Chapter 7: Misleading with Charts

Get your facts first and then you can distort
them as much as you please
– Mark Twain

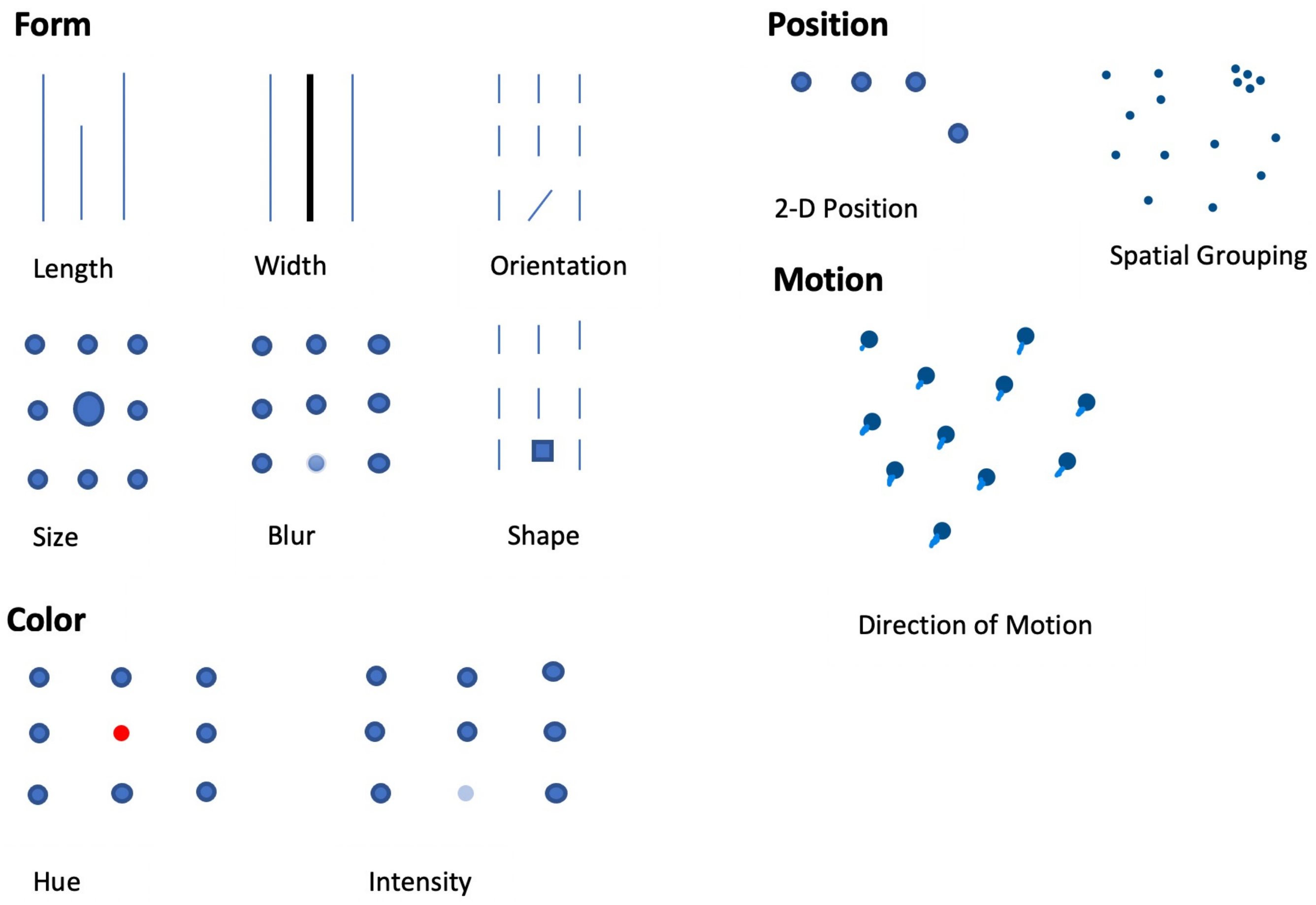
Learning Objectives

- Learn how chart can mislead us.
- Recognize conscious and unconscious brain decisions.
- Learn different ways through which we can convey incorrect information.
- Understand Simpson’s paradox and Drill down bias.

Pre-attentive Processing of Visual Attributes

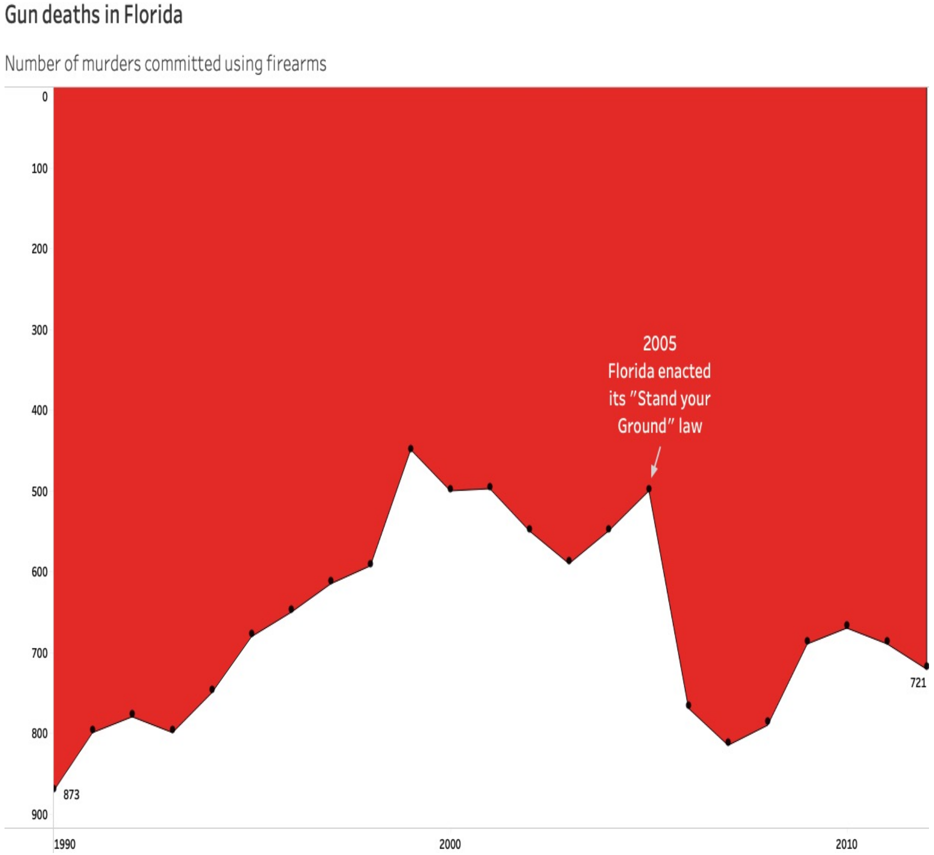
- Daniel Kahneman [Groenewegen, 2011], a psychologist and economist known for his research on the psychology of judgment and decision-making, explains that our brain has two operating systems. These systems are
  1. System 1 (Unconscious), and
  2. System 2 (Conscious).
- In his analysis of these two systems, he shows that System 2 is a slave to System 1.
- Kahneman's research revealed that our unconscious system is fast, automatic, effortless and accounts for 98% of all our thinking.
- When presented with a visual, our unconscious system quickly identifies the basic visual attributes.

Pre-attentive Processing of Visual Attributes



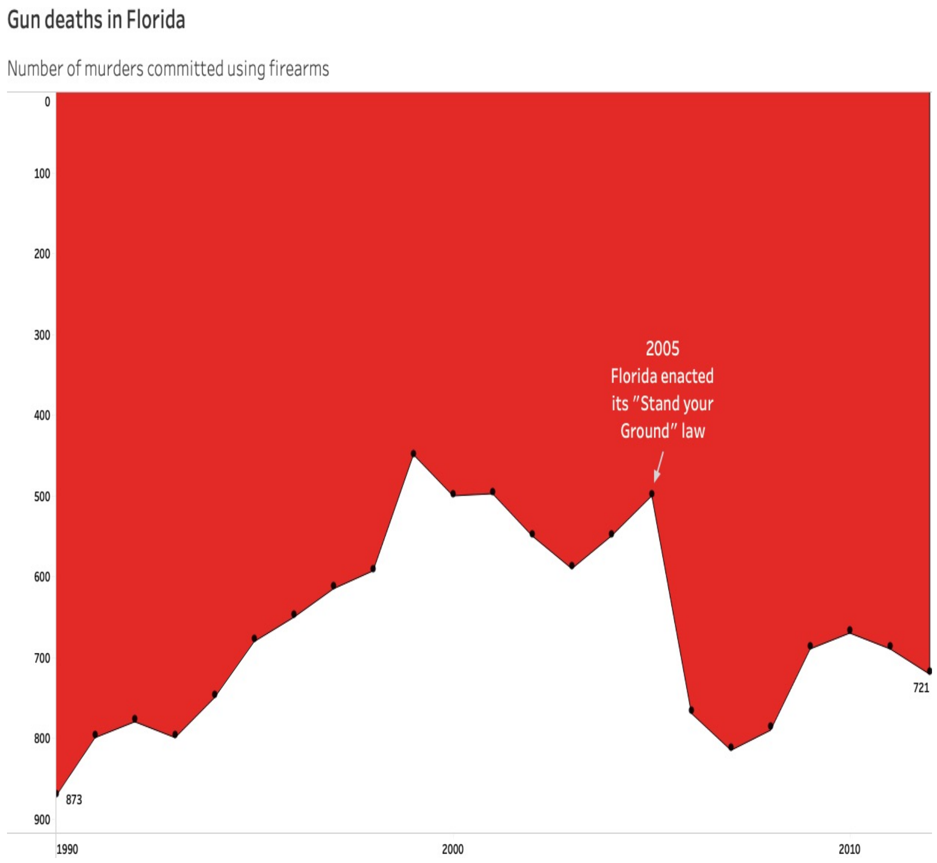
- For example, check Fig. 7.1. It is easy for our brain to distinguish the orientation of lines, colour, and position.
- On the contrary, our conscious system is slow, effortful, controlled and makes up for 2% of all our thinking.
- So, we need to actively search for a specific object within a cluster of things given a visualization.

Example of Misleading Graphs



- This chart explains how the “Stand your ground” law affected the number of deaths in the state of Florida.
- It looks as if gun deaths have dropped, after Florida adopted the law in 2005.
- When we examine the vertical axis closely, we can see that it is inverted.
- Therefore, the red area represents the deaths, which means deaths have increased after implementation of the law.

Example of Misleading Graphs

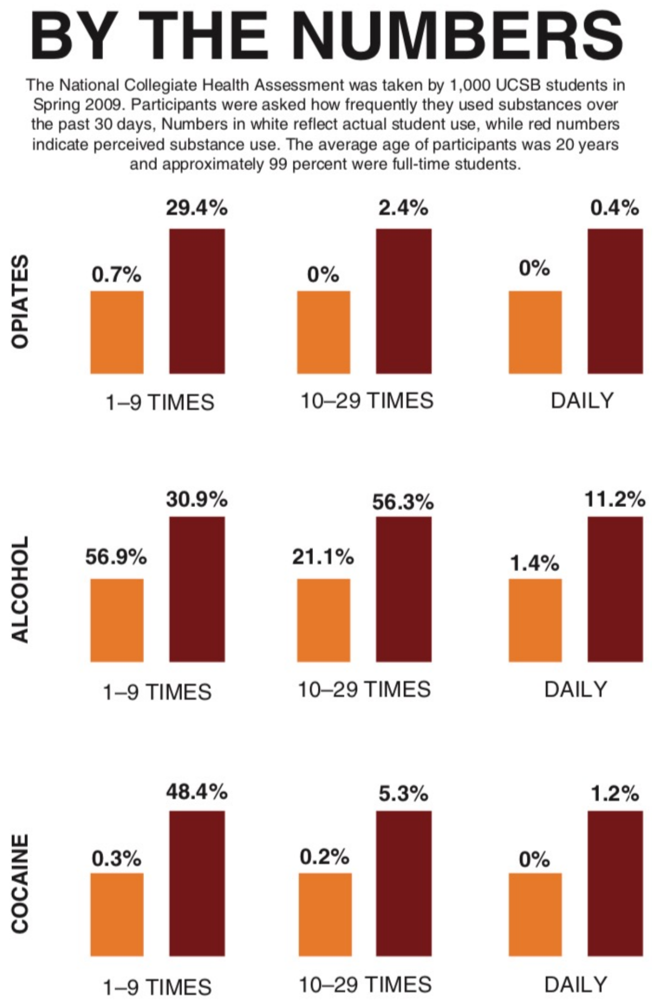


- Why did we get it wrong the first time?
- It is because the most generic convention is to read graphs from left to right and bottom to top.
- This chart goes against generic convention.
- In the case of the gun deaths of Florida chart, our unconscious brain assumed the vertical axis was non-inverted and a decreasing pattern by reading the graph from left to right.
- It then sent out those suggestions to the conscious system to become a belief.

Types of Misleading Charts

- A graph that is incorrectly represented is inefficient in communicating information or even misleading.
- It is possible to make errors during multiple stages of analysis, which can then lead to **misleading graphs**.

Visualization for Decoration



BY THE NUMBERS

The National Collegiate Health Assessment was taken by 1,000 UCSB students in Spring 2009. Participants were asked how frequently they used substances over the past 30 days, Numbers in white reflect actual student use, while red numbers indicate perceived substance use. The average age of participants was 20 years and approximately 99 percent were full-time students.

Figure 7.3 UCSB health assessment.
Source: https://dailynexus.com/2010-05-10/influence-alcohol-drugs-feature/

- Figure 7.3 shows a chart about drug use among students at the University of Santa Barbara.
- We can clearly see that the heights of the bars are not dependent on data, as irrespective of 0% or 21.1%, the height of the bar is same.
- These also known as **Visualization non sequitur.**

Visualization for Decoration

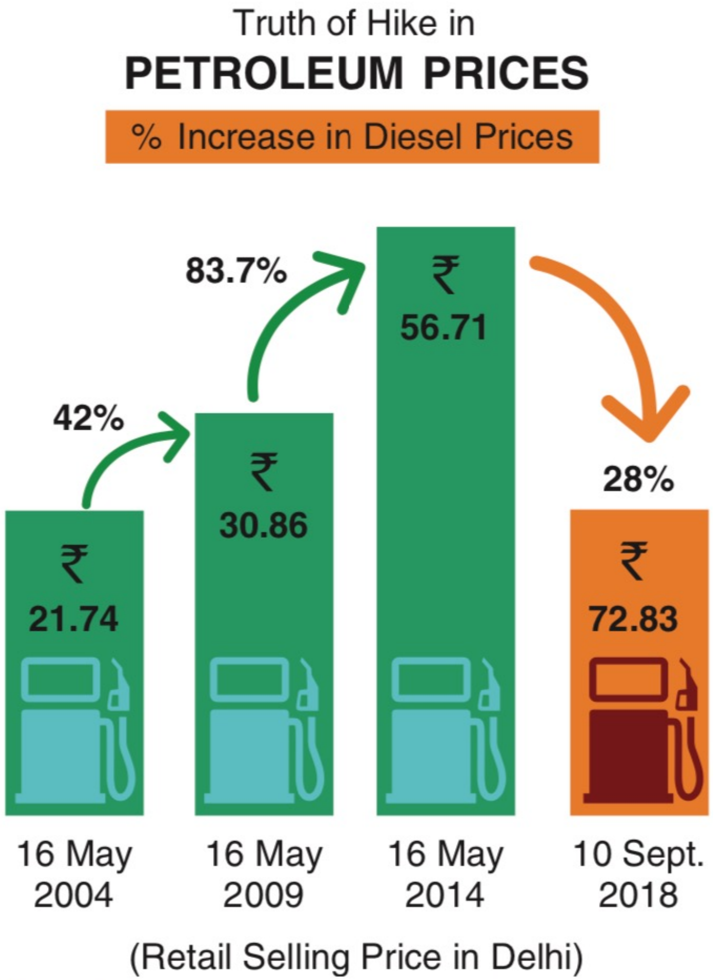


Figure 7.4 Increase in fuel price.

- The bar representing Rs72.83 is shown to be smaller than Rs56.71.
- However, upon closer inspection of the title, it appears the length of the bars represents a percentage increase in fuel price rather than the actual fuel price.
- Since the designer chose to add actual prices to the bar, it is just bad design, which can mislead the audience into wrong interpretation.

Axis Manipulation

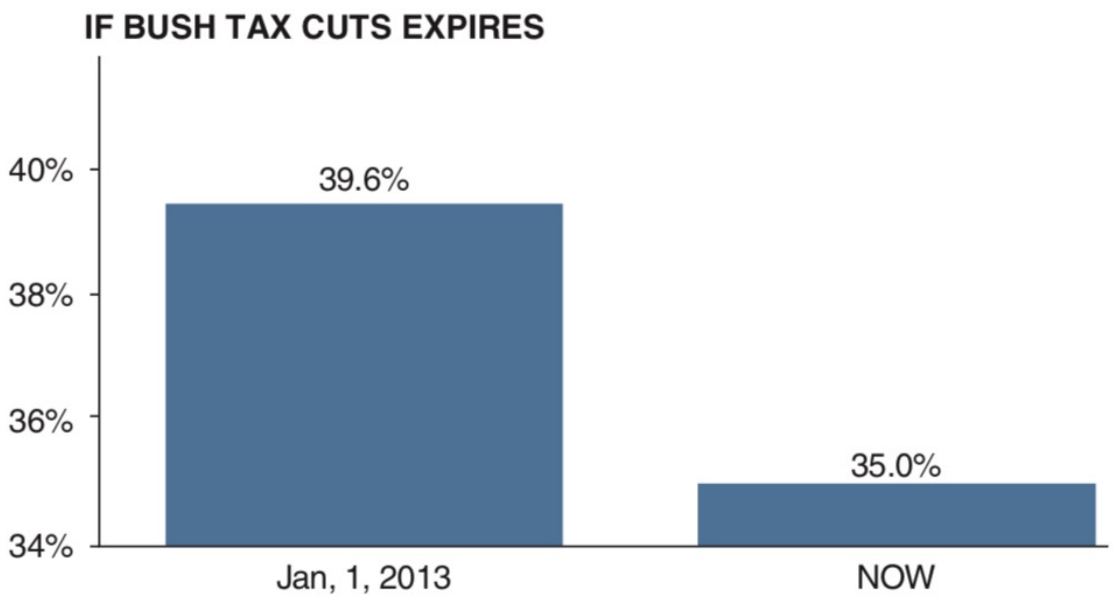


Figure 7.6 Tax rate.
Source: https://flowingdata.com/2012/08/06/fox-news-continues-charting-excellence/

- Fig. 7.6 compares how the top tax rate would fare with tax cuts and post expiry of tax cuts.
- On quick inference, it looks like the top tax rate bar post tax cut expiry is around four times higher than the bar with tax cuts.
- On another glance at the chart, we notice that the vertical axis is not starting at zero.
- Let us see if this causes any effect.

Axis Manipulation

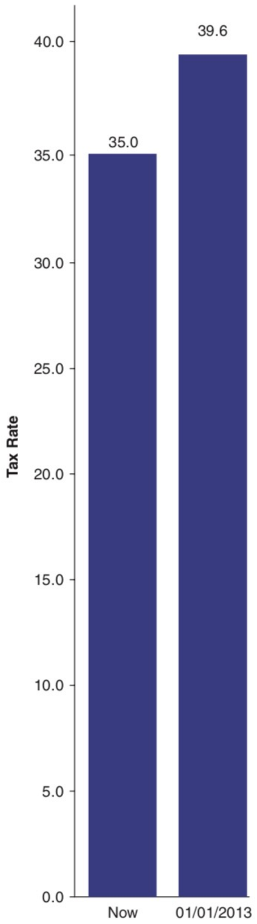


Figure 7.7 *Y*-axis from 0.

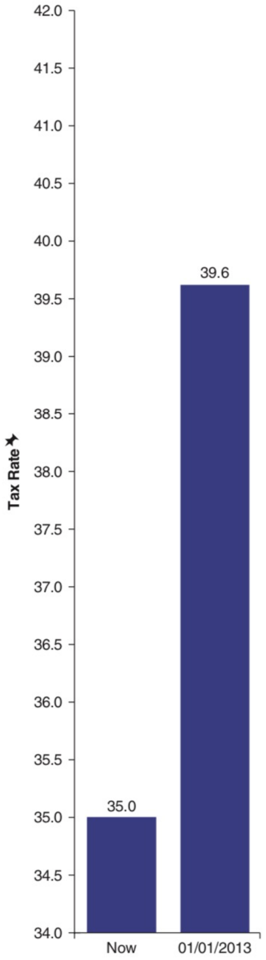


Figure 7.8 Truncated *Y*-axis.

- In Fig. 7.7, the vertical is starting from 0 and in Fig. 7.8, it is truncated.
- As we can see, the effect is less dramatic when we start the vertical from 0.
- So, the relevant effect size can be exaggerated or ignored by the user’s choice of axis.
- This can cause a misleading graph, and it is common practice that the difference in the size of the bars is considered to be directly proportional to the values.

Axis Manipulation

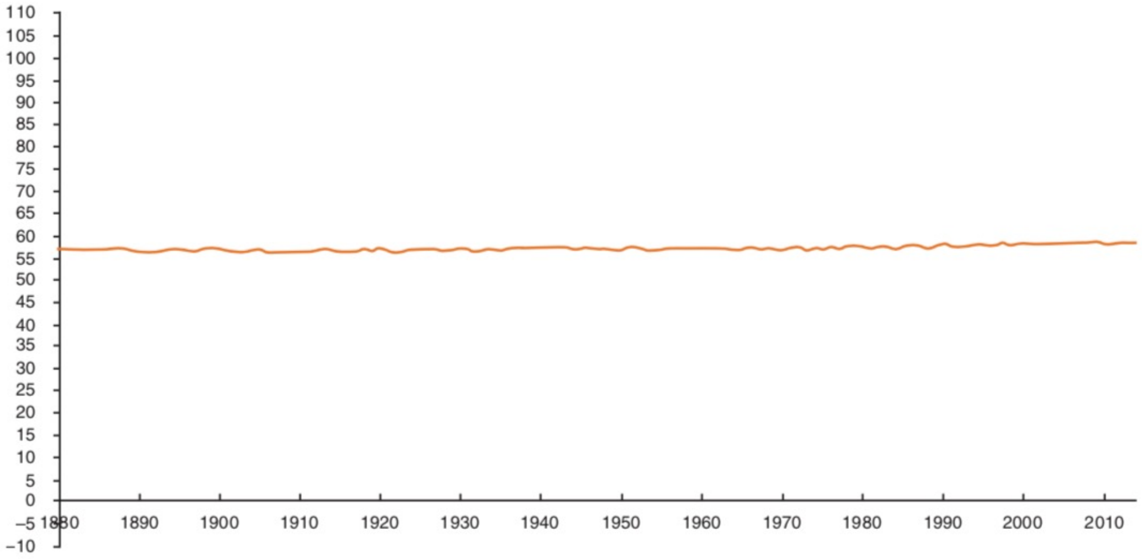


Figure 7.11 Average annual global temperature in Fahrenheit between 1880–2015.
Source: https://www.therightinsight.org/Ignorance-or-Duplicity-Average-Global-Temperature

- Here is a graph representing the average global temperature in Fahrenheit.
- It has been created to convey a message on climate change.
- Here, the Y-axis starts at 0, but this graph is still misleading.
- It is because the y-axis represents the average temperature in Fahrenheit, which cannot be 0 or 110.
- Also, this chart does not help us understand the trend.

Axis Manipulation

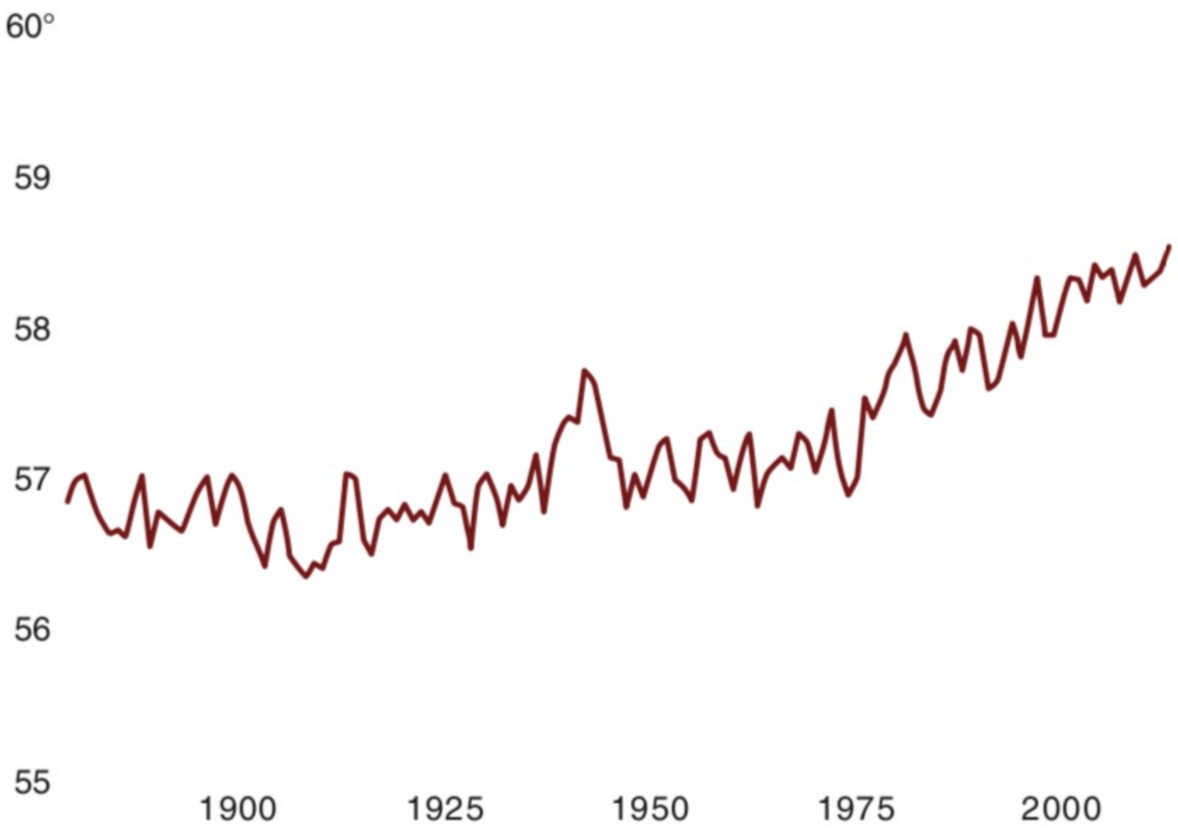
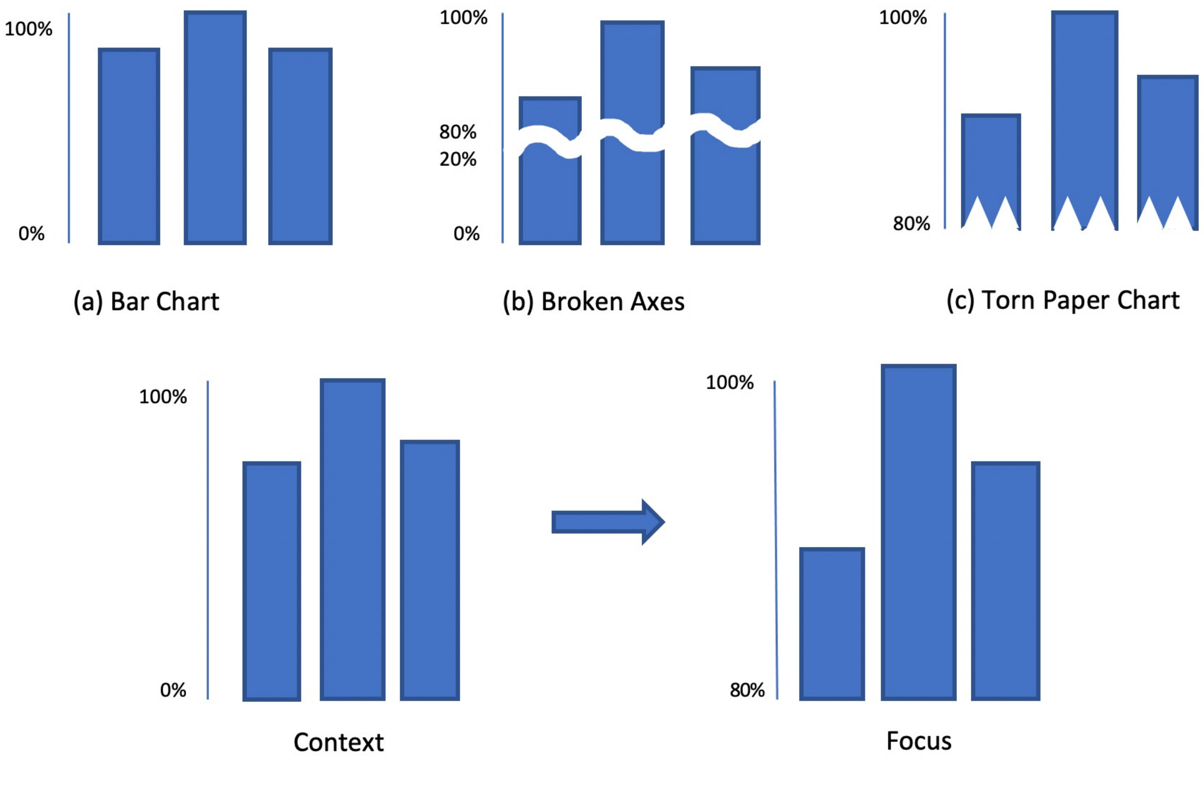


Figure 7.12 Average annual global temperature by year.

- When we are discussing temperature change, even a slight change in temperature can have a huge impact.
- Here, by starting the axis from 0, the effect of a small change in temperature is totally invisible in the chart.
- Refer to Fig. 7.12 for more an inferable representation of the chart.

Types to indicate Truncated y-axis



(a) Bar Chart

(b) Broken Axes

(c) Torn Paper Chart

(d) Interactive Focus+Context

- Figure 7.14 shows a study done by Michael Correll, Enrico Bertini, and Steven Franconeri [Correll, 2019] in a white paper “Truncating the vertical axis: Threat or Menace?[1]” which provides a few proposed ways of indicating the users of the truncated y-axis in a chart.

[1] Source:https://arxiv.org/abs/1907.02035

Axis Manipulation

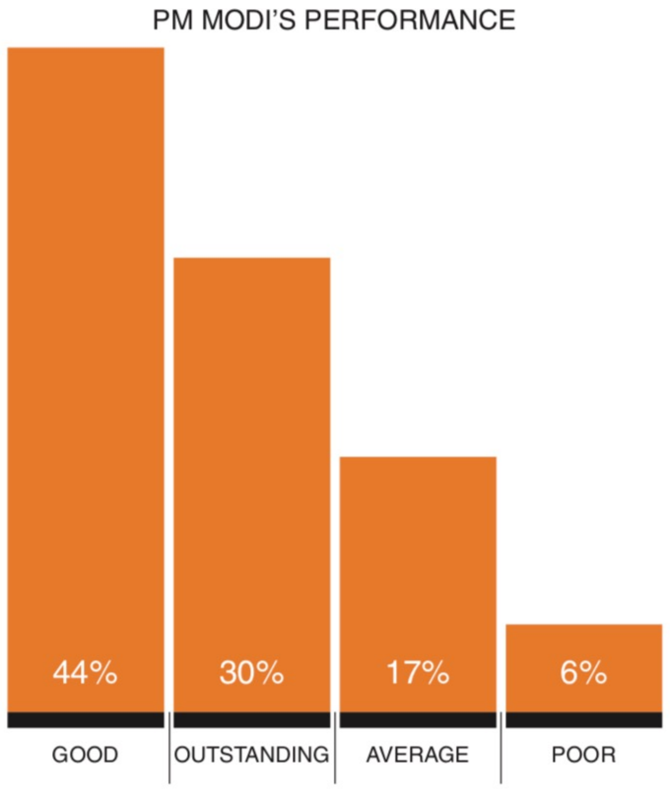


Figure 7.14 Types to indicate manipulated *x*-axis.
Source: www.indiatoday.in/mood-of-the-nation/story/74-of-india-still-with-namo-hindutva-agenda-finds-more-takers-motn-poll-1761472-2021-01-21

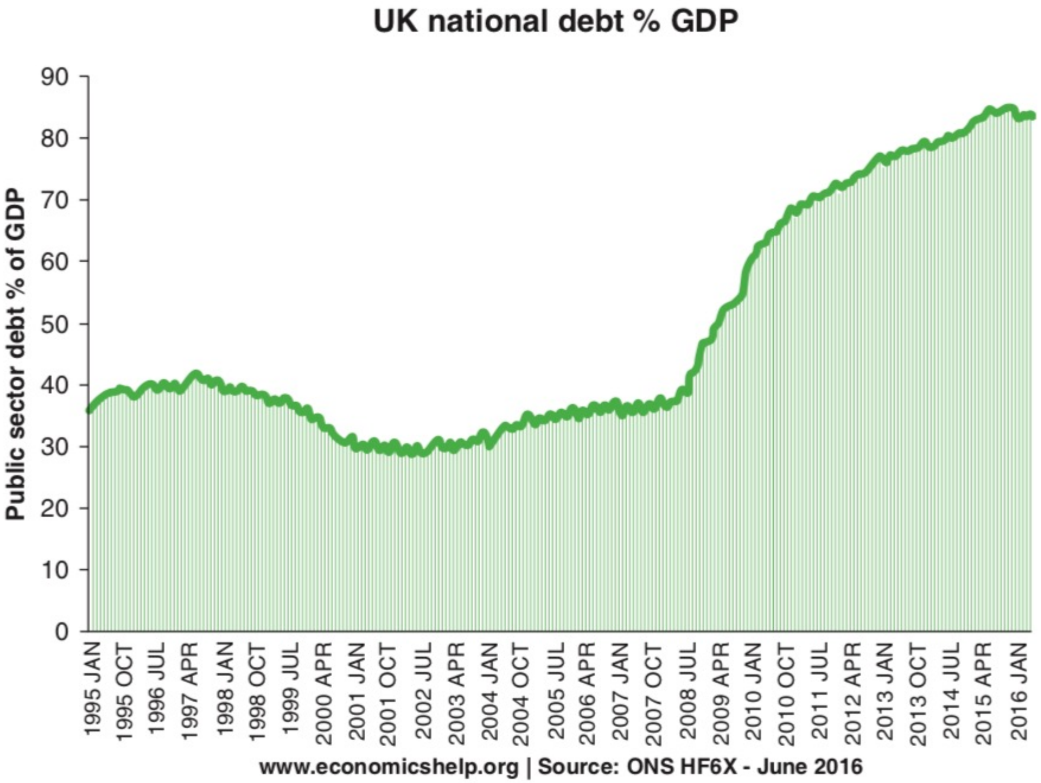
- Figure 7.14 shows the people of India how Prime Minister Modi has performed in his tenure so far.
- Do you see anything wrong with this graph?
- If you look closely, you will see that in this graph, you have been misled by horizontal axis manipulation.
- At first glance, you tend to think that the horizontal axis is in descending order i.e., outstanding, then good, then average, and poor.
- But 'out- standing' and 'good' have been interchanged and hence it becomes a common mistake by the audience to think PM Modi has been voted as outstanding by 44%.

Cherry Picking Data

- Designers often cherry pick data while building visualizations to steer the audience to their viewpoint.
- Including only a portion of the data or omitting certain parts can skew the visualization and blend it into the narrative.
- The audience may be misled as a result.
- We can also select a time range carefully to:

  1. Exclude major events that will impact the narrative.
  2. Pick only the data points that aid the narrative, thus hiding the important changes in between.

Cherry Picking Data



www.economicshelp.org | Source: ONS HF6X - June 2016

Figure 7.15 UK debt between 1995 and 2016.

- Figure 7.15 shows the UK debt crisis.
- According to this chart, it appears that the UK National debt is higher in 2016 and has been increasing over the years.
- This would help us if our narrative were to justify a policy to lower debt.
- The chart displays only a partial picture depending on the narrative.

Cherry Picking Data

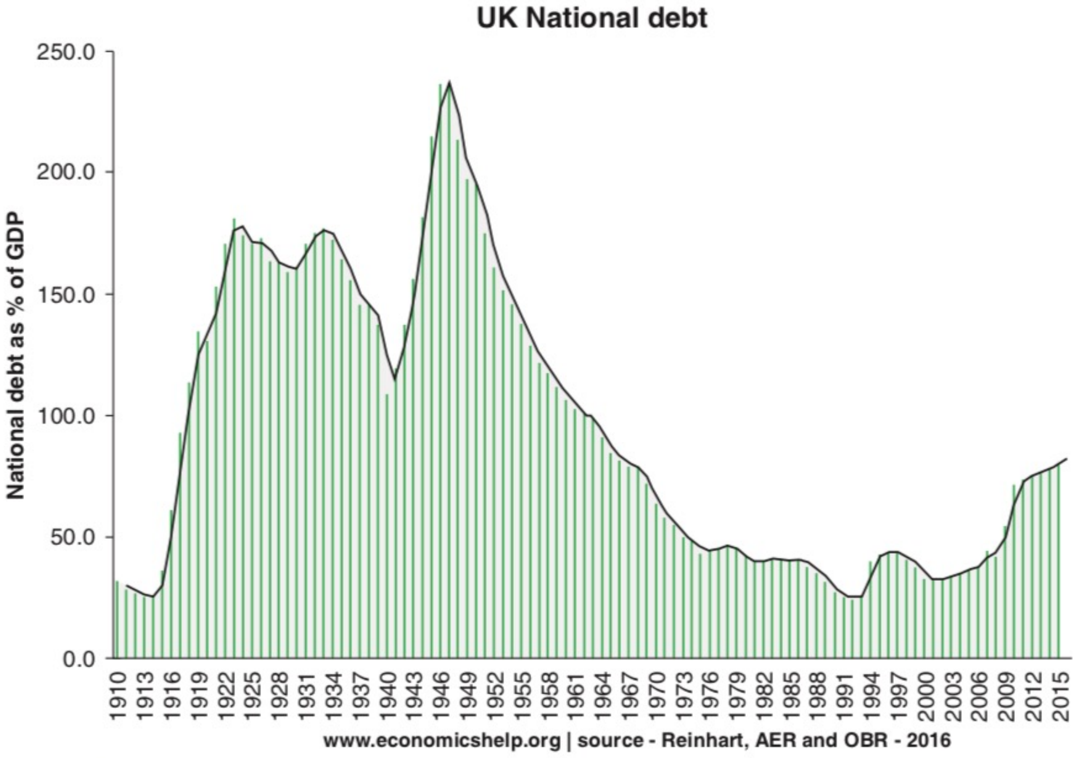


Figure 7.16 Time series on the UK debt.

- Because when we examine the full time series data (Refer to Fig. 7.16), debt is actually low in comparison with historical data.

Cherry Picking Data

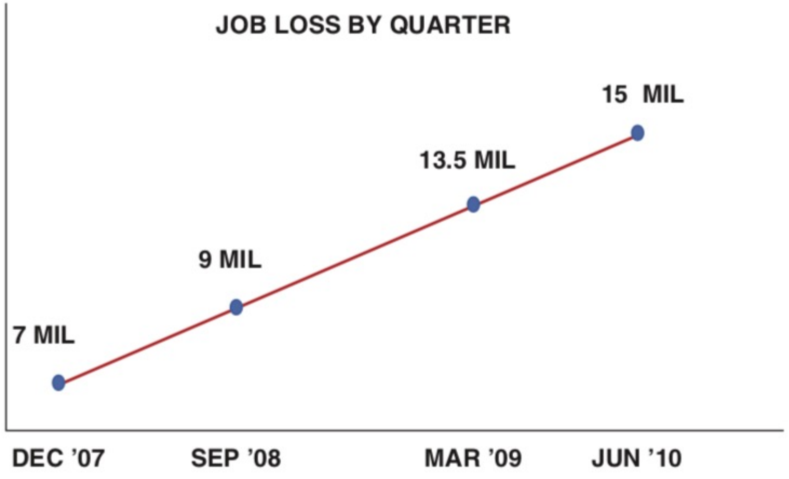


Figure 7.17 Cherry picked interval data.
Source: https://twitter.com/rvawonk/status/766224615731494912

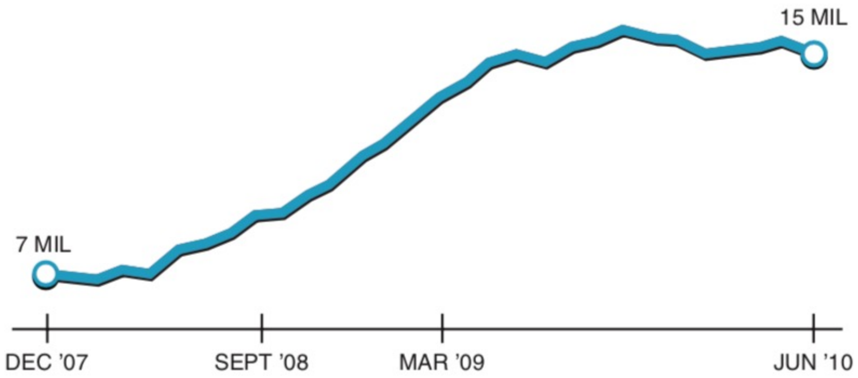


Figure 7.18 Job Loss by quarter.

- Look at a chart used by Fox News (Refer to Fig. 7.17).
- From this chart, we can infer that there is an increasing trend for job losses.
- As we look for more information from the chart, we see that there is a mismatch in the horizontal axis intervals.
- By correcting the horizontal axis intervals and including all valid data between these intervals it appears that job loss has plateaued since March 2009.

Pie Chart Blunders

- Pie charts are useful for representing composition.
- A slice of a pie represents the percentage of composition represented by a particular category.
- This is typically used when there are fewer categories.
- Since it represents the composition of a category, it is expected that adding up each slice of the pie will make up the whole pie (100%).
- Pie charts can be misused by designers when the categories are not mutually exclusive

Pie Chart Blunders

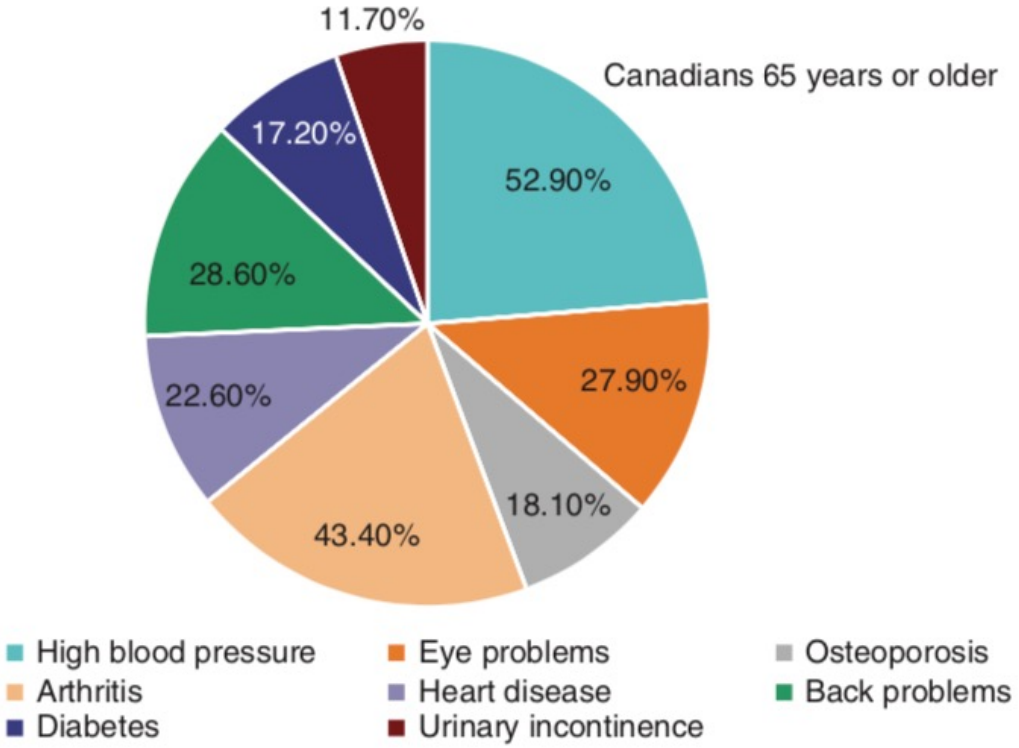


Figure 7.19 Pie chart.

- In this case, the numbers do not add up to 100%.
- With pie chart, we expect parts of a whole or composition, this chart is clearly misleading.
- Such data is best represented as a bar chart rather than a pie chart.

Pie Chart Blunders

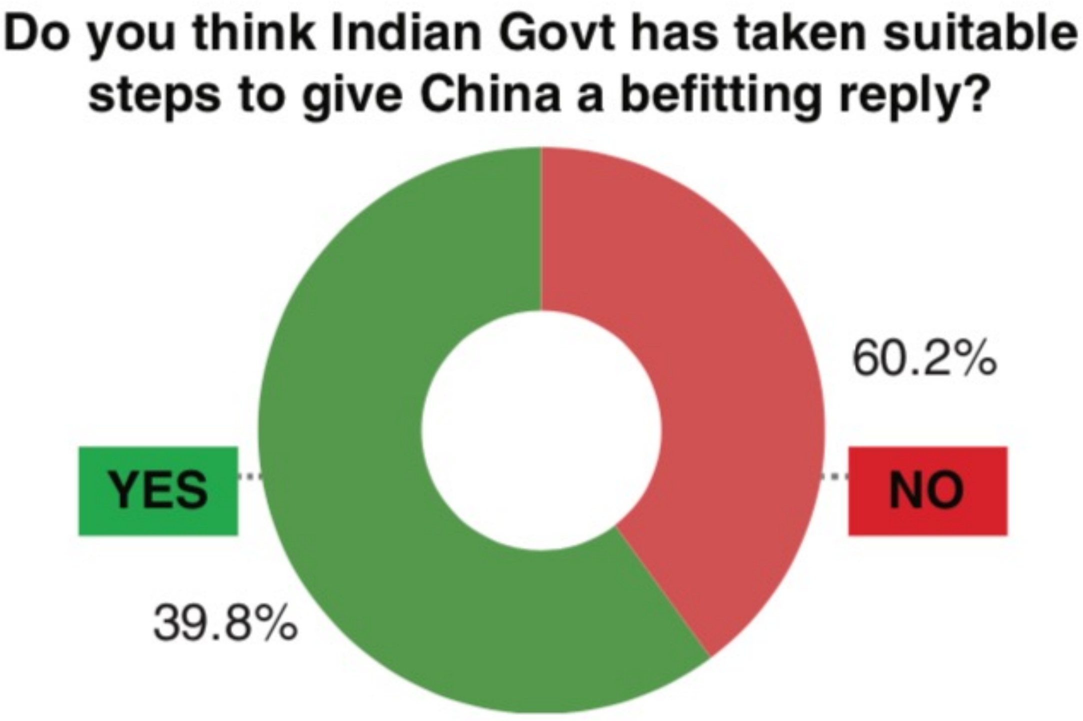


Figure 7.20 Misleading Pie chart.

Source: https://www.freepressjournal.in/india/when-398-was-greater-than-602-netizens-troll-times-now-for-pie-chart-gaffe

- Figure 7.20 is another example of how a designer can misuse the pie chart to mislead the audience.
- According to this chart which displays survey responses, 39.8% is a larger slice than 60.2%.
- At quick glance, our unconscious brain makes us believe that 39.8% is greater than 60.2%.

Misuse of Correlation and Causation

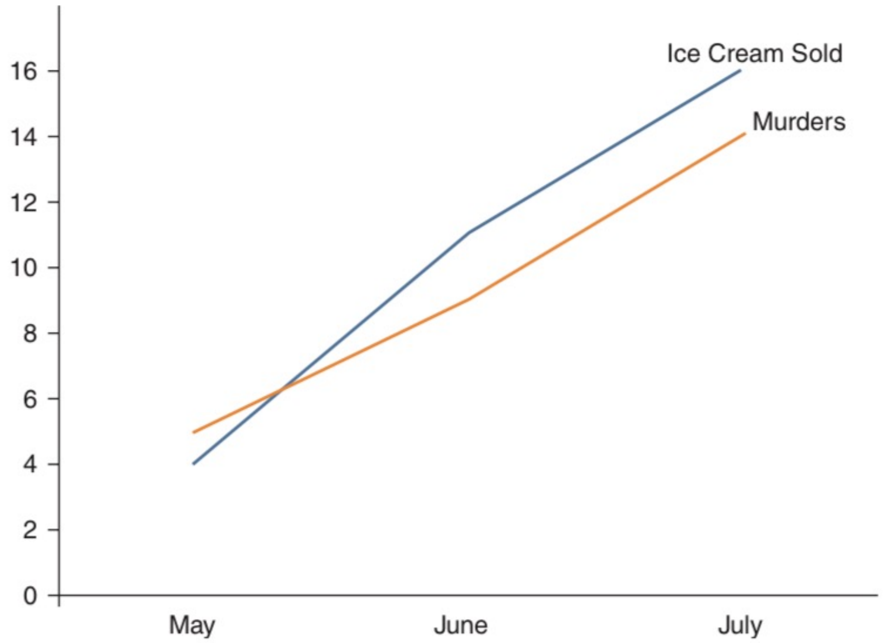


Figure 7.21 Correlation between consumption of ice cream and number of murders.
Source: https://slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html

- According to a statistical study, a positive correlation exists between the number of murders in New York and the consumption of ice cream (pints per person) (Refer to Fig. 7.21).
- The number of murders in- creases as ice cream sales per person go up.
- Does this mean your next cone of ice cream may lead to a murder?
- It is important for us to understand the difference between correlation and causation.

Misuse of Correlation and Causation

- **Correlation** and **causation** are two of the most misunderstood concepts in statistics.
- Correlation implies a linear relationship between two numerical variables.
- The important thing to note here is that a change in the value of one variable does not necessarily cause the change in value of the other variable only that it does exist.
- To understand how this relationship between two correlated variables exists and why it exists, we need to look for the causation.
- Causation states that the change in value of one variable will cause a change in value of another variable.

Misuse of Correlation and Causation

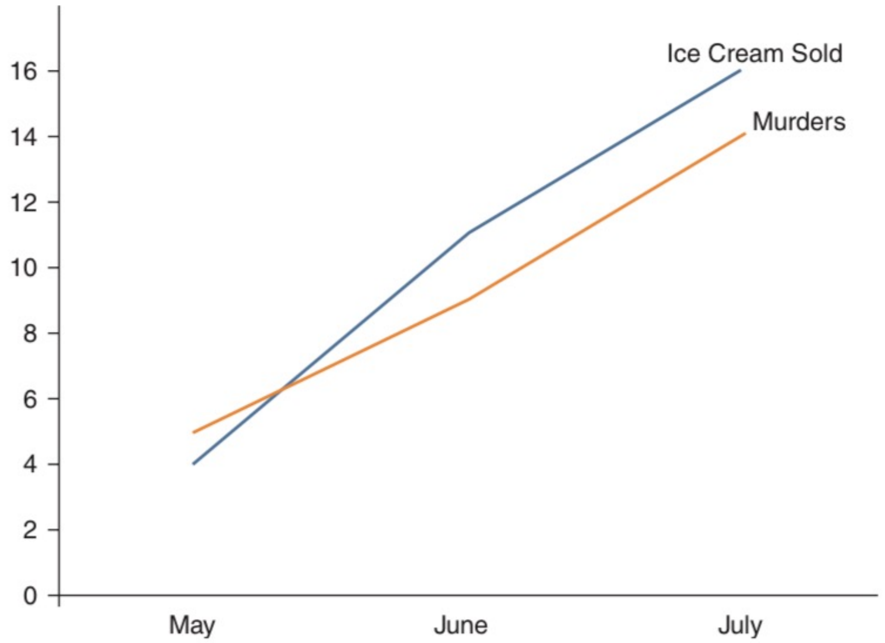


Figure 7.21 Correlation between consumption of ice cream and number of murders.
Source: https://slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html

- In this example, we cannot infer that an increased sale of ice cream causes an increased number of murders, since these two variables are correlated.
- A study explained that the consumption of ice cream and crime rates are positively correlated because both the variables are positively correlated with temperature.
- Weather is the underlying factor causing both the events.

Misuse of Correlation and Causation

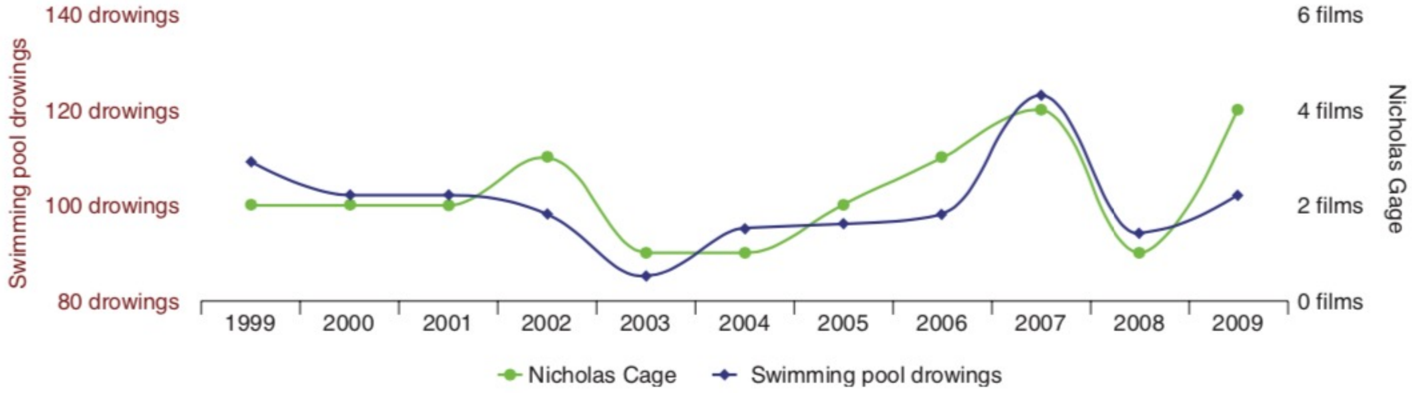


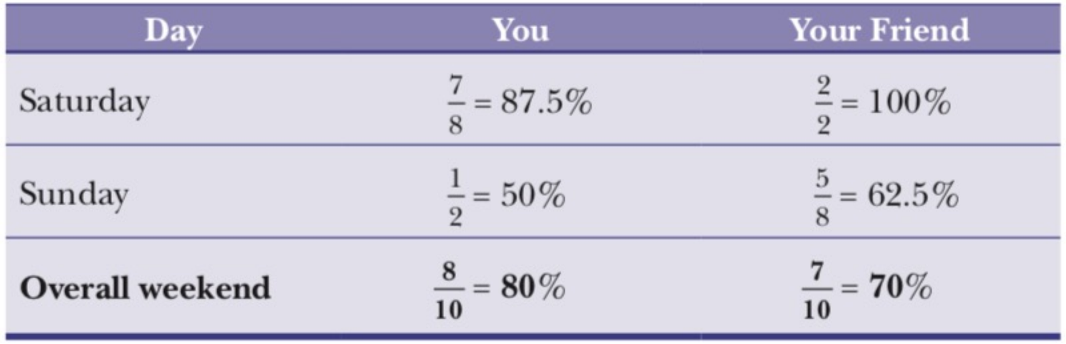
Figure 7.22 Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in.

Source: www.fastcompany.com/3030529/hilarious-graphs-prove-that-correlation-isnt-causation

- Hence, just because two variables seem to show correlation, it does not mean they are meaningfully related to one another.
- Refer to Fig. 7.22, the number of movies Nicolas Cage appeared in is spuriously correlated with the number of people drowned by falling in a swimming pool.

Simpson’s Paradox

Table 7.1 Simpson’s paradox



Day	You	Your Friend
Saturday	$\frac{7}{8} = 87.5\%$	$\frac{2}{2} = 100\%$
Sunday	$\frac{1}{2} = 50\%$	$\frac{5}{8} = 62.5\%$
Overall weekend	$\frac{8}{10} = 80\%$	$\frac{7}{10} = 70\%$

- Simpson’s paradox[2] is a phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.
- Let us consider that you and your friend took up the challenge of solving math problems for two days over a weekend.
- On each of the two days, you answered a lower proportion of questions correctly than your friend.
- But when we look at overall data for both days together, you have higher proportion of correct answers.

[2] Source:https://www.britannica.com/topic/Simpsons-paradox

Simpson’s Paradox

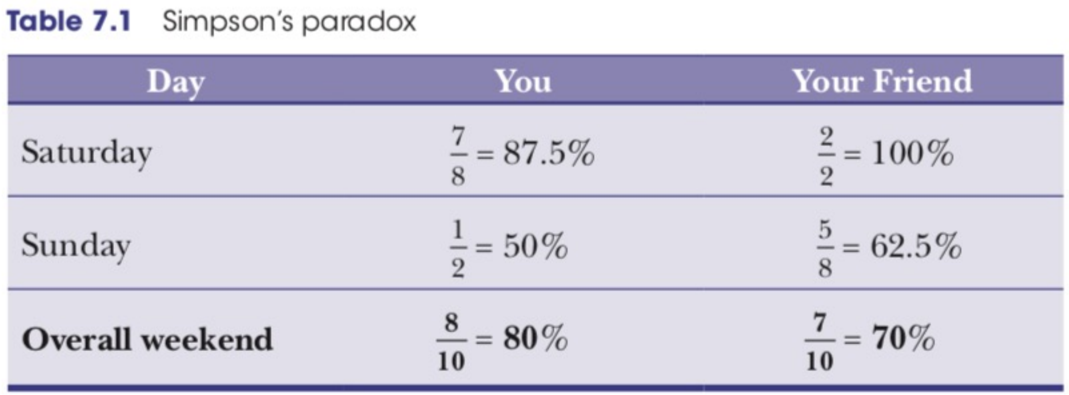


Table 7.1 Simpson’s paradox

Day	You	Your Friend
Saturday	$\frac{7}{8} = 87.5\%$	$\frac{2}{2} = 100\%$
Sunday	$\frac{1}{2} = 50\%$	$\frac{5}{8} = 62.5\%$
Overall weekend	$\frac{8}{10} = 80\%$	$\frac{7}{10} = 70\%$

- So, why does it occur?
- We assume that winning in all groups individually implies winning overall.
- However, we can arrive at this conclusion only if the group sizes are equal.
- On the other hand, if the group sizes are different, then a few groups might be dominating the total number for each side, but these groups belong to different categories.

Scaling

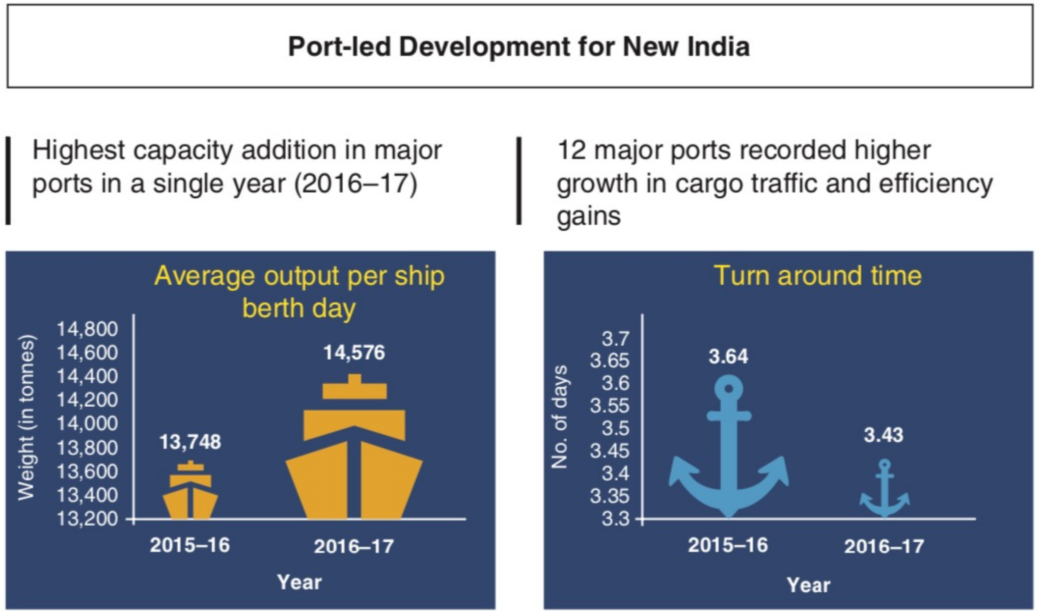


Figure 7.27 Pictorial graph with misleading scales.
Source: RBI Annual Report (2016–17).

- A graph can be altered by changing its scale.
- Fig. 7.27 is a picture chart depicting port-led development done by the Government of India.
- The average output per shipment berth has increased around 6% from 2015–16 to 2016–17.
- But the ship image for 2016–17 looks around 30 times bigger.
- Check the anchor image, even though actual decrease in turnaround time between 2015–16 to 2016–17 is around 5%, the image looks around 30 times smaller.

Scaling

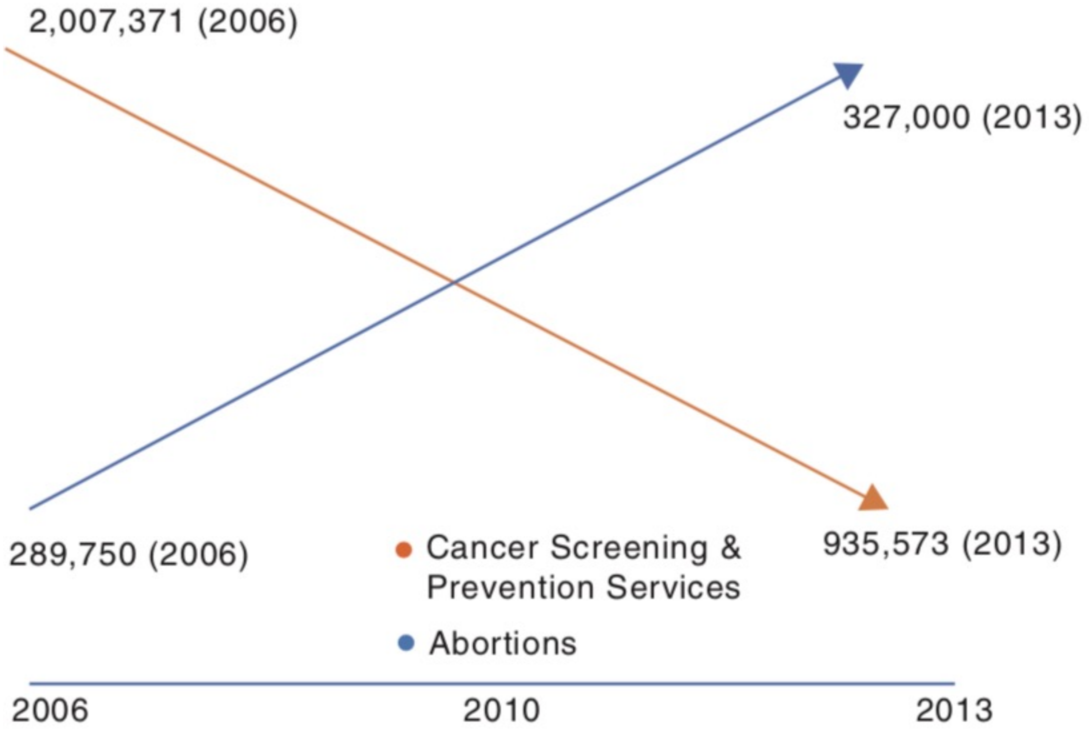


Figure 7.28 Dual axes chart with different scales.

Source: Americans United for life

- Another example of a misleading chart due to scaling is dual axis charts with different scales.
- Figure 7.28 was projected by Republican senator Jason Chaffetz during a high-profile congressional hearing investigating planned parenthood.
- This data indicates that in 2006, Planned Parenthood performed more preventive services and cancer screenings than abortion, while in 2013, abortion was more common.

Scaling

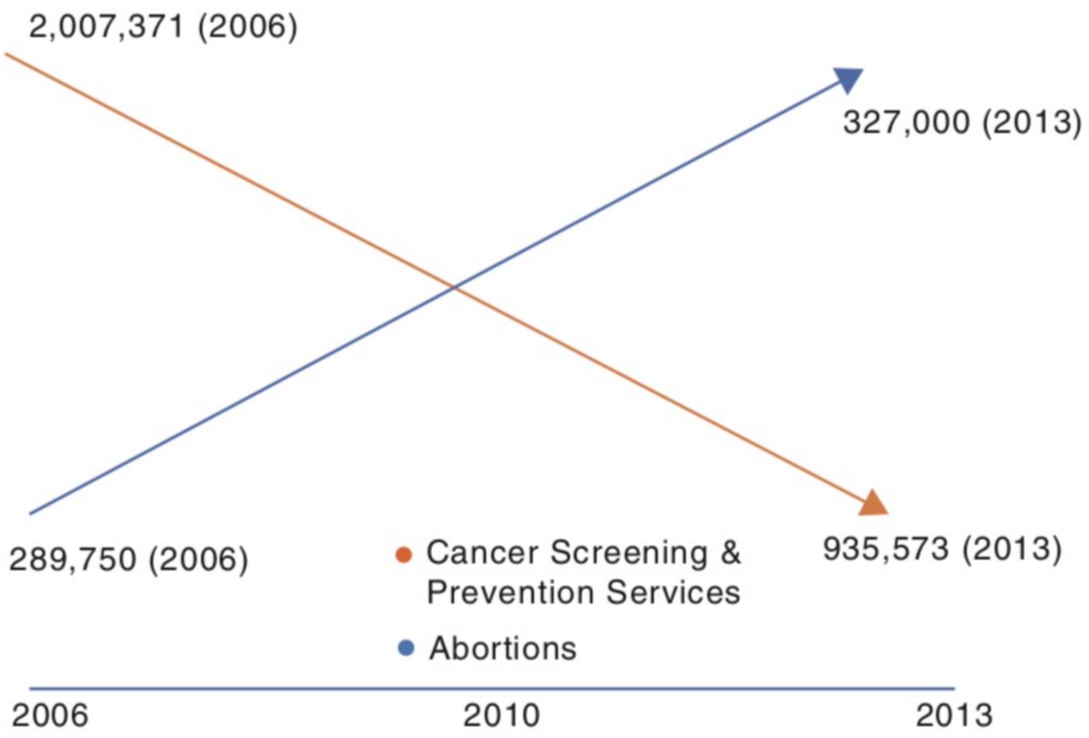


Figure 7.28 Dual axes chart with different scales.
Source: Americans United for life

- Fig. 7.28 is a dual axis chart.
- The vertical axis towards the left represents cancer screenings and prevention services, and the vertical axis towards the right represents abortions, which are two different scales.
- A graph of this kind is prone to displaying spurious correlations.
- With dual axes, it is easy to manipulate and exaggerate trend, as most people ignore the axis labels that put the numbers in context.

Scaling

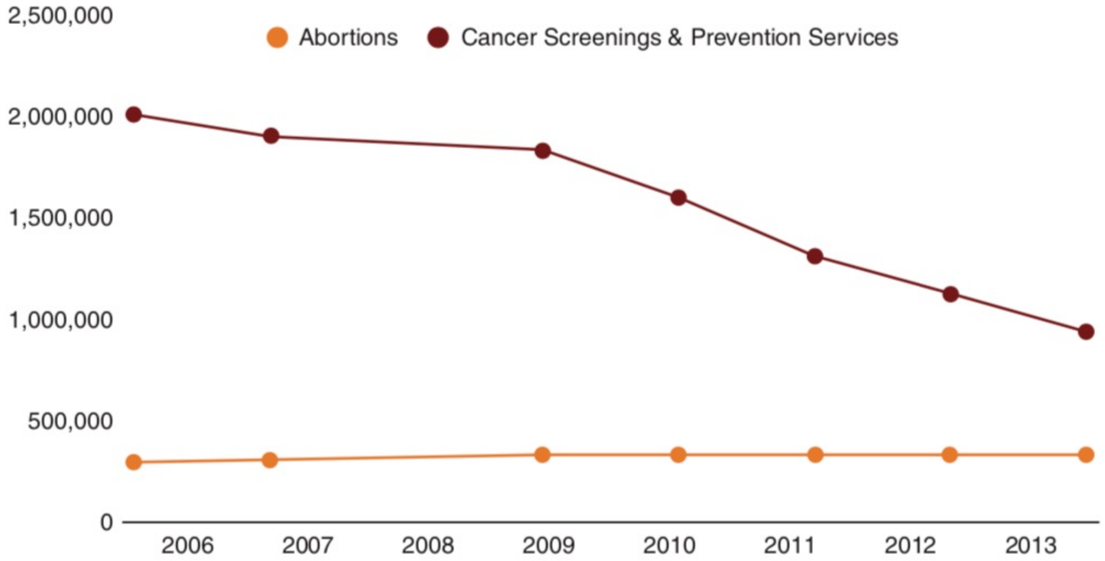


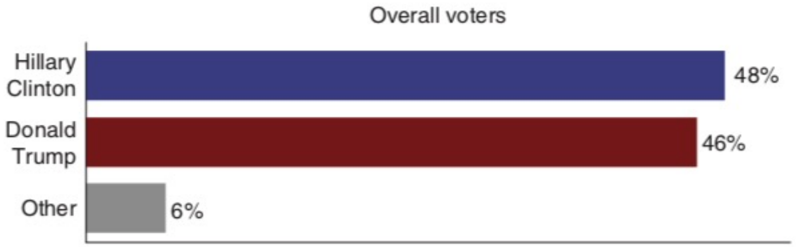
Figure 7.29 Line chart with trends.

- Check Fig. 7.29 for the trends on the same data when plotted as a simple line chart.
- A dual axis chart with two different variables with different scaled axes is misleading, irrespective of the variables in the chart.
- Such charts bring in a fallacy of variables being correlated.

Drill Down Bias

- Whenever we are provided with a complex dataset, we start with exploratory analysis through visualization to understand it.
- We usually begin by drilling down into multiple dimensions.
- But at times, the order in which we filter determines what kind of insight we uncover.
- Confounding factors in the data, coupled with an improper drill down path, can lead us into misleading insights.
- Let us understand this with an example [DrillDownData.xlsx].

Drill Down Bias



Inference: Overall, Hillary Clinton is favored by 48% voters

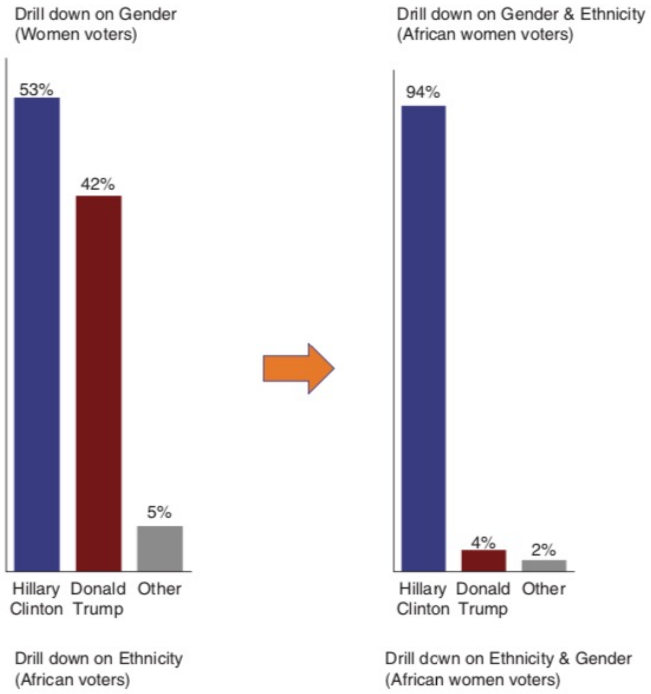
- Figure 7.30 shows a comparison of the 2016 US election results.
- In this case, insight is dependent on the order in which we apply the drill down technique to uncover insights in data.

Drill Down Bias

Drill down sequence

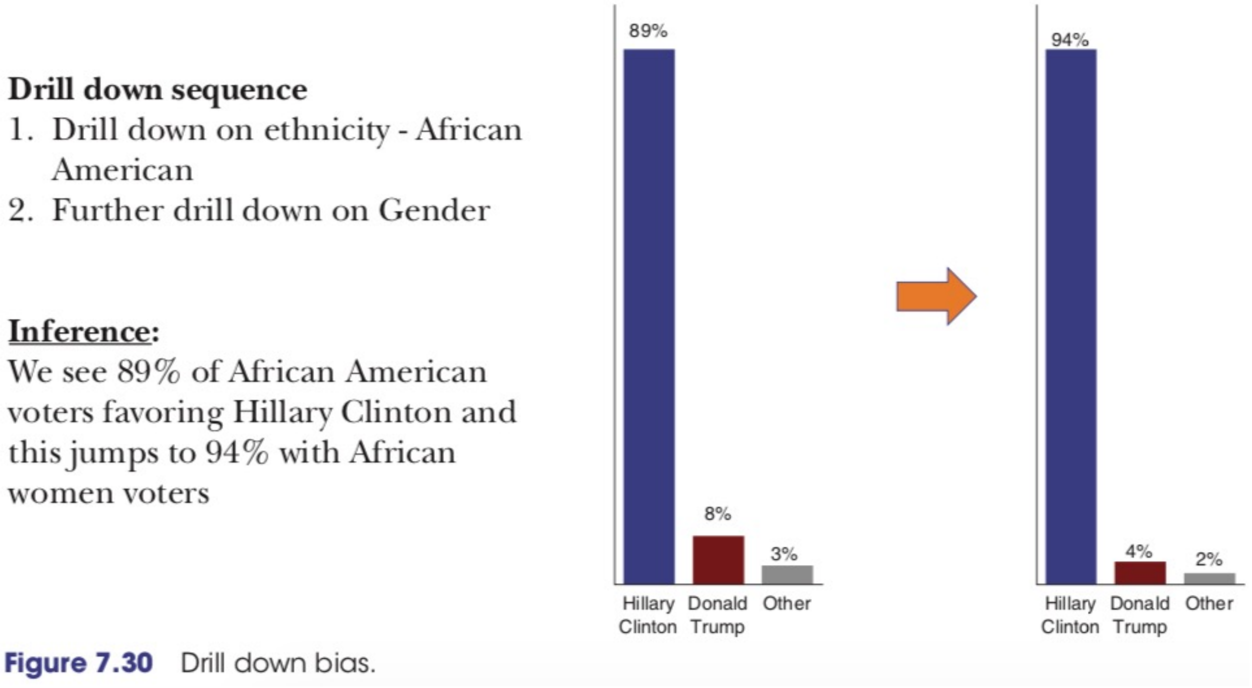
1. Drill down on Gender
2. Further drill down on Ethnicity - African American

<u>**Inference**</u>**:**
We see 53% of women favoring Hillary Clinton and it is 94% with African American women voters



- Based on the first drill down, it appears that gender and ethnicity together favour a candidate (Hillary Clinton in this case).

Drill Down Bias



Drill down sequence

1. Drill down on ethnicity - African American
2. Further drill down on Gender

<u>Inference</u>:

We see 89% of African American voters favoring Hillary Clinton and this jumps to 94% with African women voters

Figure 7.30 Drill down bias.

- If we drill down with ethnicity and then gender, we can see that ethnicity is driving this increase in the voter percent, rather than ethnicity and gender.

Drill Down Bias

- To avoid drill down bias, we should compare the distributions of parent and child, and see if the child distribution looks similar to the parent.
- Then, we can infer that the parent corresponding to the more general cause almost explains the specific case.
- Applying this to the example provided, we can see that when we filter on gender, the distribution of voters' percent, 53%, 42%, and 5% looks similar to its parent distribution of overall voters, 48%, 46%, and 6%.
- But the distribution when filtered on ethnicity, 89%, 8%, and 3% is dramatically different from the parent.

Data Discrepancy

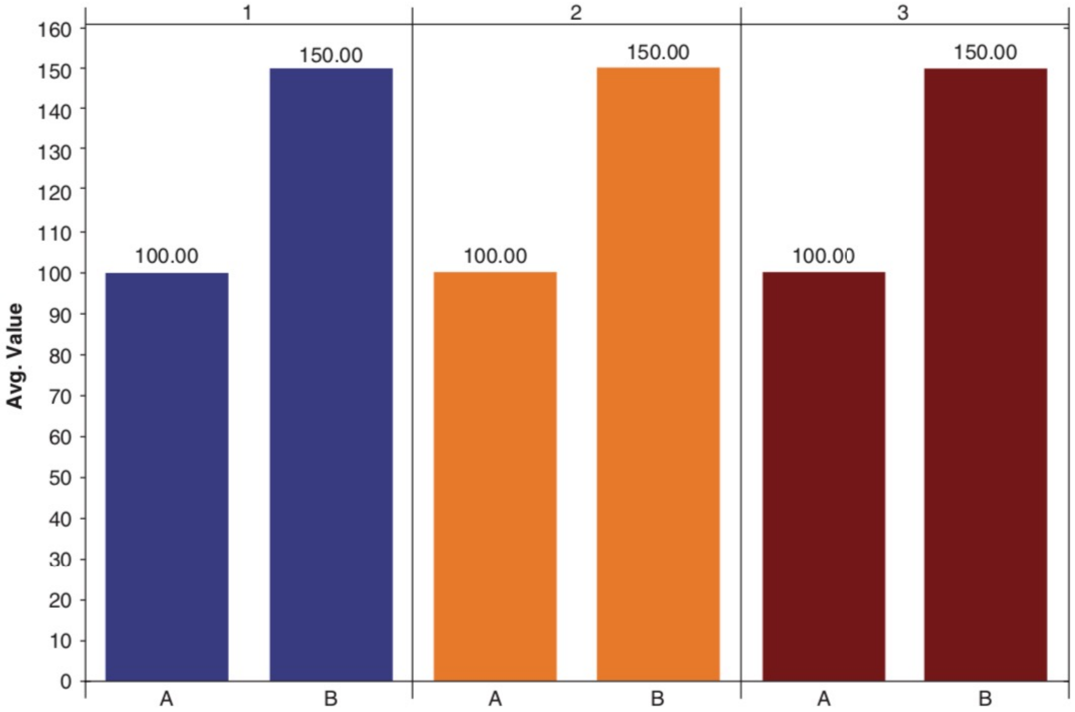


Figure 7.31 Aggregated comparison chart.

- Errors in the underlying raw data such as repeated values, missing values, outliers, or wrong data entries, if not treated correctly, will have an impact on visualizations.
- Let us assume we are comparing three states (States 1, 2, and 3) on two measures (Measures A and B).
- Figure 7.31 compares the average of measures A and B for all three states [DataDiscrepancy.xlsx].
- We can conclude that all three states are doing similarly when compared on average measures.

Data Discrepancy

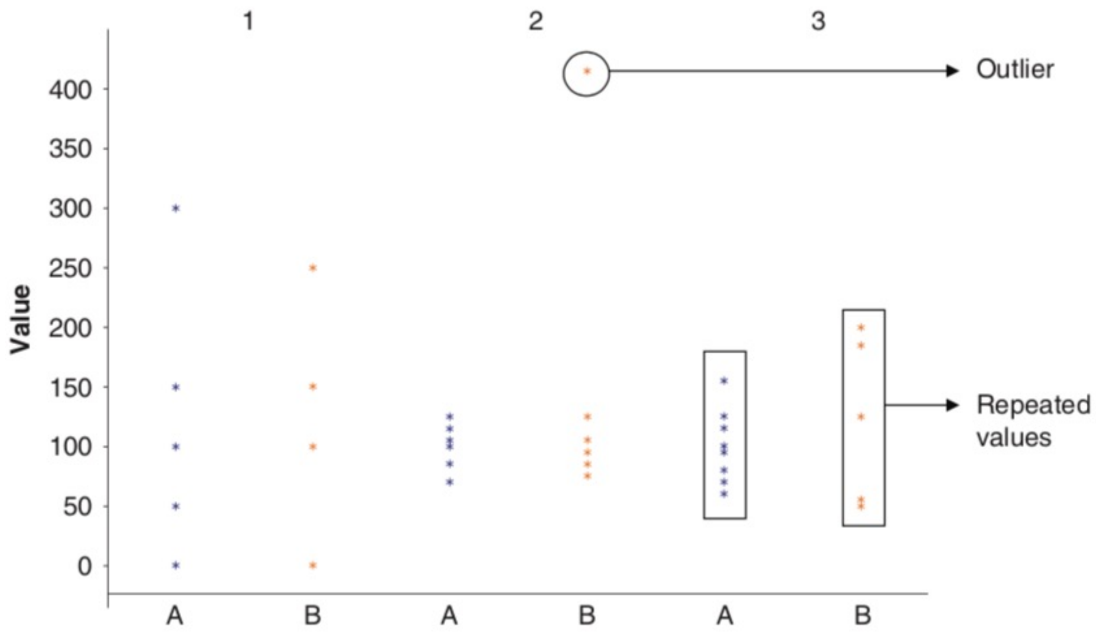


Figure 7.32 Disaggregated data comparison.

- But check Fig. 7.32, where we look at the underlying disaggregated data for these measures.
- We see that underlying **data distributions** are different. Our aggregated chart is skewed due to
  1. An outlier in State: 2 and Measure: B
  2. Repeated values in State: 3 and Measure: A and B
- In case where a decision-maker needs to understand these differences, just presenting aggregated chart, masks the relevant information from our audience.

Data Discrepancy

- Apart from the above example, there may be other data discrepancies such as:
  1. **Spelling mistakes in the underlying data.**
  2. **Sampling errors can lead to misleading charts, when unnoticed and aggregated.**
  3. **Imputed values for missing data, if not meticulous, can create spurious trends or relationships on visualizations.**

Histogram Confuser

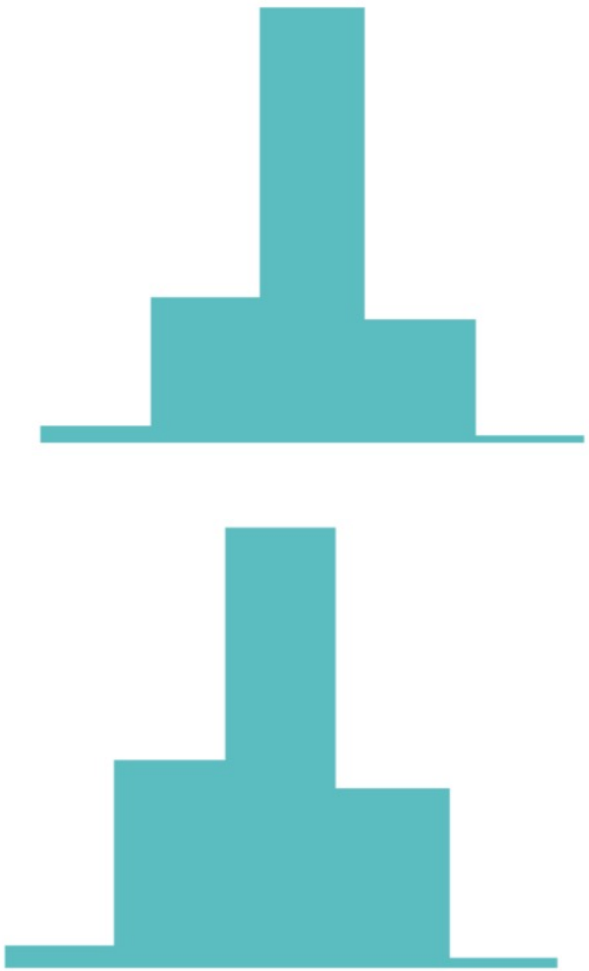


Figure 7.33 Histogram with a few bins.

- Histogram is used while exploring the distribution of numerical data.
- By plotting and analysing a histogram of a continuous numerical variable, one can discover its frequency distribution or probability distribution.
- Histograms in Fig. 7.33 look similar.
- Because these histograms are similar, we assume that the data is similar as well.

Histogram Confuser

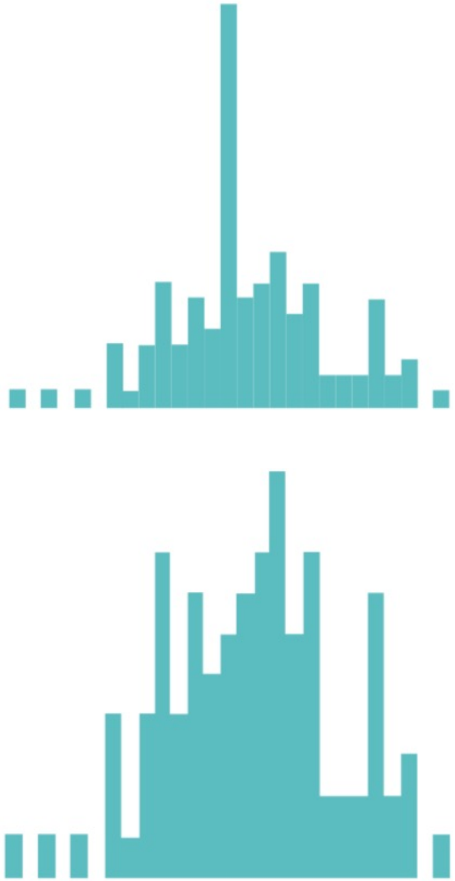


Figure 7.34 Histogram with many bins.

- But check Fig. 7.34, representing the same histograms with many bins.
- We can see that the underlying data in these histograms is different.
- This is known as histogram confuser/hallucinator.

Histogram Confuser

- Here, finding “just the right bin size” is a problem because:

   1. **If we have too few bins, then irrespective of the underlying data, histograms look similar**
   2. **If we have too many bins, the distribution will look radically different even though it might be just a sampling error.**

- In order to overcome this issue, try different chart types like density plot, violin plot or box plot to validate your inference.

More Examples of Misleading Visualization

- The World Health Organization (WHO) declared the COVID-19 out- break a pandemic in March 2020, alarming the entire world.
- Since the COVID-19 outbreak, the world has been on lockdown and people, businesses, news media, social media feeds, and political parties have shown interest in the latest and most informative trends.
- As interest in data visualization increased, misleading data visualizations also increased.
- Data visualization can be a great tool for communicating complex data succinctly, it can often be used to deliberately mislead the public and to fit a pre-set agenda.

More Examples of Misleading Visualization

- A colour gradient can be an interesting choice when it comes to data visualization (Refer to Fig. 7.35).
- Color gradients, however, can also mislead because, as is conventional in data visualization, a darker colour represents higher values, and as the number decreases, so does the shade of the colour.
- Therefore, the reader tends to associate lighter shades of colour with fewer cases and darker shades with high number of cases.

More Examples of Misleading Visualization

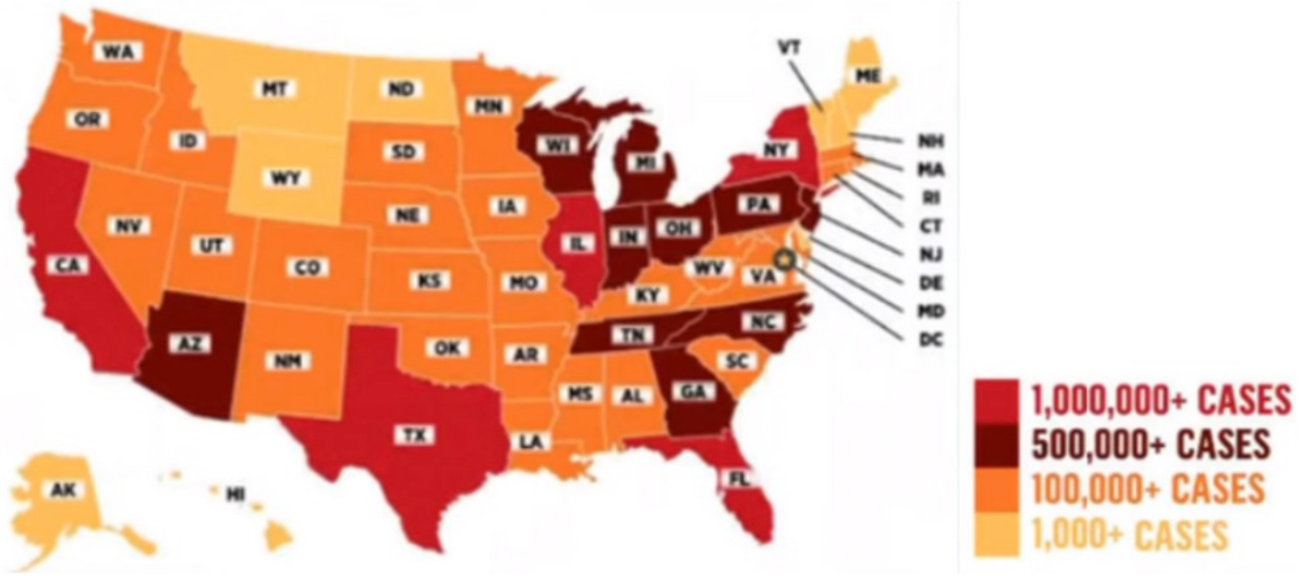


Figure 7.35 Total number of confirmed COVID-19 cases in the United States.
Source: NBC News

- But in Fig. 7.35, the darkest colour represents the second highest bin and not the highest number of cases.
- The colour gradient has either been used for decorative purposes or to mislead audience to show that the situation may not be as bad as it seems.

More Examples of Misleading Visualization

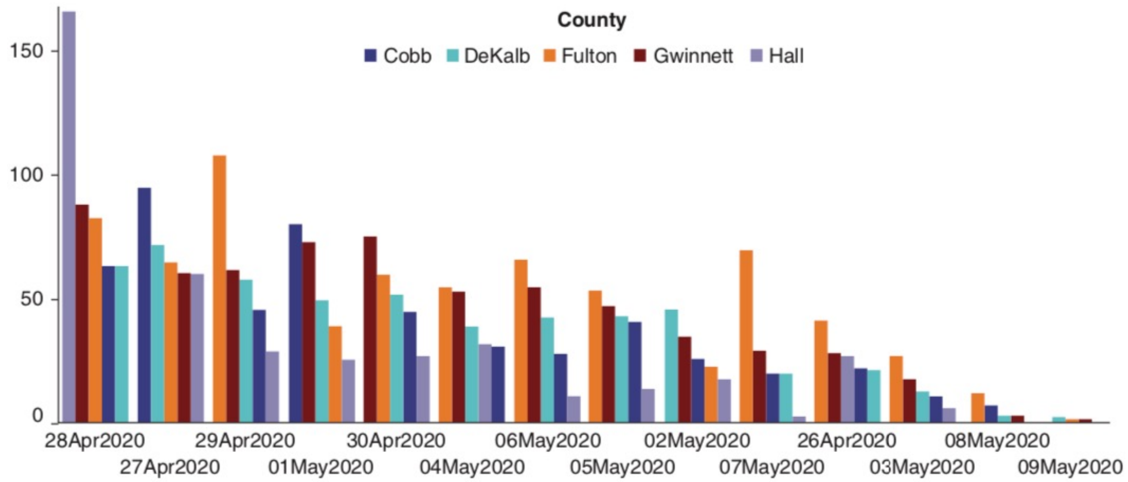


Figure 7.36 Top 5 Counties with the greatest number of COVID-19 cases in Georgia.
Source: Reddit. Originally comes from Georgia Department of Public Health

- The graph shows the number of cases over time and the counties impacted over the past 15 days, published on the Georgia Department of Public Health's website.
- A two-week period from 28 April 2020 to 9 May 2020 is represented by the horizontal axis of the graph.
- We can infer from the graph that new confirmed cases in the counties are decreasing steadily every day, and Georgia is on its way to wiping out the coronavirus completely.

More Examples of Misleading Visualization

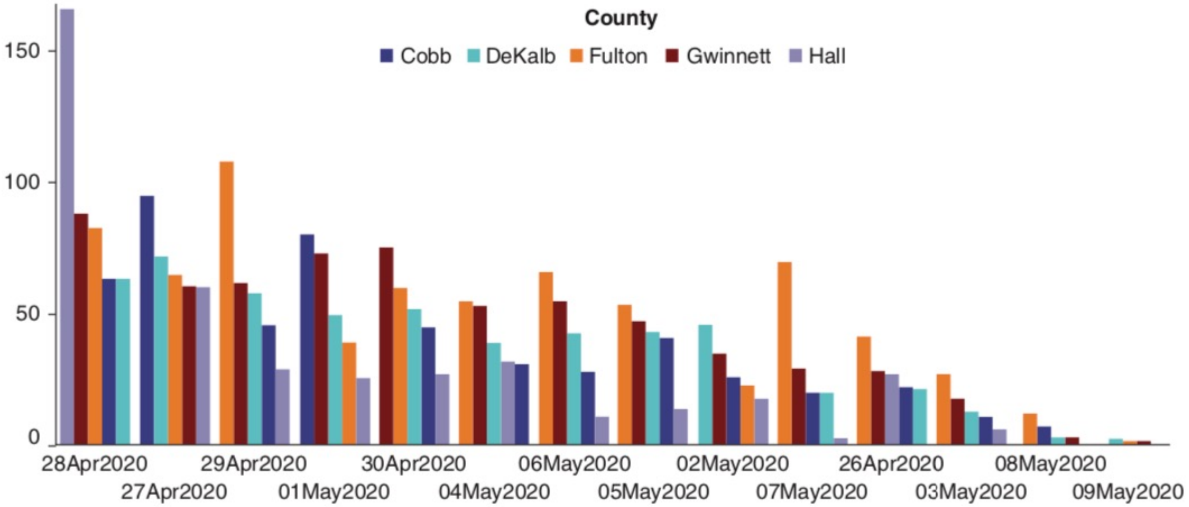


Figure 7.36 Top 5 Counties with the greatest number of COVID-19 cases in Georgia.
Source: Reddit. Originally comes from Georgia Department of Public Health

- If we look at closely, we see that the horizontal axis has been manipulated so that the dates have no set order.
- They have been arranged to a pre-set agenda to convey a downward trend.
- Since most of the world was trying to showcase that they have ‘flattened the curve’, even here, the author of the visualization tried to mislead the audience by manipulating the horizontal axis.

More Examples of Misleading Visualization

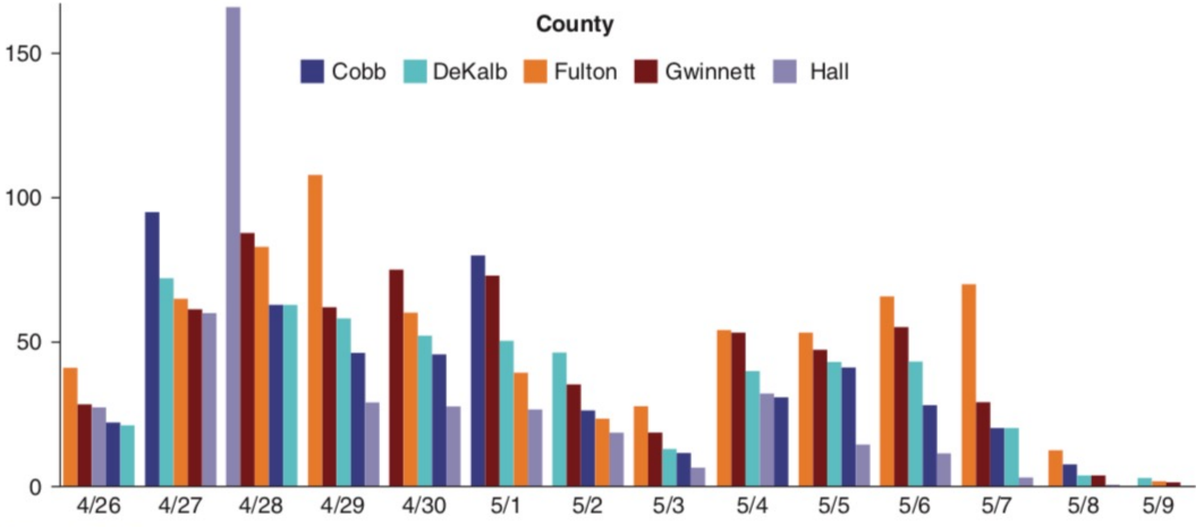


Figure 7.37 Top 5 Counties with the greatest number of COVID-19 cases in Georgia.
Source: Georgia Public Health Department

- Figure 7.37 shows the updated graph which shows the correct scenario of the number of cases over time.
- This explains that there is no downward trend in the number of cases but more or less a constant trend over those two weeks.
- This kind of misleading visualization can affect the health of millions of people as policymakers make critical decision based on such analysis.

More Examples of Misleading Visualization

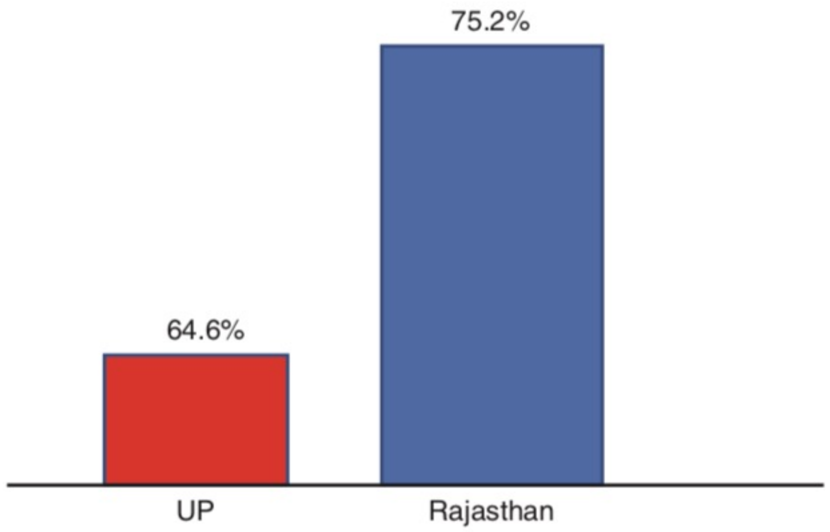


Figure 7.38 Recovery rate (in percent) comparison between UP and Rajasthan.

- In Fig. 7.38, where the vertical axis has been truncated to make it appear as if the difference between UP and Rajasthan is a lot bigger.
- The recovery rate in Rajasthan looks three times higher than UP but in reality, the difference is only 10.6%.

More Examples of Misleading Visualization

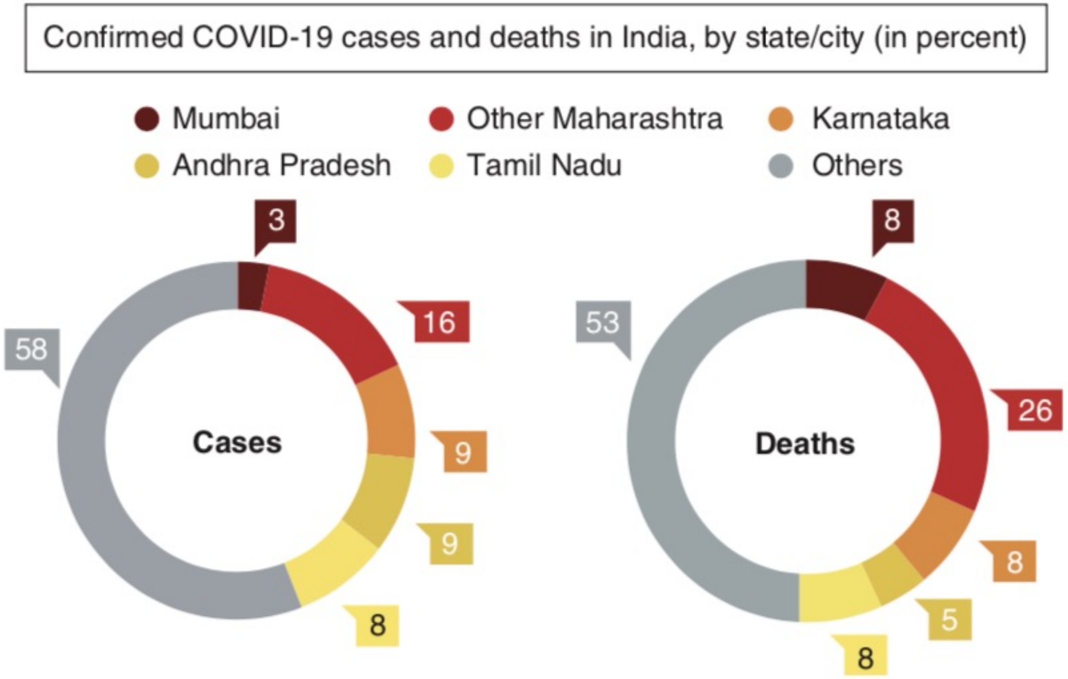


Figure 7.39 Confirmed COVID-19 cases and deaths in India.
Source: Ministry of Health and Family Welfare, Mumbai Health in India.

- In Fig. 7.39, confirmed COVID-19 cases and deaths in India are shown by state/city in percent, in the form of a pie chart.
- As discussed earlier, the sections of pie charts should add up to 100%, showing a proportion of the whole value.
- But in this infographic, when you add up the parts, it exceeds 100% which shows that either the visualization is manipulated, or the data is incorrect.

More Examples of Misleading Visualization

- When we create visualization, there is always the potential to distort or mislead.
- When critical decision-making by policymakers of a nation in the case of COVID-19 pandemic is dependent on such visualizations, it becomes a huge responsibility to show correct visualization.
- Hence, we need to critically reflect on our analysis before drawing conclusions from a visualization.

References:

- [Groenewegen, 2011] – Groenewegen A. (2011), “*Kahneman Fast and Slow Thinking Explained*”, BehavioralScience, SUE, available at https://suebehavioraldesign.com/kahneman-fast-slow-thinking/ #:~:text=Kahneman%20discovered%20not%20only%20the, systems%20arrive%20at%20different%20results, last accessed May 25, 2021.
- UCSB by the numbers available at https://twitter.com/EagerEyes/ status/13821850078, last accessed May 30, 2021.
- “Creator defends graph that appears to erroneously show a fall in Florida fun deaths”, available at https://usvsth3m.tumblr.com/ post/82779802419/creator-defends-graph-that-appears-to-errone- ously, last accessed May 15, 2021.

Thank You!