DATA VISUALIZATION

IS AN ART OR SCIENCE?

[1]Preetham Dhonekeni, [2]Sharvari Shukla, [3]Paridhi Gupta

[1]PG Student, [2]Director and Professor, [3]PG Student
[1]Applied Statistics,
[1]Symbiosis Statistical Institute, Symbiosis International (Deemed University), Pune, India

Abstract: "A picture is worth 1000 words" is an idiom in English language which says how effective a visualized picture when compared to words is; data visualization is widely used in every possible field. The ability to visualize data is crucial to scientific research. As visualization has become the key for presenting the facts, there is major chance for misleading the facts which may lead to disastrous results. This paper consists of the major reason behind how the data is misled by various visualization techniques. This paper also includes importance of virtual literacy and the role of a statistician in visualizing the complex data.

1. INTRODUCTION

The act or process of interpreting in visual terms or putting the data into visible form is said to be visualization. Humans have natural ability to visualize everything that comes to the mind, which helps in better understanding of the concept. The consumer interest in visual content isn't necessarily just a preference; it's actually easier and faster for humans to process. The right picture can go further than just telling your story visually. It enables you to highlight the most relevant conclusions from what would otherwise be considered a huge pile of documents. The three main goals of visualization are to explore the given data by using various interactive visualization, to present the data using appropriate techniques to easily communicate the results and to analyze the data and to draw the required intuitions from it. The concept of visualization in nothing new but existed in earlier stages of man when there is no proper communication medium, later during ages it slowly taken shape of graphs and charts as the revolution took place.

2. TYPES OF VISUALIZATION

Visualization can be classified into 3 types

1. **Scientific visualization:** It is concerned with database which manages both time and space information. Data is described to be continuous, however its representation is discrete. It displays data of scientific experiments generated from physical process. Now a day's most of this visualization is in 3 dimensional and related to various fields like physics, chemistry, geology, etc.

2. **Statistical visualization:** This type of visualization is mostly concerned with data sets. Visualizing the given data using various techniques which includes different types of graphs or charts of various dimensions will help us in better understanding of the data. And makes drawing the required conclusions or intuitions from the data easier when compared to the data in tabular form or numerical form.

3. Infographics: Infographics is storytelling using attractive figures. It is basically representation of processed data in a way that people can easily understand the facts mentioned. There are no limits for making an infographic we must be creative enough to include all the facts in a single infographic. This is the most effective way of presenting the data when we compare the rest two ways of visualization. These are majorly used in financial, marketing, political representation of facts etc.

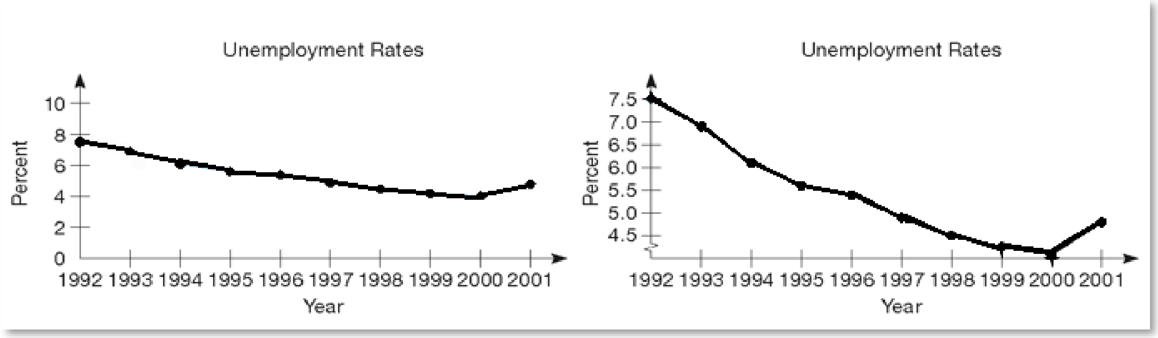
In this paper we are majorly concerned with statistical visualization, where the data is to be visualized to obtain or to learn behavior of the data which sometimes mislead due to various reasons which are being discussed further.

3. WHAT IS "MISLEADING OF DATA"?

In the modern age, data is hugely generated in across all domains. People are more interested in visualized data than numerical data, which may be cumbersome and difficult to interpret. This might lead data being easily misled. Misleading of data visualization is incorrect representation of data which may lead to wrong analysis of the data and causes the viewers to believe in results which are false or partially true. There are certain standard practices that applies the usage of mapping of data visualization techniques. Not following or deviating from these standard practices can render the chart or virtually meaningless. For example, a pie chart when includes segments of data with assigned values should add up to 100, in some misled pie charts to eliminates the influence of one segment, it is not projected in pie charts where the summation won't be 100.

There are basically four reasons for "why the data is misleading?"

- Intentionally misled.



- Unintentionally misled.
- Graph description
- Graphicacy

Lack of knowledge about the methodology of the given data also lead to making of inappropriate graph and the mapping software gives the designer the flexibility to manipulate the data and to present it in most favorable way.

3.1 Intentionally mislead

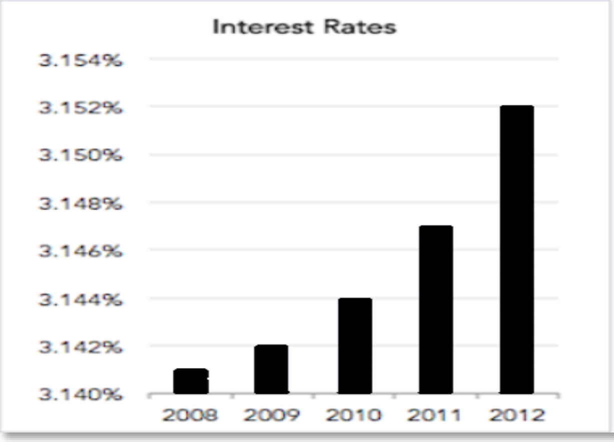
"Lies damned lies and statistics" is part of phrase attributed to Benjamin Disraeli. This statement refers to the power of numbers, the use of statistics to strengthen the weak arguments, and how effectively can statistic representation influence the peoples thought process. Various representations of data like graphs, charts, infographics doesn't lie, they are just tools in the hands of their designer. But they can lead the readers astray.

Now, why do these designers lie? There may be many reasons for which all the answers are unethical like a company wants to depict its wrong picture of growth or a politician wants to show his inexistent support in the elections and many more.

Now comes the question, how do they lie with visualization? In this paper I will discuss some examples how the graphs or charts mislead.

3.1.1 Manipulated axis

As we all know a graph (2 dimensional) has two axes namely X-axis and Y-axis. The values on these axes are manipulated to make these truncated graphs show the data which is not very significant as something which makes larger difference. This can be done by omitting the base line or starting y axis from value which is greater than zero. In a bar graph this has an effect of skewing the visual comparison in such a way that improperly emphasizes the difference between the bars which may lead to misinterpretation by the viewer. The below graph shows the information about the interest rates during 2008-2012, the graph on the left shows that there is drastic increase in interest rate from 2008-2012, but it can be seen that an actual change is from the graph on the left, this can be a good example for misinterpretation which can happen with the non-zero baseline.



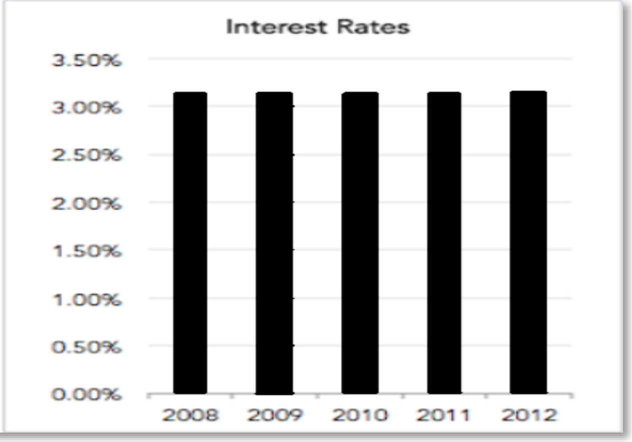


Figure 1 source: google images
Figure 2 source: google images

Manipulating the axis also includes change in the values on the axis that is change in the scale of the graph. By increasing or decreasing the difference between the values of y axis, one can show the minute change as a drastic change of the data which will affect the outcome or vice versa.

The above picture shows unemployment rates during the year 1992-2001 the graph on the left shows the original rate of change and the graph on the right is misleading and here we can see the change in the values on y axis and we can see how our impression on the decrease rate of unemployment can change with manipulating the values on the y axis.

3.1.2 Using inappropriate graphs

There are different kinds of data to be visualized and according to the type of the data there are respective visualizing techniques which should be used to depict the data in the best way. For example, if u want to observe the distribution of the data or to check the range, normal tendency or outlier we mostly use scatterplot, line chart. If we need to compare two variables, we use bar graph, box plot. Each graph has unique way of representation of data for better understanding of data. Using an inappropriate or a wrong graph is delivering misinformation that can happen through sheer incompetence. This involves picking a graph which doesn't present your graph exactly. It's because of people misconception that every graph can be used for all types of data. For example, the below graphs show students in A, B and C classes interested in 5 different games. We can observe that pie charts aren't properly showing difference when compared to bar graphs which differentiates between sizes of each section of students.

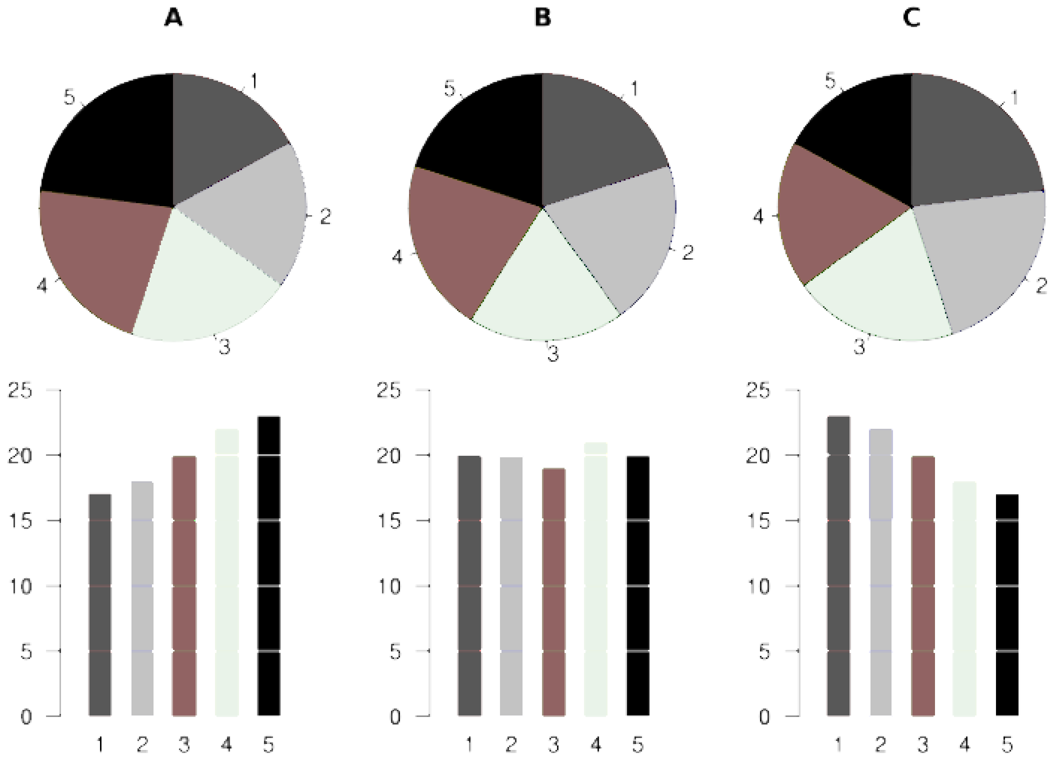


Figure 2. Source: conversionxl.com.

3.1.3 Going against convention

There is some convention which humans naturally made for example, a part in graph with higher color complexion will represent the higher quantity values and green shows the positivity while red shows the negativity. Now this may be a tool in the hands of designer to mislead the information by just reversing convection where the viewer will completely understand the data in opposite way. Choosing a proper color complexion is also important.

In figure 3, we can also notice how inappropriately the sections of the bar graph and the pie chart are colored.

3.2 Unintentionally misled.

Now according to Alberto Cairo, a Spanish data journalist mislead is different from lie. Misleading can sometimes happen unintentionally, there may be different reasons and one such reason is lack of complete data. Designers of the graphs sometimes neglect the outliers or the data which is missing and create a graph with allotted data forgetting the fact that this can affect the result. Result may be true for the data provided but it may sometimes vary from the real-life situation. Another point where the graphs are misled unintentionally is due to the malfunctioning of the software. If we merely focus on the graphs which the software produces as the result and neglect the numbers, we can never find out whether the given software is functioning properly or the given graphs are appropriate.

3.3 Graph description

Most of the time in the print media or the social media the description given about the graph will be nowhere or partially related to the given graph. These published graphs mostly consider averages of the data and averages can lie for example, we are considering average income of a specific group which will change if one super rich person enters into the group, so this average doesn't properly represent the given set of data. So, average is not always the best way to represent an entire data. This kind of misleading by improper description of graph can come under "unintentionally misled" if the publisher or the designer doesn't have proper idea about what is being visualized. Or it may also come under "intentionally misled" if the publisher or the designer wants to deceive the viewers.

3.4 Graphicacy

Graphicacy is one of the four communications (oracy, literacy, numeracy, graphicacy) which are essential for each individual in the present reality. Graphicacy has a dictionary meaning of understanding, using and generating graphic images. In this paper we consider the part of an individual's ability to understand the graphical representations and drawing interpretations from the graphically represented data. In misleading of data visualization lack of graphicacy among the people have an equal role to play compared to the mistakes made by the designer of the graphs. Consider a designer has made a visualization of a given data following all the ethics and the standards, but still it is misleading we can better say it as misinterpreted or misunderstood by few viewers who lack graphicacy and spread with same conception.

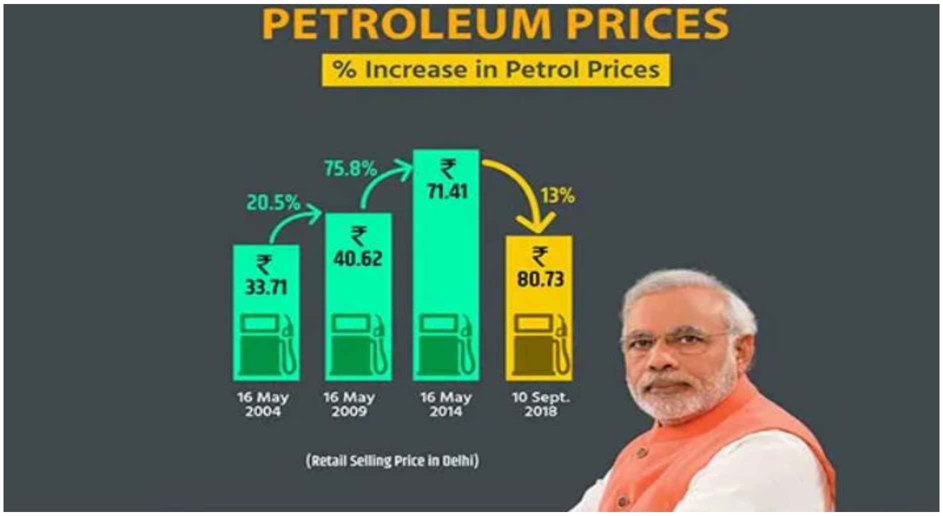


Figure 4, source: ndtv.com

The above picture shows the information about the rate of increase in petrol prices during the UPA government and BJP rule. In the above picture the rate of increase of the price to be considered instead the rate of petrol is taken into the consideration and this infographic was stated wrong and was trolled in the social media. This will be classical example which shows the lack of graphicacy among the common people. PW Wilmot in his paper of "Graphicacy as a form of communication" mentioned that knowledge of graphicacy should be developed from the initial stages of the education keeping in view the growing importance of graphicacy. So that every individual has some minimum knowledge of graphical literacy and can interpret and spread the visualized data without any misconception. Research is going on in the western countries about measuring the graphicacy among people belonging to various classes of the society, people with various educational background.

4. Role of a statistician.

A statistician is one who has a core knowledge of statistical tools, and who also prefers to have the domain knowledge from where the data is generated, so that he can get deep understanding of the data and when he/she analyses the data and makes visualizations (graphs) from the interpretation, it will be more accurate or appropriate than the visualizations made by the people who has knowledge of the software by which the graphs are made but not the data. A statistician in the role of viewer is also in better position when compared to data engineers because of high rate of graphical literacy. If a designer who only knows how to operate a software is never qualified of making graphs because he/she only has a software knowledge, where a software is just a mechanical tool which follows the commands given to and gives a result if the commands are perfect, and in some cases, if the designer thinks the output is correct and publishes it without verifying it then this will mislead the viewers. This is where the statisticians can play their role as they have both the software knowledge and the graphical knowledge. He/she can check if the graphs are properly following all the required properties without any concern regarding the outcome. A statistician plays an important role in interpreting data as they are trained to deal with different type of graphs. A person who can interpret better can detect the mistakes better, which actually helps them in making appropriate graphs.

5. Discussion and Conclusion

Let's get back to the question "Is data visualization an art or science?" An art is something out of creativity and can be done by any person. But data visualization requires domain knowledge and the knowledge of the statistics. So, it can be concluded by saying that you need to have both, the creativity to create art of graphs and the science behind it. Most importantly the science behind it without which the graphs is meaningless. The graphs should be attractive and simpler so that it catches the viewers' attention and makes the interpretation easier for them. While a designer of a graph should also possess some good knowledge about the domain of the data which is being used to make the graph, proper methodology of the respective graph. The viewers and the designer should maintain some ethics of not to mislead the data by making inappropriate graphs, having some graphicacy and not popularizing inappropriate graphs. If misled is done intentionally, then there is nothing that no one can do.

References:

[1]. Christopher Cabanski, Houston Gilbert, Sofia Mosesova, 2018 Can graphs tell lies? A tutorial on how to visualize the data. Citation: clin transl sci.
[2]. Stephen few, January 2007, Data visualization past, present and the future, journal of Perpetual edge.
[3]. Edison Zangiacomi Martinez, 2015, Description of continuous data using bar graphs a misleading approach, Journal of the Brazilian society of Tropical medicine.
[4]. Irena Bolks and Mojca Bavdax, User aspects of data visualization in official aspects, university of Ljublina.
[5].P.D Wilmot, June 1999, Graphical as form of communication, South Africa Geographical Journal.
[6]. Chun-houh-chun, Wolfgen Hardle, Antony Unwin, Handbook of data visualization.
[7]. Adebowale E. Shadre, Cajetan AKujuobi, 2016, Data visualization, journal Researchgate.
[8]. Jesse Anderson, 2018, Data Engineers vs Data scientist , journal of data science.
[9]. Darrell Huff, How to lie with Statistics.
[10]. Alex Birkett, 2018, How to avoid being deceived by data.