How To Lie With Charts

Elaboration on the presentation for the seminar
"How do I lie with statistics?"

by

Florian Fallenbüchel

Contents

1 Introduction

This presentation mostly covers the book "How to lie with Charts" by Gerald E. Jones [1], but before we can talk about how to lie with charts, we first have to give some remarks on the reliability of data. Because even if the visualization is correct in itself, distortions can spring from the very subjective data interpretation of the chart creator. Take a look at the black and blue dress displayed in Figure 1.1. The original image in the middle led to a widespread discussion about the color of the dress in early 2015, as the colors are slightly desaturated, possibly due to a bad mobile camera, distorting the dress to appear white and golden. Together with inferior mobile displays around that time, the image divided social media communities, as both groups were certain to know the "real" color of the dress.

The human perception is very subjective as the brain just interprets the signals from the different senses to form our reality and therefore data collected by humans is prone to distortions through interpretation of the collector. Even the simple act of counting things requires you to apply your own notions of reality and possibly distort the result. For example, I count the fruits in the next supermarket and give you the resulting number 42 without any label or further explanation. Did I count every fruit, or just apples, because maybe I don't eat any other fruits? And did I count the rotten ones? Some people won't eat a banana with brown spots as they already consider it rotten, but again many others would still happily eat them. And are we talking about 42 pieces of fruit? Maybe I was referring to 42 boxes of fruits, as pieces do not sound like a lot for a whole supermarket. The important point is that a visualization can only be as good as the underlying data and we

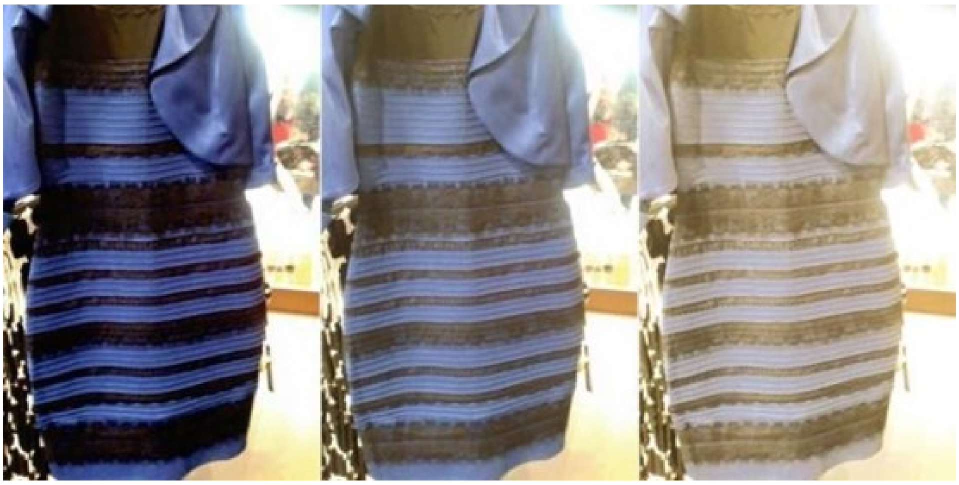


Figure 1.1: Black and blue dress which appears to be white and golden, when messing with the saturation of the image.

Figure 1.2: The internet opens up a multitude of possibilities for both helpful and abusive purposes.

have to be aware of possible distortions, as a careless chart creator might subconsciously distort the data to proof his point or as a blatant liar might try to play with our own subjective perception of the data.

In recent years, modern tools like PowerPoint have eased the creation of various types of charts to a point where it only takes a few clicks to generate visualizations from any data. The downside of these tools is that most of the design choices are made by the implemented algorithms, applying predefined criteria, which can lead to a distorted representation of the data. The usage of these tools requires little thought and, as the human brain is really efficient, most people won't give too much thought if it's not mandatory. All this favors the abuse of these tools to, intentionally or not, create misleading charts. With the invention of the internet, people gained access to a lot of information, where it is not always clear how correct or reliable the underlying data is. So everyone can reach many people through social media and everyone can create charts at a click, therefore we need to learn about the distortions in charts in order to be able to detect them, when we are confronted with a dubious visualization.

But even in the offline world, a little lie is sometimes enough to confuse the audience of a presentation by making a small adjustment to the chart, so you can quickly go over a critical slide without the audience noticing the actual message of the data. This presentation will cover some of the most common mistakes and deceptions that occur when dealing with the different types of charts. Let us start with a simple and familiar type of chart, the pie chart.

2 Pie Charts

Pie charts are among the most common types of charts and are widely used to visualize market share and other proportions. Their purpose is to visualize the share of the different instances in a whole, and this should be their only purpose. If somehow the actual amount of the different entities is important, choose another chart! A good pie chart is therefore always labeled with percentages so that it does not confuse the audience into making assumptions about the underlying data.

2.1 Pies Are For Percentages!

Take a look at the two pie charts in Figure 2.1. The purpose of these charts was to inform the management about the proportion of the different tasks to the whole workday for their average worker. The chart on the left is correctly labeled with percentages and drives the attention to the important point. The right chart on the other hand is labeled with the actual average work hours per task. Putting actual values on slices distracts the audience from the message as they are tempted to sum up the values. And even if they don't do the math, they still establish a mental impression of the data. Here the values add up to 7.5 hours, giving the wrong impression that most workers only do so much work on an average day. But most employees work 8 hours or longer, which may

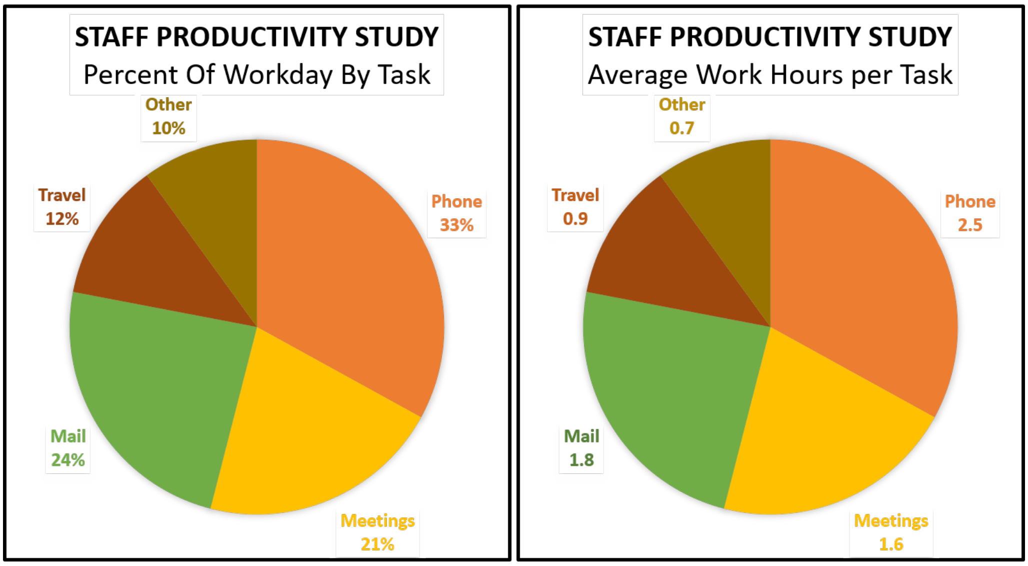


Figure 2.1: The purpose of pie charts is to focus the audience on proportions, the actual amount should not be important. Therefore they should only be labeled with percentages.

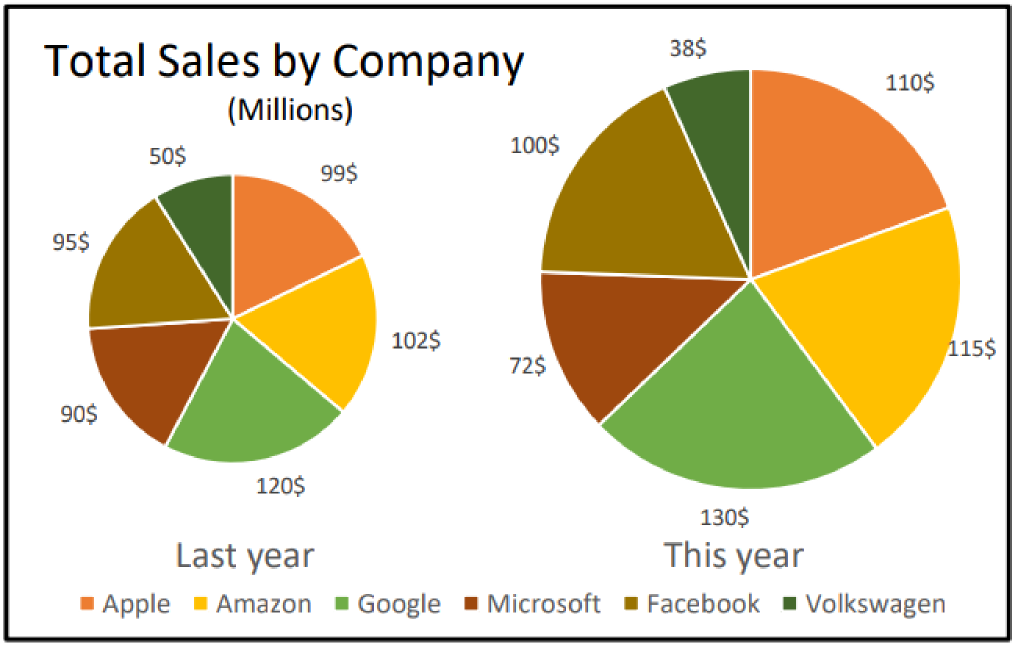


Figure 2.2: The size of the chart for the current year has been falsely adjusted to reflect the increase of the total sales compared to last year. This disguises the losses of Volkswagen and Microsoft.

worry management about how their employees spend the remaining 0.5 hours. The difference was due to the accounting of part-time workers and is therefore negligible, but at this point you might have already lost your audience, as they think about the implied messages.

2.2 Confusing With Size

As already mentioned, the actual amount should never be important for a pie chart, a single chart represents a whole. Therefore you should never adjust the size of two pie charts to reflect some difference of the total value. Human perception has difficulties to correctly correlate the magnitude of the difference with the change in the area size, the increase is always perceived smaller than it actually is. A blatant liar might abuse this to give you a wrong impression of the data. An example for this can be seen in Figure 2.2, which shows the share of some large companies in the total market sales of the previous and current year. The size of the second pie has been adjusted to reflect the increase of the total sales compared to last year, but while the increase in sales is around 1.6%, the size was more than doubled. This disproportionate scaling also increases the size of the slices of Volkswagen and Microsoft more than their size reduction through smaller sales volumes. This visually disguises their losses, making the damage appear less dramatic. This effect is combined with the labeling of the slices with actual sales values instead of percentages to further deprive the recipient of his feeling for the shares. It's small lies like these where you play with the perception of your audience which can sometimes help you disguise inconvenient data.

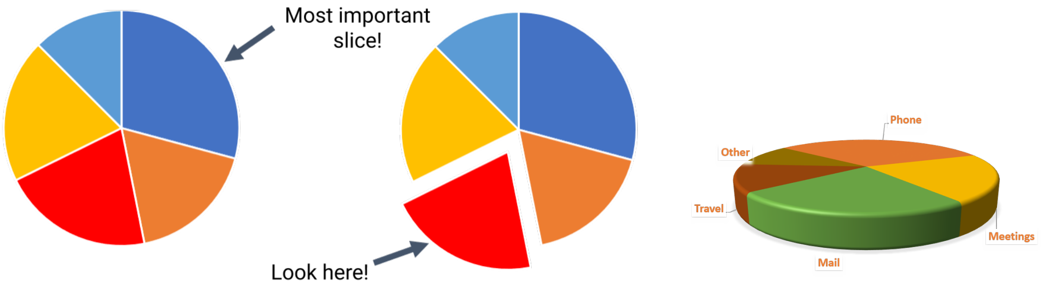


Figure 2.3: The perception of the importance of slices is based on their position. The top right slice is usually perceived as the most important one, except another slice is exploded from the pie. In 3D the bottom slice becomes the most important as it is visually enlarged by the tilt.

2.3 Position And Perception

The positioning of the slice in the pie also has a big influence on how important the respective slice is perceived and can be used to guide the attention to certain parts. Figure 2.3 gives an overview on the common perception of pie charts. In general, most audience would give most importance to the upper right slice, with decreasing importance clockwise. Notice how the top scoring companies were also placed in the upper right part of the charts in Figure 2.2 to emphasize their increase in sales. This rule applies for flat charts unless another slice is exploded from the pie, as this emphasis is stronger than the natural one through placement. In the 3-dimensional case, the bottom slice becomes the most important one, as in addition it is visually enlarged by the inclination of the plot, literally giving it an edge over the other slices. By increasing the tilt and shifting the center of the pie, we can maximize this effect to totally confuse the actual proportions.

A professional liar would not exaggerate this too much, of course, but this trick is widely used in public presentations and advertising to subtly highlight the preferred slice. The left image in Figure 2.4 shows a screenshot from an Apple presentation, showing their share on the phone market. They positioned their slice in the most prominent place to make it appear bigger compared to the slice of IBM with double the share, but to be fair, they chose a rather small tilt.

They also made use of another trick, the "Others" category. This mystery category can be abused to hide all kinds of inconvenient data as it is up to the creator of the chart to define what is important and what is negligible. Notice how they placed this category, which is bigger than the

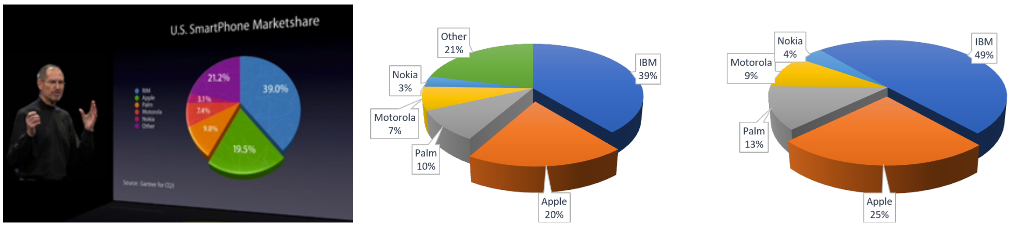


Figure 2.4: The mystery category "Other" can be abused to hide inconvenient data. Dropping this category can be used to enlarge the remaining slices in your favor.

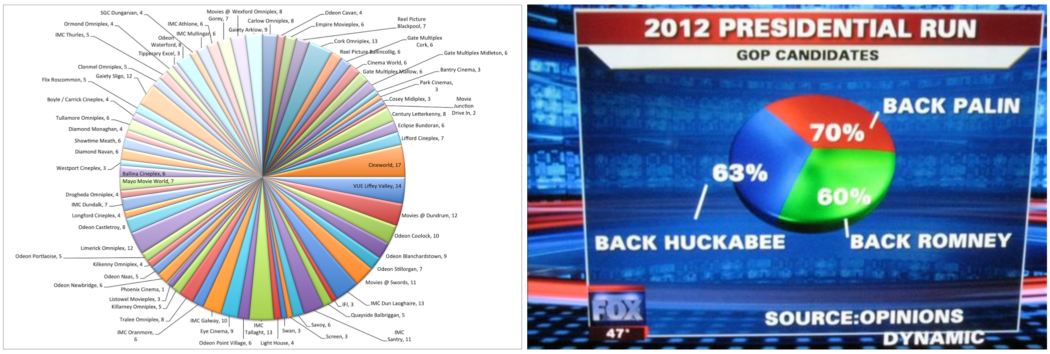


Figure 2.5: Accounting for too many categories can clutter the pie, making it almost impossible to get any valuable information from it. With percentage labels which do not sum up to 100 you can give a false impression of accuracy.

Apple slice itself, in the upper left corner, the least important place. There are valid reasons for an "Other" category of this size, like a heavily fragmented rest, but in general you should become suspicious if someone tries to feed you such a big slice without telling you what it consists of. On the other side, if you want to avoid unpleasant questions about your mystery category or if it complicates your story, you can simply drop the whole slice like in the right image of Figure 2.4. You not only remove the awkward data, you also increase the share of the remaining slices as the whole is reduced.

2.4 Other Examples of Misuse

There are various other ways to confuse about your pie chart, which we will not cover to full detail now. Just to give you two more quick examples, take a look at the pies in Figure 2.5. The pie on the left includes a huge variety of categories, which are all individually labeled. While the pie is not incorrect in itself, due to the heavy fragmentation, most of the information is inaccessible without a detailed study of the plot. If you grant your audience only a quick glance, you could hide a lot of inconvenient data in there.

Another common trick to give a false impression of exactness is to label the slices with percentage values which do not sum up to 100, like in the pie chart from Fox News in the right image of Figure 2.5. For the survey behind this chart, they allowed multiple answers for the favorite candidate. Without any further explanation on the voting behavior, this plot yields basically no information, while it still gives you the impression to inform you about the state of the elections. In the next chapter we will talk about the most common type of chart in western societies, the XY-chart.

3 XY-Charts

As we have already mentioned, the position of a slice in a pie chart influences the perception of its importance and guides the focus. Every audience carries subconscious assumptions about the meaning of position and orientation, usually influenced by their cultural background. It is important to be aware of these biases as they strongly influence the impression your listeners get from the chart, whether you want to lie or not. Most cultures associate upwards movement with increase and gain, while downwards movement is associated with loss. Western audiences read from left to right. As a consequence, these audiences tend to associate motion to the right with progress or positive movement, as well as the flow of time, and, on the other hand, leftward motion is usually considered backward or bad. Of course, these effects combine, like visible in the right image of Figure 3.1, showing a positive, future-facing arrow pointing upward right, and a terrifying, hideous arrow pointing downward left. These assumptions are the basis for XY-charts, the most common type of chart in western societies.

3.1 Orientation

Applying to the biases we have just mentioned is very important to transport your message, as the orientation of your plot already implies a message without much context. Take a look at the plots in Figure 3.2. The left chart only shows a series of data points connected by a line. There are no labels on the axes, only the line is labeled with "U.S National Debt" next to an arrow reaching way outside the range of the y-axis. Without further description, one has the impression that the national debt has exploded in recent years and is getting out of control. A simple way to hide this development would be to simply switch the axes of the chart, resulting in the plot on the right. This disguises the original message, as it confuses the audience about their intuitive interpretation

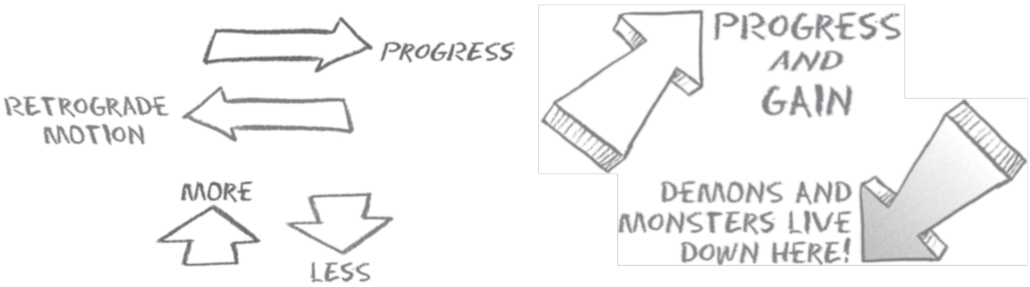


Figure 3.1: Most cultures associate upwards movement with increase, downwards movement with decrease. Western audiences relate rightward motion with progress, and conversely leftward motion is considered backwards. These effects also combine.

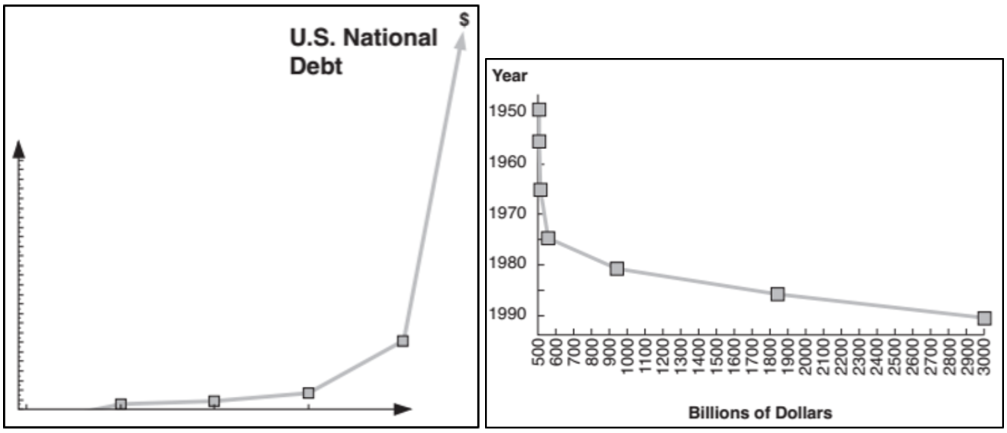


Figure 3.2: The orientation of the left chart implies the message, that the U.S national debt is skyrocketing, without any label on the axes. Flipping the axes results in the plot on the right. Despite the detailed axis description, the plot already appears a lot less concerning, due to the downward orientation alone.

of the directions, while still being perfectly valid math-wise. Even though the right plot shows a more detailed and accurate description of the change in debt through the axes labeling, the plot gives a way less concerning impression, and sometimes this is all you need. While they are busy trying to understand your unconventional chart, they are distracted from the important point, giving you again the possibility to quickly sweep over some unpleasant slides.
Many companies like to abuse this positive association by displaying cumulative sales volumes for their shareholders. An example for this can be seen in the Apple presentations in Figure 3.3, in which they showed cumulative sales for their iPhones and iPads. By the very definition, these volumes can only increase and therefore the representation always show upward oriented lines with a positive connotation. This gives you the incentive to get the impression that these sales can only go up! And this is not even a lie. These charts can also be exploited to conceal the fluctuations in sales over the quarters as they make it difficult to reconstruct this information from the chart.

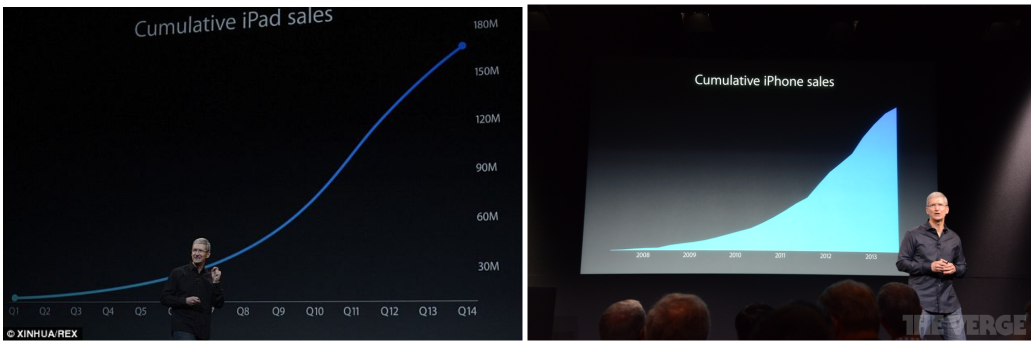


Figure 3.3: Charts of cumulative sales abuse the positive impression of upward motion, as, by definition, they can only grow.

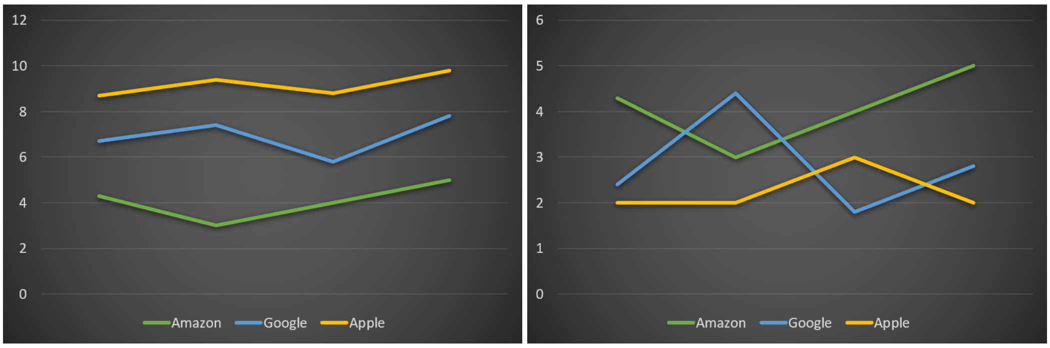


Figure 3.4: The stacked chart on the left gives a wrong impression about the sales volumes of the different companies. The strong sales of Amazon were chosen as the bottom line to diminish the fluctuations of Google and to push the declining sales of Apple to the end of the year. The right chart shows the values with a common baseline at y=0.

3.2 Stacked Charts

There are several ways of displaying your chart in a way that is technically correct, but plays with the expectations of the audience to make them misinterpret the data. A common assumption of most audiences is that every line of a chart has the same baseline at $y = 0$, the x-axis. A deceptive liar could abuse this presumption and present his audience a stacked chart without telling them about it. In a stacked chart, each line is the baseline of the next instance and therefore influences its fluctuations. Take a look at the example charts in Figure 3.4. The left graph could give the impression that Apple had the best sales during the year, followed by Google, and Amazon with the lowest sales, but all companies had a strong end with Christmas sales. The right chart tells a more accurate story. Apple was actually at the bottom with the least sales and a minor decline to the end of the year. The strong finish of Amazon was the baseline for the other two companies and pushed the appearance of their sales to the end of the year and the strong fluctuations of Google are mitigated in the stacked chart by the opposite behavior of the Amazon sales. The difference between a stacked and a regular chart is almost impossible to notice if you do not inform your audience about it.

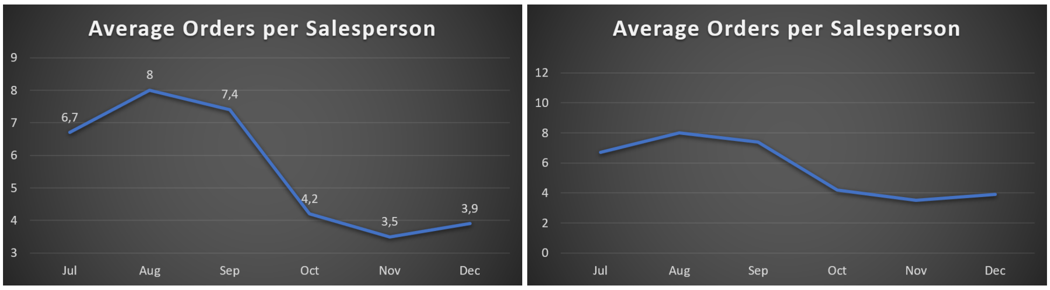


Figure 3.5: By increasing the displayed range on the y-axis we can minimize fluctuations of the representation. This can be used to make changes appear less dramatic

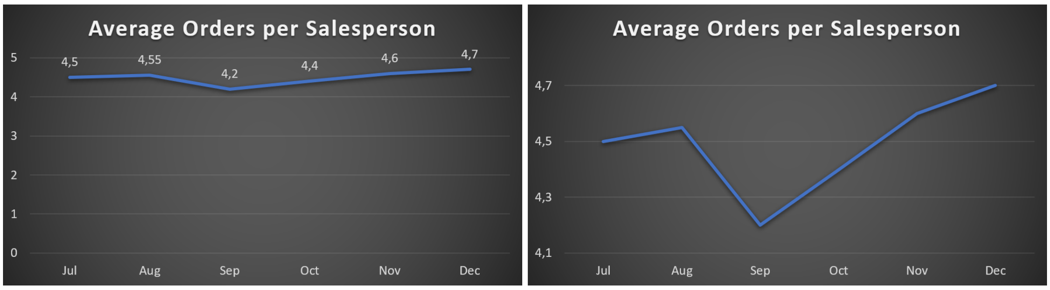
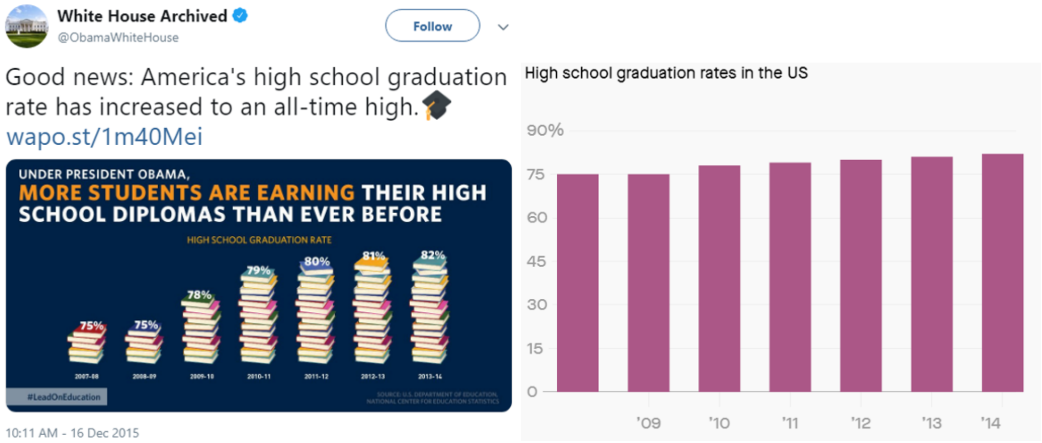


Figure 3.6: By decreasing the displayed range on the y-axis we can maximize fluctuations of the representation.

3.3 Messing With Axes

Adjusting the range displayed on the axes is another way to change the appearance of the chart in your favor. This effect can be used in both directions, depending on what you desire. Figure 3.5 shows an example where the increased range on the y-axis minimizes the visual variation of the plot. While the appropriately scaled chart on the left might worry the management about the heavy decline in sales to the end of the year, the rescaled chart on the right can be marketed as a "mild seasonal adjustment" and you quickly turn to the next slide. On the other hand, decreasing the displayed range can exaggerate a small increase like visible in Figure 3.6. The small range maximizes the small fluctuations of the data, giving the impression that, after a collapse in September, the sales got a big boost to the end of the year and finished strong.

This is a widely used trick, even from "reputable" sources. The left image of Figure 3.7 shows a tweet from the White House under the Obama presidency, which contains a chart visualizing the change of high school graduation rates over the years. They visualized the percentages with different amounts of books, which do not really correlate with the difference in value. For the 3% difference between 78 and 75%, they added 5 books to the tower, for the next 1% difference


White House Archived
@ObamaWhiteHouse

Good news: America's high school graduation rate has increased to an all-time high.🎓 wapo.st/1m40Mei

10:11 AM - 16 Dec 2015

Figure 3.7: Tweet from the White House under the Obama presidency, 2015.

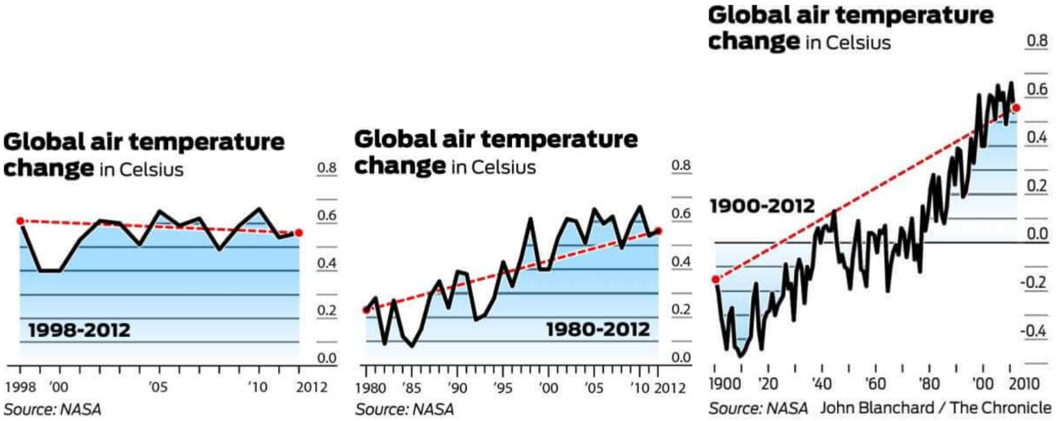


Figure 3.8: Depending on the time slice selected, the change in the global air temperature appears to be more or less problematic.

they added another 5. Displaying the percentages in a plot with consistent heights like in the right chart appears way less impressive. Figure 3.8 shows an example where this trick was applied to the time axis to confuse about the dimensions of climate change. Because 1998 was an exceptionally hot year, the displayed section can be chosen in such a way to imply that there is actually a slight downward trend in temperature.

3.4 Correlating Axes

Plotting several data series along in the same chart with multiple axes is another way to imply all sorts of things. This trick plays with the common misconception that correlation equals causation. Take a look at the chart in Figure 3.9, which states that there is a 66% correlation between the

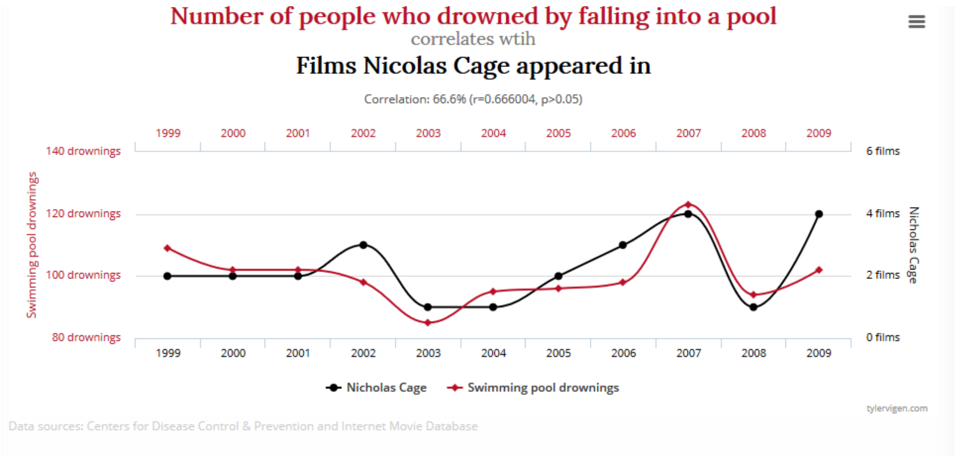


Figure 3.9: Unified plotting with multiple axes can be abused to imply correlation where none exists, it is only important that the data appears to be correlating.

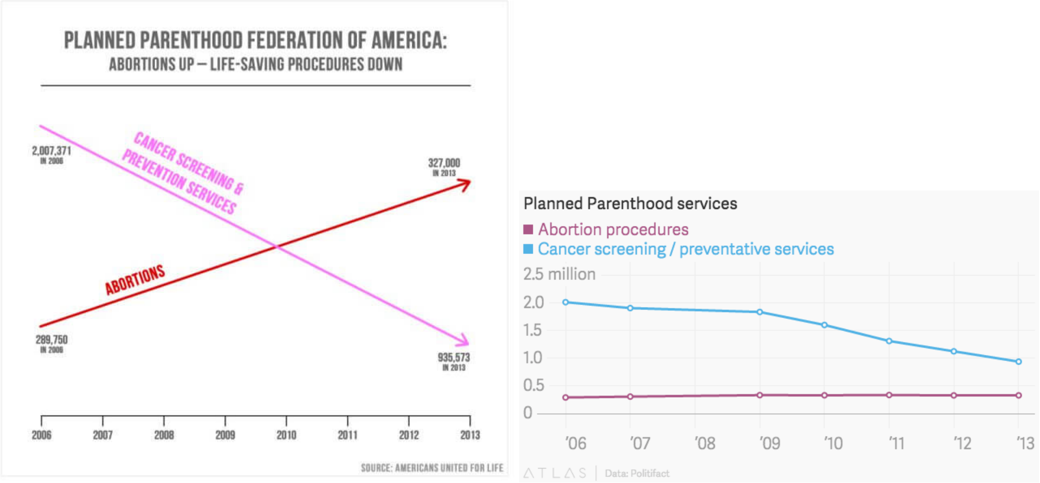


Figure 3.10: Chart from a republican discussion about rising abortion rates compared to declining cancer screening and prevention services. The adjacent graph shows both rates on a comparable scale, whereby the enormous implied increase in abortions is almost undetectable.

number of people who drowned by falling into a pool and the number of movies in which Nicolas Cage appeared during that year. At first glance, both lines seem to coincide quite well, especially in the beginning and end segment of the lines, which gives the impression that there might actually be a connection. While this relation might feel entertaining for you, never forget that the displayed range for the drownings got chosen in such a way that they better fit to the movie count, and that there are several other factors, like weather conditions, which have a greater influence on pool drownings than Nicolas Cage. Or are his movies really that bad?

A more insidious way to lie is to use an individual scale for the different axes if they both have the same value unit. Figure 3.10 shows an example for this from a republican discussion about rising abortion rates. They presented this chart and used it as an argument against planned parenthood funding, saying that the money of tax payers is rather used to prevent life instead of saving lives with cancer prevention. Even though both numbers could be displayed on a single axis, they chose multiple ones with different scaling to visually exaggerate the comparatively small change in abortions. Abortion rates raised by around 38000, while cancer prevention services declined by around 1 million. Plotting the actual values over the years in a comparable diagram as in the right image of Figure 3.10 makes the change of abortion rates almost invisible in contrast to the services.

4 Trends

Foreseeing the future has always been an inexact science. Just remember famous words like "I believe in the horse. The automobile is a temporary appearance!" by Wilhelm the Second in 1916, "There is no reason anyone would want a computer in their home!" by Ken Olsen in 1977, or "There is no chance that the iPhone is going to get any significant market share!" by Steve Ballmer in 2007. Most predictions are based on an abstraction of reality, as it is almost impossible to take every possibility into account, and are influenced by some subjective bias. As it is very complicated to account for reality in its completeness and as this can sometimes be more confusing than enlightening, people tend to oversimplify relations and discard outcomes with low probabilities. You should always be cautious when someone presents you his prediction of the future as it is very likely that the person has not considered every option.

Many people would argue that the average is a good approximation for a future outcome of an event. But from a statistical standpoint, there are multiple ways to define the average. Here lies another possibility to lie with your chart, as you could choose an average which is more suitable for the trend you want to imply to your audience. Besides, the average is only a good approximation if the values are consistent, as high variance can give a wrong impression about the probability of

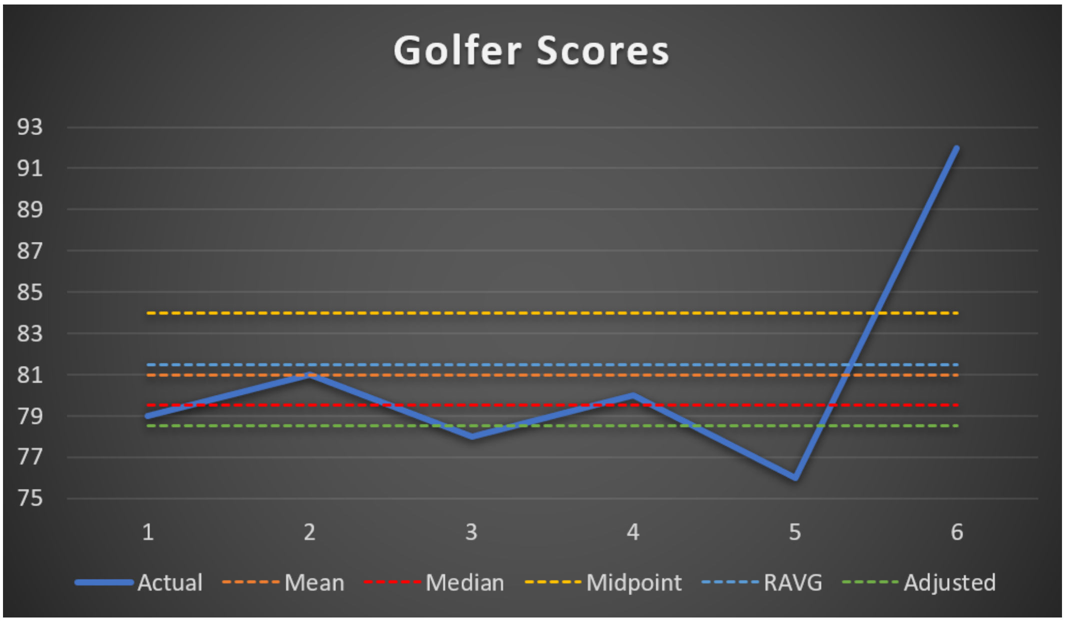


Figure 4.1: Comparison of the different averages for the same six golf scores. Simply choose the definition which best fits your story.

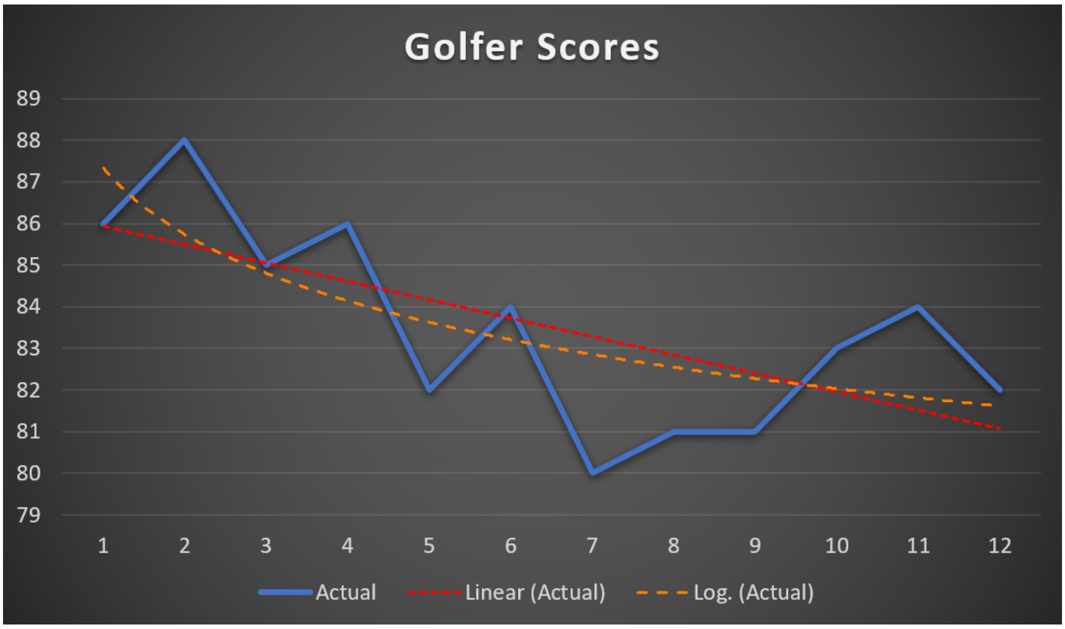


Figure 4.2: Trend lines can be fit to the data points with regression. Omitting unpleasant data points can shift the curves in your favor.

certain values. Take a look at the chart of golf scores in Figure 4.1. The golfer scored 79, 81, 78, 80, 76, 92 in six games, shown by the jagged blue line. The arithmetic mean for his games would now be $\frac{79+81+78+80+76+92}{6} = 81$, the slightly more favorable median $\frac{79+80}{2} = 79.5$, and the awkward midpoint up at $\frac{76+92}{2} = 84$. Each of these values can be presented as the average, depending on what best fits the message you want to convey to your audience. And it does not stop there, as for example you could use a 4 game running average to stress on the current trend of the data, resulting in a score of 81.5. You could also argue that the 92 score game is due to the golfer smashing his hand in the car door on the way to the course and as a result deleting the score as an outlier from your calculations, giving a positive score of 78.5.

Methods from regression analysis can be used to fit trend lines to the data points as visible in Figure 4.2. There are several types of trend lines, while the most common ones are linear, exponential and logarithmic curves. Here lies another bias, as people tend to assume that the line that best matches the data points must be the best estimate of future outcomes. But again, keep in mind that these lines are a huge abstraction of reality and do not account for short term developments. Also a bold liar could discard data points from the regression models as outliers to reinforce his hypothesis with a trend line that better fits his story.

Bibliography

1. G. E. Jones. *How to Lie with Charts, 4th Edition*. LaPuerta Books and Media, 2018. ISBN: 0996543864.