The Art of Data Visualization & Why You Should Care

Utsav Vachhani

Data has been present since the ancient civilizations. For well over 5000 years, we have applied it to acquire knowledge of all aspects of life and has drove our curiosity towards new discoveries. The creative nature of humans has also led us to visualize this information in variety of ways that have transcended languages. Mithun Sridharan offers many notable examples of visualizations in his article with some dating all the way back to ancient Egypt and Babylonia.[1] The figures below highlight some of these intriguing data visualizations. The first figure represents the oldest map, one made from the skin of Papyrus, it displays mining information in the nearby wadi ('valley' in Arabic).

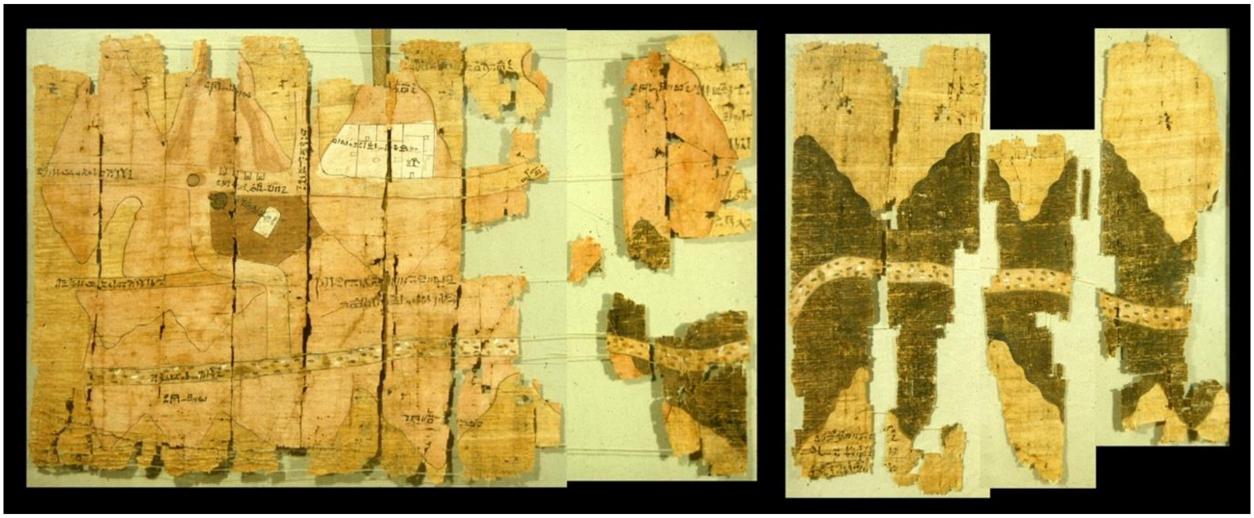


Figure 1.0 – Turin Papyrus Map (Egypt, 1150 BCE)

[1] https://thinkinsights.net/data-literacy/data-visualization-history/

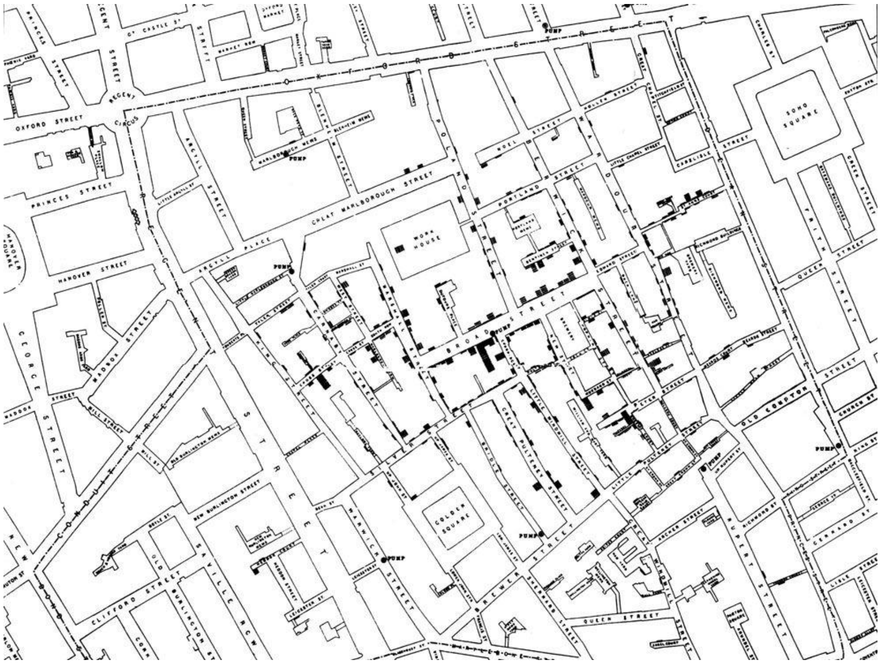


Figure 2.0 – Cholera Outbreak of London (John Snow, 1854).

In Figure 2.0, the map by John Snow gave critical insights on the potential causes of Cholera by plotting the collected infection data with the location information and drastically reduced the severity of the disease.[2] Visualizations such as these paved an important pathway towards advancement in human knowledge. Within the digital era today, data visualization today has evolved into being universal available and widely utilized from academic institutions to large organizations across the world. Visualizations from data have many uses, one being helping students learn visually concepts that are difficult to grasp.

The Central Limit Theorem in Statistics is a central (pun intended) concept which is easier to understand for students when its properties are visualized by the density plots in *Figure 3.0*.[3] This means high school teachers to university professors can now use this a great resource in teaching – where students previously may have solely relied on text to engage with new knowledge. Similarly, in academic research, data visualization represents a way for researchers to convey their ideas in a manner that appeals to a

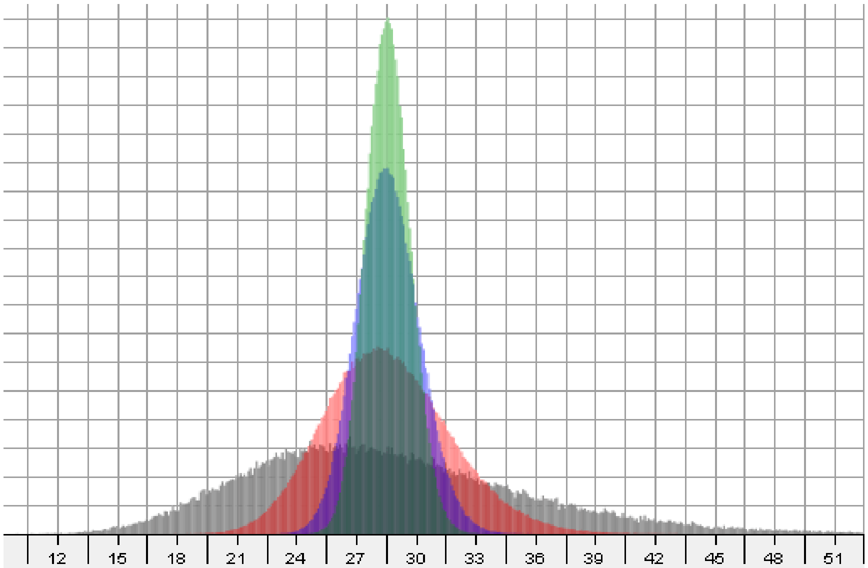


Figure 3.0 – Central Limit Theorem

[2] https://www.nationalgeographic.org/activity/mapping-london-epidemic/
[3] https://statisticsbyjim.com/basics/central-limit-theorem/

larger audience. Many properties of a data visualization, such as the type of graph and color palette used by the author, help bring the data to life from an otherwise mute text. Enago Academy summarizes this well that for researchers, “it facilitates theory refining, uncovers holes in data sets, and supports model development. It also summarizes data-supported conclusions, promotes online sharing, and becomes a method by which scientists can advocate for policy changes.”[4] As one can see, data visualization’s powerful techniques serve as a cornerstone in one’s ability to expand on their existing knowledge, much like in ancient times.

Outside of academia and research, data visualization has taken a pivotal role in helping organizations grow. From Fortune 500 companies to global humanitarian funds, visualizations are being used to inform and convince its audience. When I was in undergraduate studies at University of Illinois Urbana-Champaign, I was part of UNICEF USA’s chapter and I vividly remember using graphs and other data metrics in presentations. It was an active part of our fundraising campaign and was also heavily present at UNICEF Summit in Washington D.C. that I attended. When I was elected as the Treasurer, informing the general members on the progress of our annual donation goal was often in the form of graph(s) that were more appealing (notice the pattern) for communication. In the industry, there are countless examples of data visualization, but I want to highlight one of my favorites, Spotify.

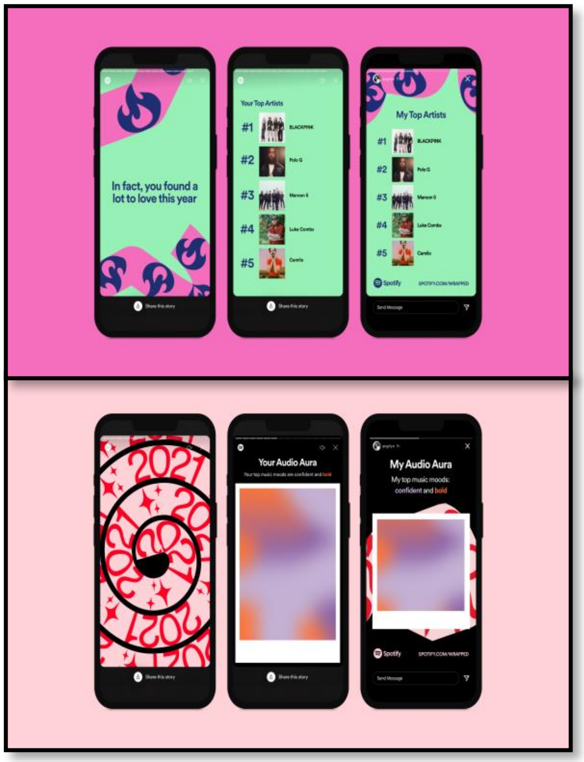

In fact, you found a lot to love this year
Your Top Artists
My Top Artists
#1
#2
#3
#4
#5
Spotify
2021
Your Audio Aura
My Audio Aura
My top music moods: confident and bold

Figure 4.0 – Examples of Spotify Wrapped

[4] https://www.enago.com/academy/why-is-data-visualization-important-in-academic-research/

Even if you are not one of the millions of users that uses the platform for music, you have seen Spotify Wrapped series on a friend's social media account. The Wrapped campaign started by Spotify has been extremely successful and unsurprisingly, one of the biggest factors for its success comes down to data visualization. When you are spending thousands of minutes listening to music, it is natural to be curious the statistics. Instead of showing a dull screen filled with numbers (which I am certain no one wants to see), Spotify brilliantly took this opportunity to display beautiful visualizations of the *same* data. The impact? It became a popular trend on social media while helping Spotify cement its position as the leader in the streaming music industry. Simona Galant's article on Springboard, she mentioned that although Spotify is just giving the data back to its users, it is "the way the data is presented is what gets people excited in a similar way that a personality test might." They also have added other fun aspects such as telling the user that they are in "the top 1% of a band s most loyal followers or that they are among the bold, non-mainstream music listeners."[5] This is a powerful example of why data visualization is an *integral* part of organization today and the overall data science field.

Now that I have provided the ubiquitous use cases of data visualization, I would like to introduce key visualization knowledge that everyone should have, regardless of one's technical background. While you may not necessarily apply any of this to your profession, it is helpful to know, as you will see from the latter instances. All the information below is directly from the book *Fundamentals of Data Visualization* by Claus O. Wilke.[6] I highly recommend reading it if you are going into data science—or any data-related profession.

1. Aesthetics – these describe "how" the graph looks to the reader. The figure below states what falls under this category. From a prior math class, you might be familiar with the two-dimensional Cartesian coordinate system (x, y), which represent the position of a data point. Other important aesthetics include shapes used in the graph, adding sizes to shapes, colors as well as several types of lines. Different combinations of these aesthetics help create the foundation of many data visualizations that you may have encountered.

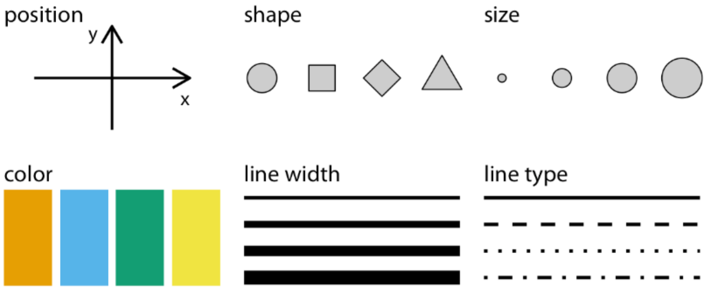


Figure 5.0 – Aesthetics of a graph

[5] https://www.springboard.com/blog/data-science/spotify-data-insights/
[6] https://clauswilke.com/dataviz/

2. Scale: A scale is what we use to assign a specific value of data to an aesthetic. If we want to record the times for a race lap, we might assign the x-axis to each driver, while using y-axis for the lap times. For instance, figure 6.0 has three scales. Two of these are position scales, month on x-axis and temperature on y-axis while the last scale being the color, which highlights the location.

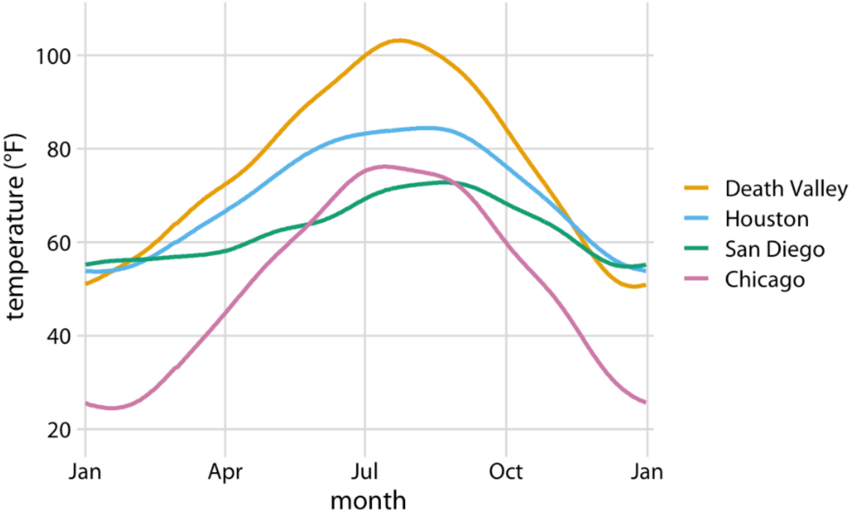
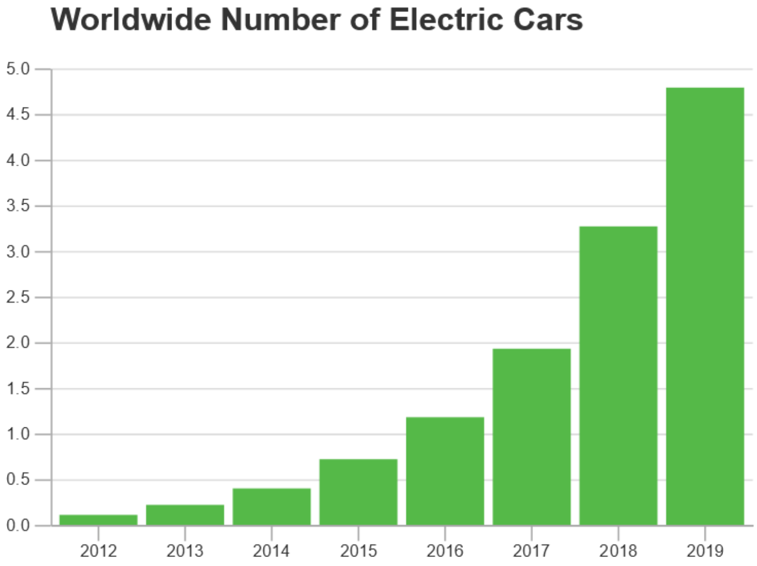


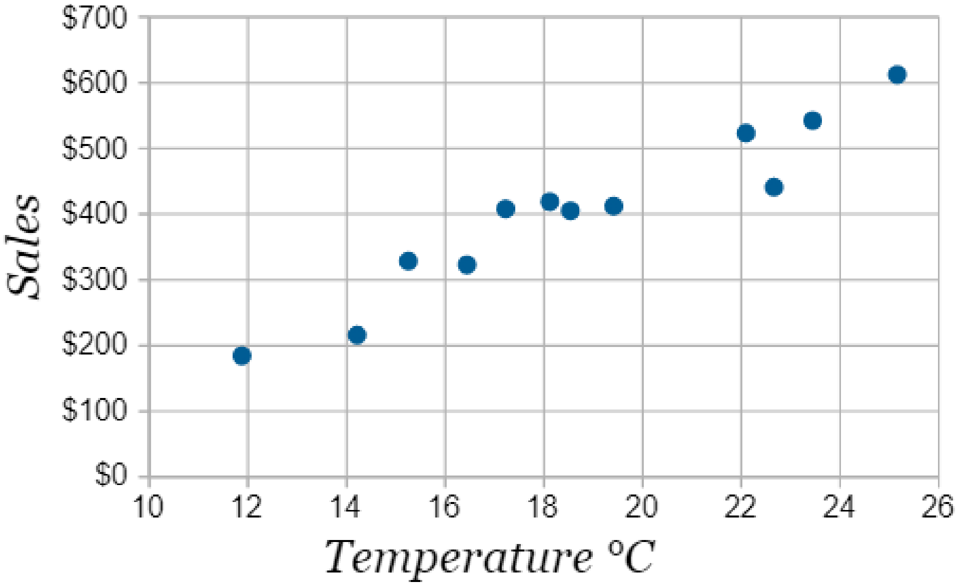
Figure 6.0 – 3-scale graph

3. Graphs: These are some of the widely used graphs in data visualization and for what they are used. Just for reference, a variable is a "thing" of interest that we want to visualize. Each graph below is represented with a real-world example and give you further insight on trends that take place within the data.
    a. [7]Bar Charts: shows the amount of a variable

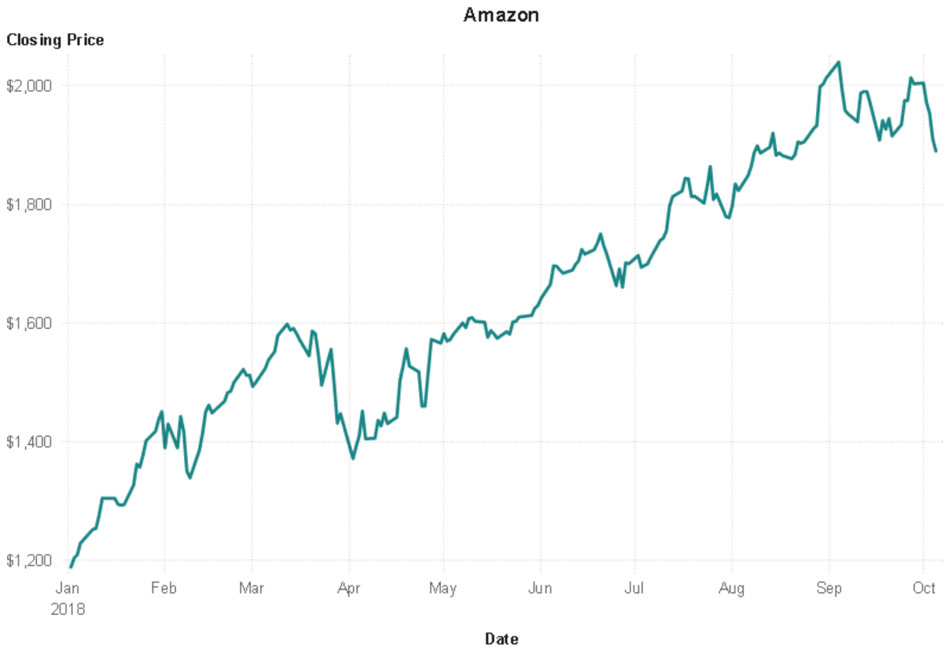


[7] https://www.smartdraw.com/bar-graph/

b. [8]Scatter Plot: shows relationship between two variables



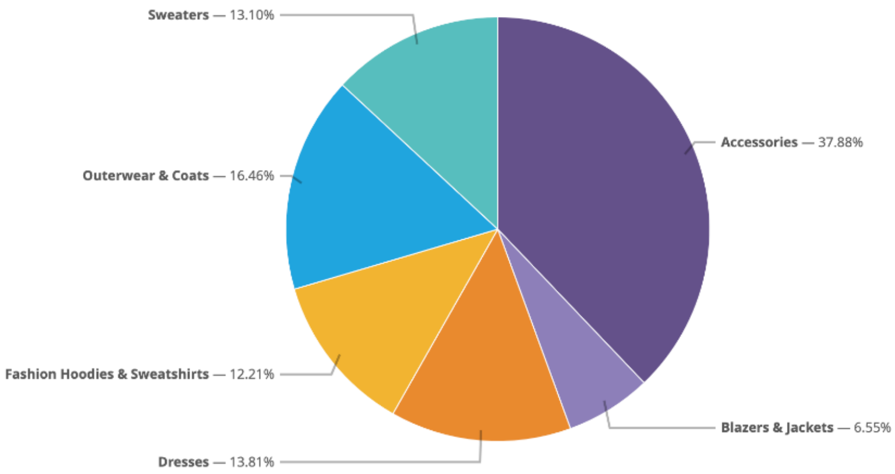
c. [9]Line Graph: like scatter plot but also has a line connecting the dots



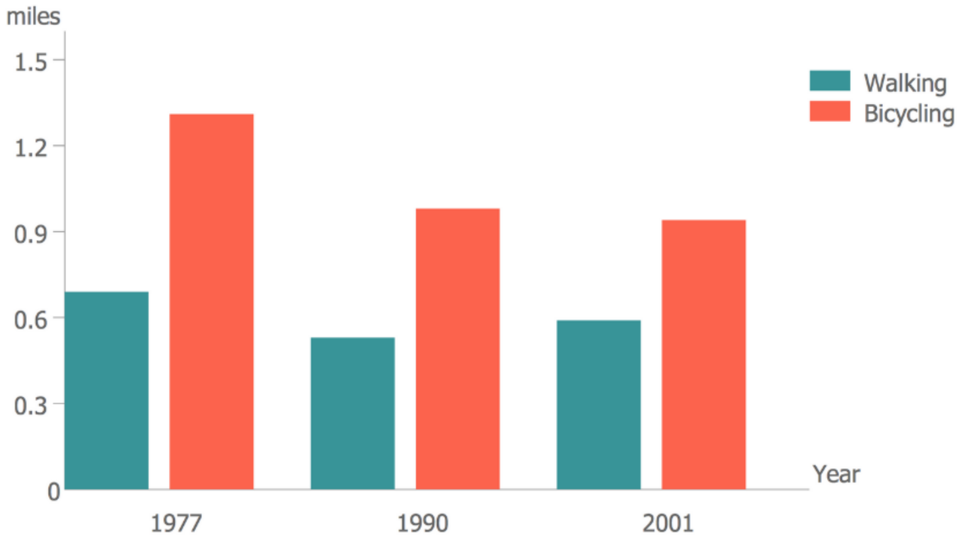
[8] https://www.mathsisfun.com/data/scatter-xy-plots.html

[9] https://www.perkins.org/resource/activity-creating-line-charts-yahoo-finance-stock-market-data/

d. [10]Pie Chart: shows proportion of one variable (usually in %) compared to other variables



e. [11]Grouped Bar Chart: shows the changes in each variable, can also be used to compare with changes in other variables

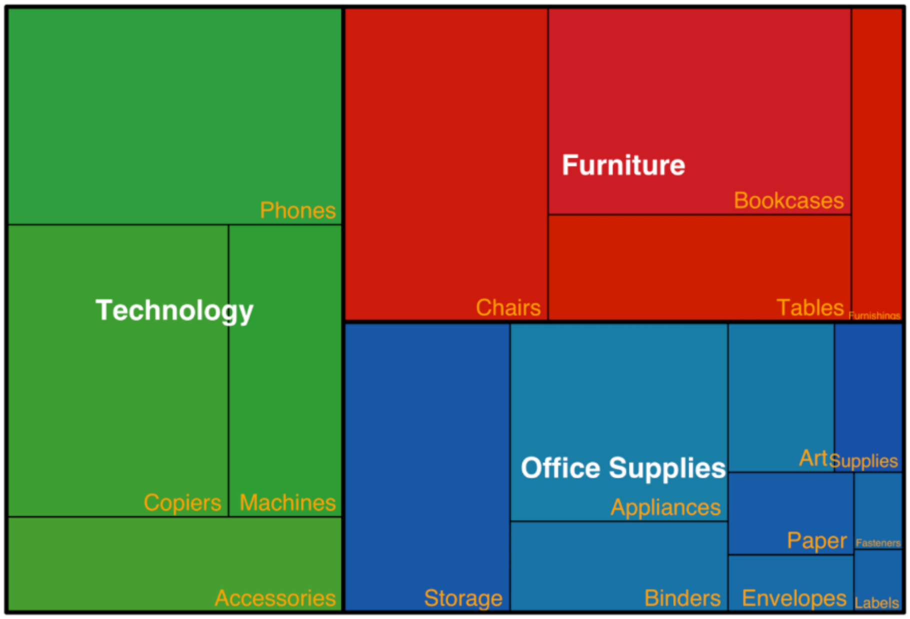


Average trip length (in miles) among U.S. children aged five to 15 years.
Author's analysis based on data from National Personal Transportation Surveys
for 1977 and 1990 and the National Household Travel Survey for 2001.

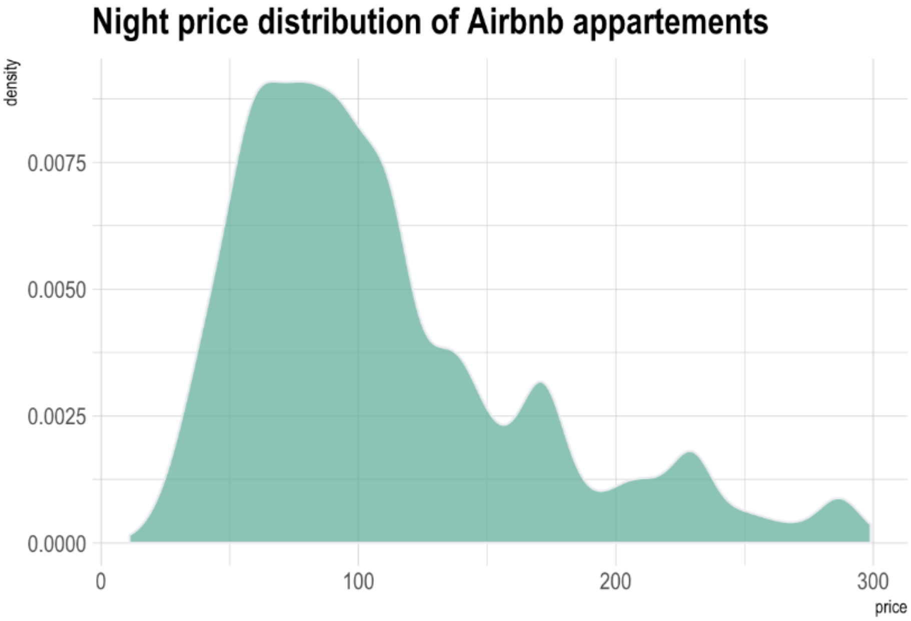
[10] https://cloud.google.com/looker/docs/reference/param-lookml-dashboard-pie-chart

[11] https://www.conceptdraw.com/solution-park/charts-bar-graphs

f. [12]Tree Map: also shows proportions, can be customized for more complex visualizations



g. [13]Density Plot: shows the distribution of the variable(s)



[12] https://exploratory.io/note/BWz1Bar4JF/How-to-create-a-Treemap-in-Exploratory-uMS0CNW3rZ
[13] https://www.data-to-viz.com/graph/density.html

Now that you are familiar with the frequently used graphs, let us analyze more examples. So far, all the cases have been effective use cases of visualizations but what happens when they are incorrectly utilized? There are plenty of real-life instances were graphs are misleading, sometimes on purpose. The figure above is from the time when Barack Obama was still president. Look at the line graph from Fox News in the below figure and you will notice something is off. The 8.6% unemployment rate in November is not correctly plotted! The graph on the left from *StatisticsHowTo* on the right shows what the graph *should* look like.[14] You can see that there is a drop-in unemployment rate that is not previously noticeable in the original graph. Apart from this, you can also argue that the aesthetics of the line graph from Fox News is far from perfect with the y-axis scale not truly clear in addition to the grid lines. The COVID-19 pandemic gave another opportunity for people to incorporate graphs that did not adhere to the fundamentals of data visualization. One such case was from *C5N,* an Argentine television news company. There are two bar graphs, the first from C5N and a corrected one from the article by Nikita Kotsehub.[15]

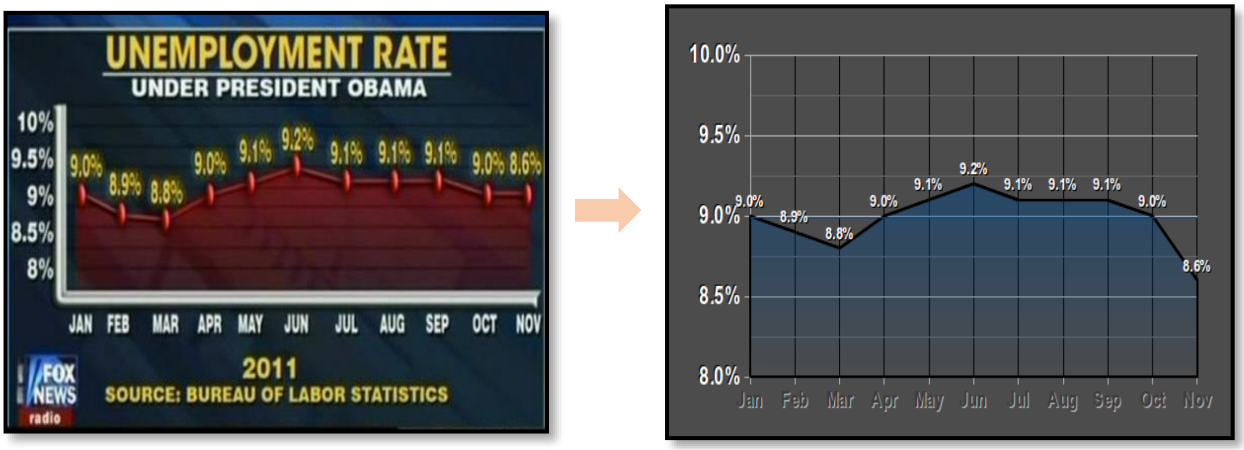


Figure 7.0 – Original line graph (left) and correct line graph (right)

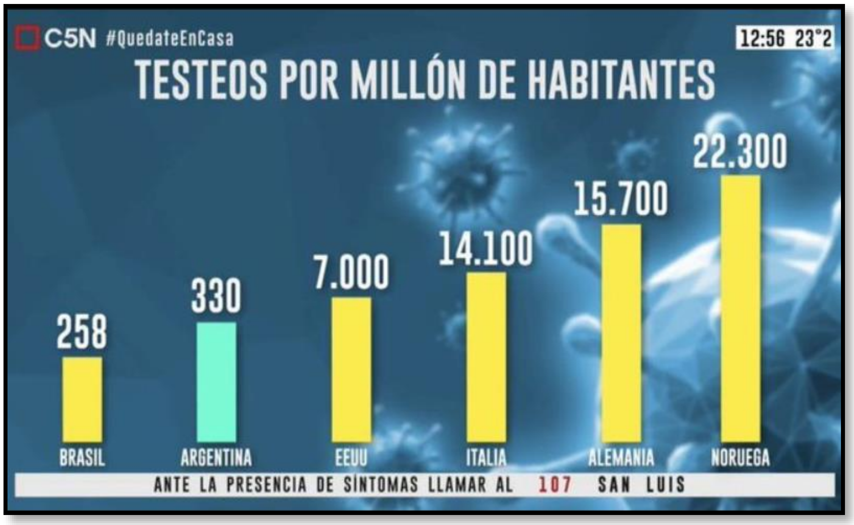


Figure 8.0 – Original bar graph from C5N

[14] https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/

[15] https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

Remember, the main use case of bar graphs to show amounts of a variable, which in this case is countries. However, when you take a closer look at the top bar chart, you will notice how the bar for Argentina is so close to EEUU (United States), yet the numbers (330 vs 7000) are wildly different. That is a difference of more than twenty times! The bars certainly do not show this and creates an illusion that Argentina is keeping up with other countries such as US, Germany, or Norway. To be fair, they also misrepresented the US numbers compared to Norway which is over three times higher. The corrected bar chart above is an accurate data visualization. The data has not changed at all, but you can see just how far behind Brazil and Argentina were in terms of COVID-19 testing. Of course, this graph would not be too appealing to someone watching news in their home in Buenos Aires during the pandemic, hence misleading the audience. Now, for both examples, you could decipher what the data is telling you without the need to focus on the graphs, but it is a lot more difficult. The illusion created by these misleading graphs are dangerous and can create panic and mistrust, which are completely avoidable.

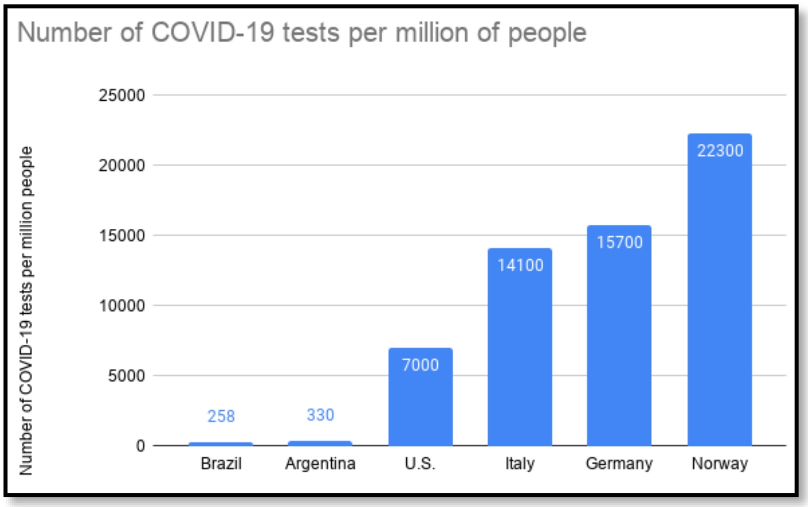


Figure 9.0 – Correct bar graph of COVID-19 testing

In conclusion, I hope this article will help you look at data visualization in a new light. While it has been around for thousands of years, it is ever evolving and becoming increasingly important. Knowing the basics can help you correctly analyze information around you and even help in your professional goals. Knowledge has been creatively visualized by humans for a long time and data visualization today ensures we carry that forward into the future.