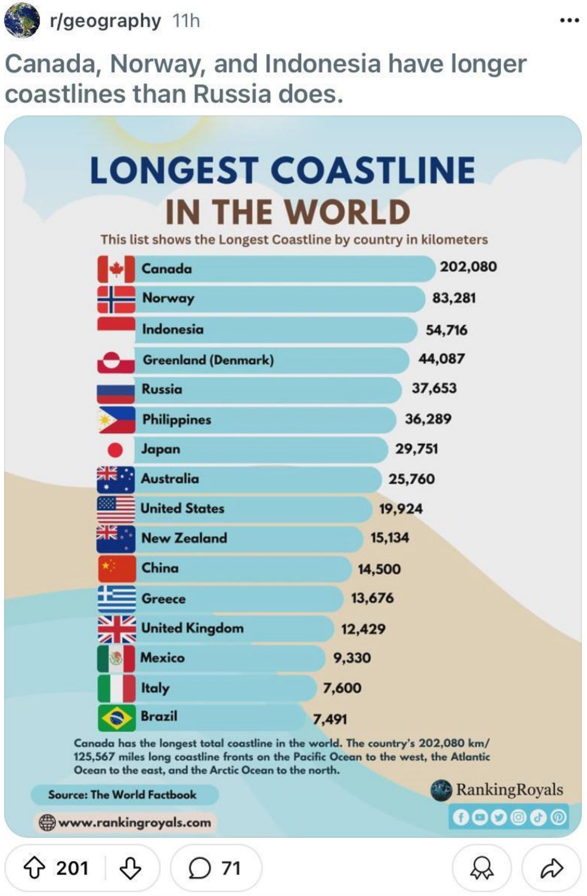
2. What are we supposed to be comparing in this graph?

3. What do the different colors represent?

4. What is the scale for the bar chart?

None of the answers to these questions are clear. In particular, note that the bar representing 42.8 deaths per live births is nearly as long as the one representing 76.5. This graph may be misleading, but it is difficult to understand how to interpret death rates “rising” with no baseline to use for comparison.

Similar to the last example. Discuss this among yourselves.



r/geography 11h

Canada, Norway, and Indonesia have longer coastlines than Russia does.

LONGEST COASTLINE IN THE WORLD

This list shows the Longest Coastline by country in kilometers

Canada	202,080
Norway	83,281
Indonesia	54,716
Greenland (Denmark)	44,087
Russia	37,653
Philippines	36,289
Japan	29,751
Australia	25,760
United States	19,924
New Zealand	15,134
China	14,500
Greece	13,676
United Kingdom	12,429
Mexico	9,330
Italy	7,600
Brazil	7,491

Canada has the longest total coastline in the world. The country's 202,080 km/ 125,567 miles long coastline fronts on the Pacific Ocean to the west, the Atlantic Ocean to the east, and the Arctic Ocean to the north.

Source: The World Factbook

www.rankingroyals.com

RankingRoyals

201 71

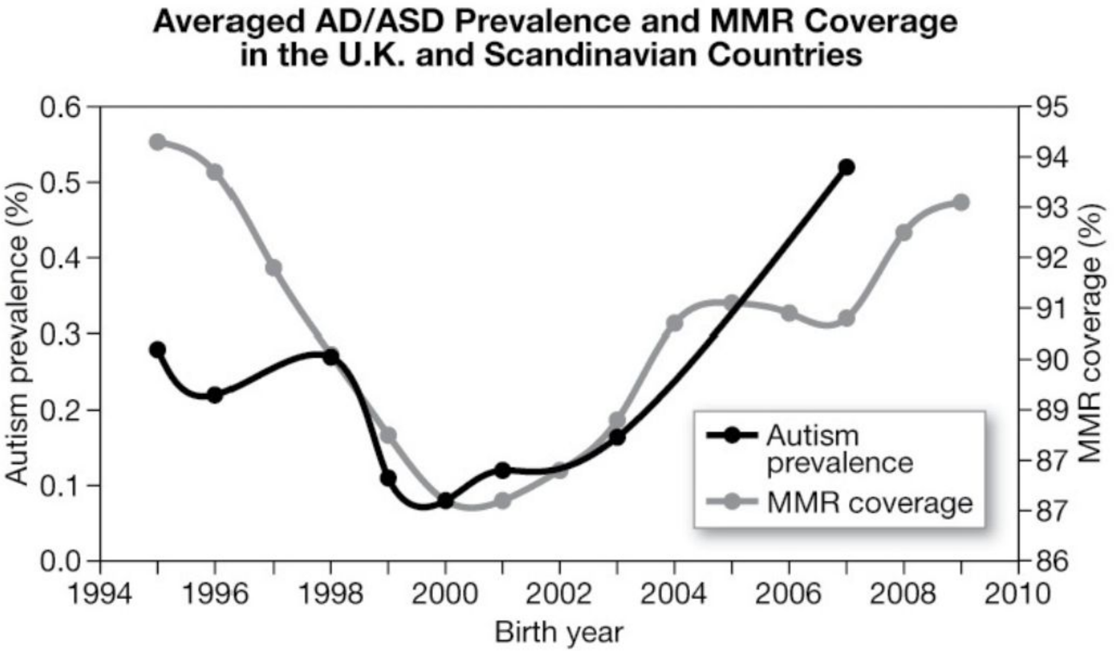
- The lengths of the bars are not proportionate to the values they represent. For example,

a) the bar for Indonesia (54,716 km) appears to be almost the same length as the bar for Norway (83,281 km), even though Norway's coastline is significantly longer. This misrepresents the relative differences between the countries' coastlines.

b) the difference between Russia (37,653 km) and the Philippines (36,289 km) seems negligible, but the visual difference on the graph is exaggerated or minimized based on inconsistent scaling.

- Any data visualization that is not drawn up to scale is immediately a misleading one!

What do you think about this graph?

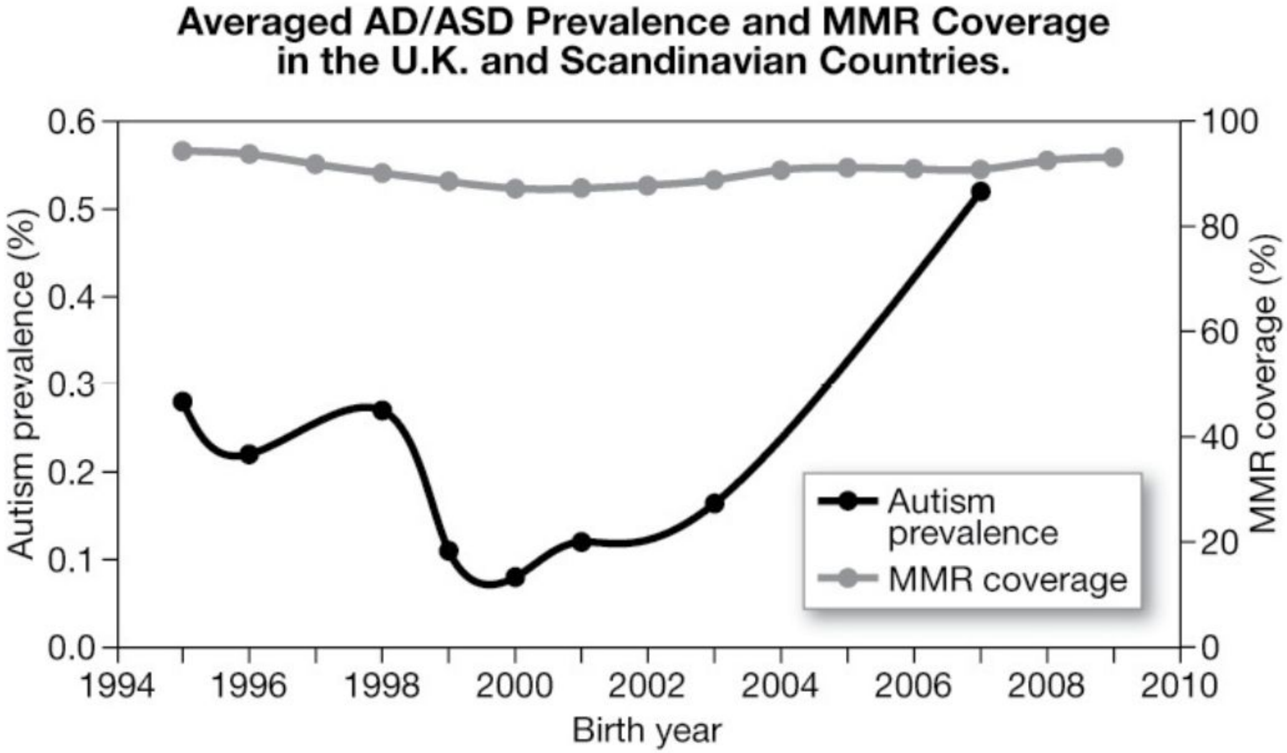


- A research paper in a less well-known journal tried to bring back the already disproven idea that the MMR vaccine is connected to autism.

- It looks like autism rates seem to track vaccination rates closely. But look at the axes. Autism prevalence is plotted from 0 to 0.6 percent. MMR coverage is plotted from 86 percent to 95 percent.

- During this time, autism rates increased significantly—about ten times from 2000 to 2007—while MMR vaccination rates changed very little. This difference becomes obvious if we adjust the graph. We don’t need to use the same scale for both trends, but we should make sure that both axes start at zero.

- This is how the corrected graph would look like. Notice the significant difference between the misleading graph and this one!


Averaged AD/ASD Prevalence and MMR Coverage
in the U.K. and Scandinavian Countries.
Autism prevalence (%)
MMR coverage (%)
Birth year
Autism prevalence
MMR coverage