# A Reader on Data Visualization

*MSIS 2629 Spring 2018*

*2018-04-30*

# Contents

# Chapter 1

# Preface

This is a collaborative writing project as part of the course MSIS 2629 "Data Visualization" at Santa Clara University. The purpose of the class reader is to collaboratively engage with and reflect on data visualizations, to establish a solid theoretical background, and to collect useful practices and showcases. More information on the background of this project is available in the syllabus.

The following text serves explains how we organize ourselves.

## 1.1 References

**EVERY** references must be included in the `book.bib` file. This file uses the bibtex notation (Learn how to use bibtex here.). Most literature search engines allow you to export the reference information in Bibtex. For websites we use the following minimal notation (you may add further information - usually the more the better is a good strategy):

```
@misc{great_viz,
  author = {{A great visualizer}},
  year = {1982},
  title = {A ficticious web page title},
  howpublished = {\url{http://great_viz_org/}},
  note = {Accessed: 2018-04-26}
}
```

Particularly important is the `note` field. Websites change frequently, so links will break. If we do this correctly, `[@great_viz]` will produce (A great visualizer, 1982).

## 1.2 Images

Images should not be loaded from external website because the links may change. Instead download a version of the image and create a reference that contains the link to the image. For example the following image is a deceptive visualization (the bars do start at zero).

Source: (Halper, 2012) referenced in (Andalde, 2014)

The citation for the image looks like this.

```
@misc{halper_2012,
  author={Halper, Daniel},
  year={2012},
```

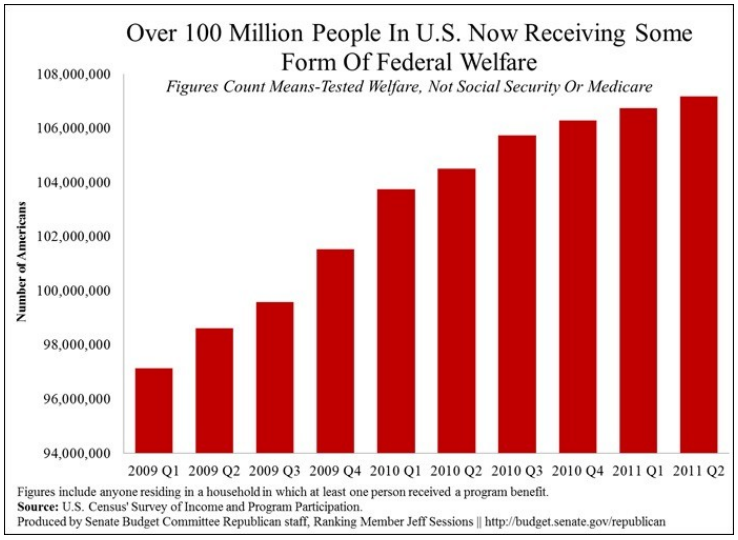



Figure 1.1: An Example of a deceptive visualization

```
  title = {Over 100 Million Now Receiving Federal Welfare},
  url={https://www.weeklystandard.com/daniel-halper/over-100-million-now-receiving-federal-welfare},
  note = {Accessed: 2018-04-26}
}
```

You have probably found this image through a different website that explains the visualization. For example the following website explains some problematic aspects of this visualization:

```
@misc{andale_2014,
  author={Andalde, Stephanie},
  year={2014},
  title = {Misleading Graphs: Real Life Examples},
  url={http://www.statisticshowto.com/misleading-graphs/},
  note = {Accessed: 2018-04-26}
}
```

# Chapter 2

# Introduction

**Data Visualization** Data visualization refers to representing data in a visual context to help people understand the significance of that data. A way so that information, numbers, and measurements makes sense is a form of art – the art of data visualization. Graphs do that for us.Below is a link for different types of plots used in data visualization.

Plot links: https://datavizcatalogue.com/search.html

Infogram helps a user make different types of plots and learn the art of visualization. Below is the link: https://infogram.com/page/data-visualization

Some useful links are mentioned below:

https://eagereyes.org/ https://research.tableau.com/user/robert-kosara https://twitter.com/eagereyes?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor http://flowingdata.com/

**Why is data visualization important?**

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**. https://research.tableau.com/user/robert-kosara https://twitter.com/eagereyes?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

https://data-visualization.cioreview.com/cxoinsight/what-is-data-visualization-and-why-is-it-important-nid-11806-cid-163.html

The article , written by Chris Pittenturf, VP-Data & Analytics, Palace Sports & Entertainment, talks about what data visualization is and its importance to the businesses today. The article begins with a definition of data visualization in simple terms and goes on to explain how a good data visualization should be visually engaging to the reader. Chris goes on to explain the basic criterias that a data visualization should satisfy to be an effective visualization. These criterias and their brief meanings are as follows: 1. Informative: The visualization should be able to convey the information of the data to the reader 2. Efficient: The visualization should not be ambiguous. 3. Appealing: The visualization should be captivating and visually pleasing. 4. (Optional) Interactive and Predictive: The visualizations can contain variables and filters for the users to interact with the visualizations in order to predict results of different scenarios.

Chris goes on to give various day-to-day examples where visualization gives a better understanding of the data. One extremely simple example used by Chris is that of an energy bill. Chris states that as a consumer, when we receive an energy bill, we normally look at the graph in the bill first before proceeding to read the text in the bill. Chris states that consumers are more likely to analyze and understand the visualizations before reading further along. The article ends with Chris emphasizing the importance of data visualizations in our businesses as well as in our daily lives. According to me, the article gives a simple, short and crisp understanding of what data visualization is and how it is relevant to everyone. It shows that data visualization

is an aid to get a better understanding of the complex insights that any business data provides. Most of the data used by the businesses is highly unstructured and these businesses can get a better understanding of their businesses by visualizing their data.

https://www.interaction-design.org/literature/article/information-visualization-a-brief-introduction This article is a brief introduction to Information Visualization. It explains briefly how information visualization helps to make sense of data, how it helps to find relationships between data and confirm ideas.

About David McCandless's TED talk on data visualization: https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization

Visuals help us understand concepts that would otherwise be difficult to contextualize—for example, expenditures or valuations of extremely large amounts of money are represented in the billion dollar-o-gram by color-coded, relatively-sized boxes. Furthermore, it allows synthesis of a breadth of information to be delivered in a small, easily-digestible, aesthetically pleasing way. Visuals serve as a sort of map for a vast landscape of information—they direct your eyes to the important places and details. And the eye, as McCandless notes, is uniquely suited among our senses to process large amounts of information and detect patterns.

The billion dollar-o-gram is extremely readable and rather pretty, but it seems a bit dubious to compare the predicted Iraq War cost to the "mushroomed" actual cost of Iraq and Afghanistan wars, since its purpose seems only to conflate two wars for dramatic effect.

Beyond its ability to make information from several different sources and in large amounts more quickly and easily understood, data visualization can also reveal smaller interesting patterns—allowing us to play the "data detective" as McCandless calls it. In other words, as we have already discussed, data visualization can not only be extremely effective in a declarative manner, but can also be used as an exploratory tool.

McCandless also postulates that we all have a latent "design literacy" that is being developed every day as we are constantly bombarded with visuals, and that our minds and our eyes are taking in this information and processing it so that we all have an intuitive sense of design, and have actually begun to demand a visual aspect to our information. This is an interesting perspective, since everyone does seem to have a sense of visual aspects—space, color, etc., but of course the time-honored adage tells us that beauty is in the eye of the beholder. So while it might be whimsical to claim that we are all designers, there is still, of course, great value in learning formal principles of design.

Basic Guidelines: Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

**Key Figures in the History of Data Visualization**

**Reference**

Infogram,Jun 14,2016.Key Figures in the History of Data Visualization,Medium. https://medium.com/@Infogram/key-figures-in-the-history-of-data-visualization-30486681844c

The history of data visualization is full of incredible stories marked by major events, led by a few key players. This article introduces some of the amazing men and women who paved the way by combining art, science, and statistics.And one of them is Charles Joseph Minard whose most famous work is the map of Napoleon's Russian campaign of 1812 displayed in our class. I present several figures names with their famous works and you can find other stories in the article.

**William Playfair (1759–1823)**

William Playfair is considered the father of statistical graphics, having invented the line and bar chart we use so often today. He is also credited with having created the area and pie chart. Playfair was a Scottish engineer and political economist who published The Commercial and Political Atlas in 1786.

This book featured a variety of graphs including the image below. In this famous example, he compares exports from England with imports into England from Denmark and Norway from 1700 to 1780.

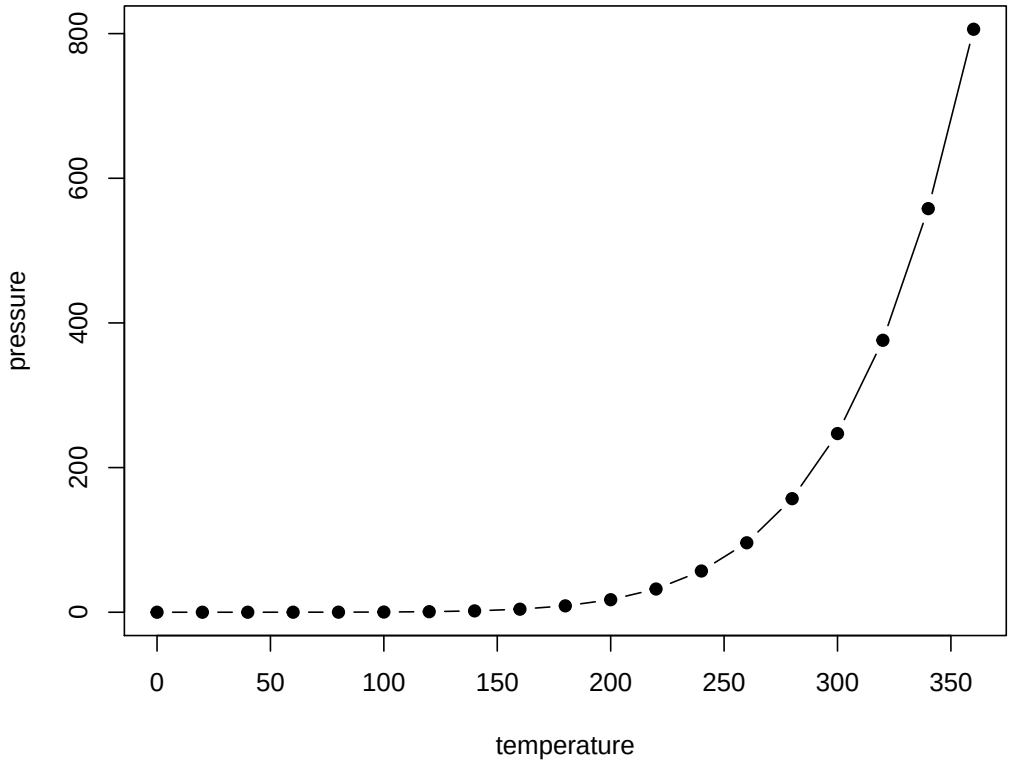



Figure 2.1: Here is a nice figure!

**John Snow (1813–1858)**

In 1854, a cholera epidemic spread quickly through Soho in London. The Broad Street area had seen over 600 dead, and the remaining residents and business owners had largely fled the terrible disease.

Physician John Snow plotted the locations of cholera deaths on a map. The surviving maps of his work show a method of tallying the death counts, drawn as lines parallel to the street, at the appropriate addresses. Snow's research revealed a pattern. He saw a clear concentration around the water pump on Broad Street, helping to find the cause of the infection.

**Charles Joseph Minard (1781–1870)**

Charles Joseph Minard was a French civil engineer famous for his representation of numerical data on maps. His most famous work is the map of Napoleon's Russian campaign of 1812 displaying the dramatic loss of his army over the advance on Moscow and the following retreat.

You can see how many soldiers are still marching and how many died. Drawn in 1869, it is described by many as the best statistical graphic ever drawn. It represents the earliest beginnings of data journalism.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample

Table 2.1: Here is a nice table!

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 4.3 | 3.0 | 1.1 | 0.1 | setosa |
| 5.8 | 4.0 | 1.2 | 0.2 | setosa |
| 5.7 | 4.4 | 1.5 | 0.4 | setosa |
| 5.4 | 3.9 | 1.3 | 0.4 | setosa |
| 5.1 | 3.5 | 1.4 | 0.3 | setosa |
| 5.7 | 3.8 | 1.7 | 0.3 | setosa |
| 5.1 | 3.8 | 1.5 | 0.3 | setosa |

book, which was built on top of R Markdown and **knitr** (Xie, 2015).

# Chapter 3

# Fundamentals

https://www.educba.com/data-mining-vs-data-visualization/

This article gives me a clear understanding of data mining and data visualization.

https://www.educba.com/data-mining-vs-data-visualization/

This article gives me a clear understanding of data mining and data visualization.

In Data Mining, there are different processes involve carrying out the data mining process such as data extraction, data management, data transformations, data pre-processing, etc. In Data Visualization, the primary goal is to convey the information efficiently and clearly without any deviations or complexities in the form of statistical graphs, information graphs, and plots. Also, the author listed the top 7 comparisons between data mining and data visualization, and 12 key differences between data mining and data visualization. After reading the article, you will have a very clear understanding of what are data mining and data visualization and the characters for those two techniques.

In Data Mining, there are different processes involve carrying out the data mining process such as data extraction, data management, data transformations, data pre-processing, etc. In Data Visualization, the primary goal is to convey the information efficiently and clearly without any deviations or complexities in the form of statistical graphs, information graphs, and plots. Also, the author listed the top 7 comparisons between data mining and data visualization, and 12 key differences between data mining and data visualization. After reading the article, you will have a very clear understanding of what are data mining and data visualization and the characters for those two techniques.

- Theoretical background of data visualization

History Data visualization has comes a long way. Prior to the 17th century, data visualization already exists. Though display in other format such as maps, the content are much similar to today's visualization, which mostly presented geologic, economic, and medical data. Here is useful link: http://www.dashboardinsight.com/news/news-articles/the-history-of-data-visualization.aspx

Current research: Deceptive visualizations Data visualization is a powerful communication tool to support arguments with numbers in a way that is accessible and engaging. More people than ever before are making their own charts and infographics, which is presenting a unique problem. Despite the availability of some great charting resources, we are witnessing an influx of poorly-designed misleading or downright deceptive data visualizations. Here are useful links: https://medium.com/@Infogram/study-asks-how-deceptive-are-deceptive-visualizations-8ff52fd81239 https://www.datapine.com/blog/misleading-data-visualization-examples/

## 3.1 1. A Brief History of Data Visualization

Michael Friendly,2006,A Brief History of Data Visualization,York University.http://www.datavis.ca/papers/hbook.pdf

The only new thing in the world is the history you don't know. - Harry S Truman

This paper provides an overview of the intellectual history of data visualization from medieval to mode describing and illustrating some significant advances along the way.

1. Data Visualization: modern product?

   It is common to think of statistical graphics and data visualization as relatively modern developments in statistics. In fact, the graphic representation of quantitative information has deep roots.These roots reach into the histories of the earliest map-making and visual depiction, and later into thematic cartography, statistics and statistical graphics, medicine, and other fields.

   Developments in technologies (printing, reproduction) mathematical theory and practice, and empirical observation and
   recording, enabled the wider use of graphics and new advances in form and content.

2. Milestones Tour

   2.1 Pre-17th Century: Early maps and diagrams

   The earliest seeds of visualization arose in geometric diagrams, in tables of the positions of st celestial bodies, and in the making of maps to aid in navigation and exploration.

   2.2 1600-1699: Measurement and theory

   Among the most important problems of the 17th century were those concerned with physical measurem of time,
   distance,and space- for astronomy, surveying, map making, navigation and territorial expansion. T saw great new growth in theory and the dawnof practical application.

   2.3 1700-1799: New graphic forms

   With some rudiments of statistical theory, data of interest and importance, and the idea of graph at least somewhat established, the 18th century witnessed the expansion of these aspects to new c graphic forms.

   2.4 1800-1850: Beginnings of modern graphics

   With the fertilization provided by the previous innovations of design and technique, the first ha century witnessed explosive growth in statistical graphics and thematic mapping, at a rate which equalled until modern times.

   2.5 1850–1900: The Golden Age of statistical graphics

   By the mid-1800s, all the conditions for the rapid growth of visualization had been established- a "perfect storm"
   for data graphics. Official state statistical offices were established throughout Europe, in reco growing importance of numerical information for social planning,industrialization, commerce, and

   2.5.1 Escaping flatland
   2.5.2 Graphical innovations
   2.5.3 Galton's contributions
   2.5.4 Statistical Atlases

   2.6 1900-1950: The modern dark ages

```
If the late 1800s were the "golden age" of statistical graphics and thematic cartography, the ear
called the "modern dark ages" of visualization. There were few graphical innovations, and, by the
enthusiasm for visualization which characterized the late 1800s had been supplanted by the rise o
and formal, often statistical, models in the social sciences.
```

2.7 1950–1975: Re-birth of data visualization

```
Still under the influence of the formal and numerical zeitgeist from the mid-1930s on, data visua
rise from dormancy in the mid 1960s.
```

2.8 1975–present: High-D, interactive and dynamic data visualization

```
During the last quarter of the 20th century data visualization has blossomed into a mature, vibra
disciplinary research area, as may be seen in this Handbook, and software tools for a wide range
methods and data types are available for every desktop computer.
```

## 3.2 2. Fundamental Components of Design

Artists use balance, emphasis, movement, pattern, repetition, proportion, rhythm, variety, and unity as the design foundation of any work. If you want to take your data visualization from an everyday dashboard to a compelling data story, incorporate the 9 principles of design from graphic designer Melissa Anderson's article: https://www.idashboards.com/blog/2017/07/26/data-visualization-and-the-9-fundamental-design-principles/

Balance doesn't mean that each side of the visualization needs perfect symmetry, but it is important to have the elements of the dashboard/visualiaztion distributed evenly. And it important to remember the non-data elements, such as a logo, title, caption, etc., that can affect the balance of the display.

Another closely related component to balance is variety which could seem counter to balance, but when done correctly, variety can help increase the recall of information. However if overdone, too much variety can feel cluttered and blur together the images and data in the mind of the viewer.

Arguably the most critical of the components is proportion. Proportion can be subtle but it can go a long way to enhancing a viewer's experience and understanding of the data. The danger of proportion though is that it can be easy to deceive people subconsciously. Naturally images will have a greater impact on how our brains perceive the dashboard or visualization. For example, someone can change the scale of a graph or images to inflate their results and even if they write the numbers next to it, the shortcut many people will take is to interpret the data based on the image. This is why it is important we take care to accurately reflect proportion in our data visualization and remain critical of how others use proportion in their visualization.

Emphasis was the component that I most related to when reading through the nine principles of design in this article. From prior experience with art through photography I understand it is key to be concious of what I am drawing the viewers attention to in my art. When thinking about the art design of data visualization it is also very important to remain keen on the main point of your story and how the entire visualization is either drawing the viewer to that point of emphasis or how they are being distracted or drawn elsewhere.

## 3.3 3. Guide to Best Practices in Data Visualization

These are the best practices of data visualization. Anticipate in advance what kind of questions the viewers will ask and then focus your visualization with respect to those questions.

The brain processes stimuli from our environment to process what is important in 2 ways – unconscious (System 1 represents uncontrolled functions such as facial expressions, reactions) and conscious (System 2 – represents controlled function such as solving math problems). Data Visualization leverage attributes of System -1 which can have a quick and correct impact in a most efficient manner. The three best practices of data visualization are as follows:

** 1. Design and layout matter **The design and layout should facilitate ease of understanding to convey your message to the viewer.** 2. Avoid Clutter **Keep it simple. To implement this always keep into account the data-ink ratio – the ratio of ink required to convey the intended meaning to the total amount of ink used in the table or chart should be as close to 1 as possible. That means, avoid ink which do not add any information.** 3. Use color purposely and effectively ** Use of color may be prettier and attractive but can be distractive too. Thus, color should be used only if it assists in conveying your message. The above three principles are illustrated with the help of scenarios and examples which helps to comprehend the topic in more meaningful and practical way in the article. It also gives various advantages of using the above principles.And the above best practices could be applied to all the 3 types of analytics: descriptive, predictive and prescriptive.

**Reference** Jeffrey D. Camm, Michael J. Fry, Jeffrey Shaffer (2017) A Practitioner's Guide to Best Practices in Data Visualization.Interfaces 47(6):473-488. https://doi.org/10.1287/inte.2017.0916

## 3.4 4. Survey of Popular Data Visualization Tools

Due to the rise of big data analytics, there has been an increased need for data visualization tools to help understand the data. Besides Tableau, there are several other software tools one can use for data visualization like Sisense, Plotly, FusionCharts, Highcharts, Datawrapper, and Qlikview. This article is from forbes and has a brief, clear introduction about these 7 powerful software options for data visualization. This could be helpful for future reference because for different purposes I may need to use different tools. Each option has its advantages and disadvantages and this article helps highlight them.

**Tableau** is the most popular of the group and has many users. It is simple to use, making it easy to learn and can handle large datasets. Tableau can handle big data thanks to integration with database handling applications such as MySQL, Hadoop, and Amazon AWS.

**Qlikview** is the main competitor to Tableau and is also quite popular. Qlikview is customizable and has a wide range of features which can be a double-edged sword. These features take more time to learn and get acquianted with. However, once one gets past the learning curve, they have a powerful tool at their disposal.

The distinctive aspect of **FusionCharts** is that graphics do not have to be created from scratch. Users can start with a template and insert their own data from their project.

**Highcharts** proudly claims to be used by 72% of the 100 biggest companies in the world. It is a simple tool that does not require specialized training and quickly generates the desired output. Unlike some tools, Highcharts focuses on cross- browser support, allowing for greater access and use.

**Datawrapper** is making a name for itself in the media industry. It has a simple user interface making it easy to generate charts and embed into reports.

**Plotly** can create more sophisticated visuals thanks to integration with programming languages such as Python and R. The danger is creating something more complicated than necessary. The whole point of data visualization is to quickly and clearly convey information.

**Sisense** can bring together multiple sources of data for easier access. It can even work with large datasets. Sisense makes it easy to share finished products across departments, ensuring everyone can get the information they need.

https://www.forbes.com/sites/bernardmarr/2017/07/20/the-7-best-data-visualization-tools-in-2017/#3a12b8ea6c30

## 3.5 5. Pick the Right Chart Type!

Data divusalization is combining the art and science. As for the art, we can say there are no correct answers for doing the visualization. There are many ways to present the data. However, how to making sense of facts, numbers and measurement for better understanding is still have a logical path to follow.

To determine which kind of chart is hard for those people new to data visulization. Most people learn it by refering some other people's work without understanding the logic behind. So they don't have the theory in their mind to make the judgement. Here , I will introduce some guidance to choose the charts.

When we about to choose the type of chart, we need to answer some questions. - How many features would you like to show in a chart? - how many data points do you want to display for each variable? - Will you display time serious data or among items or groups.

After answered this question, you shoul able to get a better imagenation of your ideal graph. The simple guidance for using different type of chart is line charts for tracking trends over time, bar charts to compare quantities, scatter plots for joint variation of two data items, bubble charts showing joint variation of three data items, and pie charts to compare parts of a whole.

Let's review the most commonly used chart types and expalin what circumstance should better use typical chart and the pros and conts of each type of chart. Before introduce differnt types of charts, you can use the following website to familiar with different types of charts. The Data Visualisation Catalogue

**Type 1 Column Charts.** This should be the most popular chart type. This chart is good to do comparison between different values when specific values are important. TBD

Still have hard time to choose? There are many resources on line can help you do the decision. For example, Dr. Andre Abela create a chart selection diagram that is helpful to pick the right chart depends on the data type. The link of website is http://extremepresentation.typepad.com/blog/2015/01/announcing-the-slide-chooser.html

Reference: Data Visualization – How to Pick the Right Chart Type? , By Jānis Gulbis https://eazybi.com/blog/data_visualization_and_chart_types/

Data Visualization Best Practices by melindasantos | Sep 19, 2017 http://paristech.com/blog/data-visualization-best-practices/

http://paristech.com/blog/data-visualization-best-practices/ http://extremepresentation.typepad.com/blog/2015/01/announcing-the-slide-chooser.html

## 3.6 6. Guide for Developing Dashboards

https://www.klipfolio.com/blog/intuitive-dashboard-design Three rules to follow in order to develop intuitive dashboards:

```
1. the dashboard should read left to right
2. group related information together
3. find relationships between seemingly unrelated areas and display visuals together to show the relati
```

Often a designer can become too concerned with coming up with a visual that is too intricate and overly complicated. A dashboard should be appealing but also easy to understand. Following these rules will lead to effective presentation of the data.

Because we read from top to bottom and left to right, a reader's eyes will naturally look in the upper left of a page. The content should therefore flow like words in a book. It is important to note that the information at the top of the page does not always have to be the most important. Annual data is usually more important to a business but daily or weekly data could be used more often for day to day work. This should be kept in mind when designing a dashboard as dashboards are often used as a quick convenient way to look up data.

Grouping related data together is an intuitive way to help the flow of the visual. It does not make sense for a user to have to search in different areas to find the information they need.

Grouping unrelated data seems contradictory to the second rule, but the important thing is to tell a story not previously observed. Data analytics is all about finding stories the data is trying to tell. Once they are discovered, the stories need to be presented in an effective manner. Grouping unrelated data together makes it easier to see how they change together.

## 3.7 7. Definions of Date Deception and Graphic Integrity

Data visualization becomes more and more pupular to communicate and support arguments nowdays. There are lots of great resources online to create and design amazing data products, in the same time, there are some poorly-designed misleading deceptive data visualizations.

So what does **data deception** mean? Data deception, defined by School of Law at the New York University, as "a graphical depiction of information, designed with or without an intent to deceive, that may create a belief about the message and/or its components, which varies from the actual message."

In reality, decades ago, Edward Tufte already introduced the concept of graphical intergrity in his book and presented six principles of graphic integrity. Here are the principles from book:

```
1. The representation of numbers, as physically measured on the surface of the graphic itself, should b
proportional to the numerical quantities measured.

2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity.
explanations of the data on the graphic itself. Label important events in the data.

3. Show data variation, not design variation.

4. In time-series displays of money, de ated and standardized units of monetary measurement are nearly a
nominal units.

5. The number of information-carrying (variable) dimensions depicted should not exceed the number of di
data.

6.Graphics must not quote data out of context.
```

**Reference** (1) Pandey, A. V., Rall, K., Satterthwaite, M. L., Nov, O., & Bertini, E. (2015). How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques. In CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems: Crossings (Vol. 2015-April, pp. 1469-1478). Association for Computing Machinery. DOI: 10.1145/2702123.2702608 (2) Tufte, E. R., and Graves-Morris, P. The visual display of quantitative information, vol. 2. Graphics press Cheshire, CT, 1983.

**Misleading graphs:**

Misleading graphs or distorted graphs, are graphs created which skews the data, intentionally or unintentionally, resulting in a representation of incorrect conclusions.

There are some ways in which distorted graphs can be created: 1. Improper scaling of y axis: This is one of the classic misleading graphs. Instead of scale starting from zero or a baseline, y axis is scaled conveniently to highlight the differences among bins. 2. Improper labelling of graphs: Lack of labels make the graph hard to interpret for the reader and lead to wrong conclusions. 3. Paired graphs on different scale: It is not a fair comparison if two elements are plotted side-by-side, on a different scale and compared. This makes one graph look better than the other, even when it is not. 4. Dual axis with different scales: If we are plotting two elements on the same graph with different scales, even if the axes are properly labeled, it is assumed

that both axes are on the same scale. 5. Incomplete data: Short-term graphs are made to manipulate the trend, which will not be seen otherwise. Time-series data are cut intentionally to just show a trend within a particular period to create a more favorable visual impression.

Please find the references below. http://hypsypops.com/axes-evil-lie-graphs/ http://www.statisticshowto.com/misleading-graphs/

## 3.8 8. Contemporary Research Results & What's Next

Next Steps for Data Visualization Research

With the development, studies and new tools applied in data visualization, more people understand it matters. But given its youth and interdisciplinary nature, research methods and training in the field of data visualization are still developing. So, we asked ourselves: what steps might help accelerate the development of the field? Based on a group brainstorm and discussion, this article shares some of the proposals of ongoing discussion and experiment with new approaches:

1. Adapting the Publication and Review Process As the article states, "both 'good' and 'bad' reviews could serve as valuable guides", so providing reviewer guidelines could be helpful for fledgling practitioners in the field.
2. Promoting Discussion and Accretion Discussion of research papers actively occurs at conferences, on social media, and within research groups. Much of this discussion is either ephemeral or non-public. So ongoing discussion might explicitly transition to the online forum.
3. Research Methods Training Developing a core curriculum for data visualization research might help both cases, guiding students and instructors alike. For example, recognizing that empirical methods were critical to multiple areas of computer science, Stanford CS faculty organized a new course on Designing Computer Science Experiments(http://sing.stanford.edu/cs303-sp11/). Also, online resources could be reinforced with a catalog of learning resources, ranging from tutorials and self-guided study to online courses. Useful examples include Jake Wobbrock's Practical Statistics for HCI and Pierre Dragicevic's resources for reforming statistical practice.

Reference: https://medium.com/@uwdata/next-steps-for-data-visualization-research-3ef5e1a5e349

## 3.9 Typography and Data Visualization

This article discusses less common applications of typography in data visualization. While data components such as quantitative or categorical data are commonly represented by visual features like colors, sizes or shapes, utilization of boldface, font variation, and other typographic elements in data visualization are less prevalent.

Highlighted in the article are preattentive visual attributes; preattentive attributes are those that perceptual psychologists have determined to be easily recognized by the human brain irrespective of how many items are displayed. Therefore, "preattentive visual attributes are desirable in data visualization as they can demand attention only when a target is present, can be difficult to ignore, and are virtually unaffected by load." Examples of preattentive attributes are size/area, hue, and curvature.

This brings us to the disparateness of the popularity of visual aspects like color and size and typographic aspects such as font variation, capitalization and bold. The authors present several possible reasons for this, beginning with the preattentiveness of visual attributes like size and hue. However, some typographic attributes such as line width or size, intensity, or font weight (a combination of the two) are considered preattentive as well.
Furthermore, these visual attributes are inherently more viscerally powerful, and they are easy to code in a variety of programming languages. Technology has also perhaps previously limited the use of typographic

attributes, for only recently have fine details such as serifs, italics, etc. been made readily visible to the audiences of data visualizations by technological advances.

Lastly, the authors remark that it is possible the lack of variety of typographic elements used in data visualizations is due to the limited knowledge of computer scientists and other individuals pursuing data visualization in how to apply these elements effectively. While the first few proposed explanations make sense from personal experience with technology and exposure to data visualizations and design in general, the hypothesis that lack of knowledge of typographic elements in data visualization seems more plausible if it was being applied to a small group of people rather than all of the data visualization design community. I would say that it is more likely that the use of typographic elements in data visualization is less popular because there are fewer instances in which it can be used appropriately, or a status quo bias—if current visual attributes are received well, the prevailing attitude may be not to fix what is not broken.
However, the authors also point out that despite the dearth of typographic attributes in data visualization, other spheres like typography, cartography, mathematics, chemistry, and programming "have a rich history with type and font attributes that informs the scope of the parameter space."

The authors continue by pointing out some tips for using typographic attributes to encode different data types, since certain attributes may be suited to particular purposes. For example, font weight (size and intensity) is ideal for representing quantitative or ordered data, and font type (shape) is better suited to denote categories in the data.
Furthermore, as in typography and cartography, use of typographic attributes in data visualization raises concerns of legibility, the ability to understand both individual characters and commonalities that identify a font family, and readability, the ability to read lines and blocks of words. Often, interactivity of a visualization will not only improve functionality, but also provide a solution to readability issues by providing a means to zoom in on small text.

There are a few examples of unusual/innovative use of typography for data visualization in the article, not all of which I agree are made more effective by the interesting utilization of typographic attributes, but the "Who Survived the Titanic" visualization's use of typographic attributes allowed it to not only answer macro-questions very quickly, such as if women and children were actually first to be evacuated across classes, but also to provide answers to micro-questions, like whether or not the Astors survived. It used common visual elements like color and area to indicate whether or not a person survived and number/proportion of people, as well as typographic aspects like italic and simple text replacement to indicate gender and the passengers' names.

The authors round out the article by addressing the most common criticisms of typography in data visualization, the foremost one being whether or not text should even be considered an element of data visualization, since visualization connotes preattentive visual encoding of information, and text or sequential information necessitates more investment of attention to understand. Another criticism is that textual representations are not as visually appealing even when used effectively. However, the authors counter that "this criticism indicates both the strength and weakness of type," that while text may not be suited for adding style or drama to a visualization, it can be particularly powerful in situations where a finer level of detail is needed, without sacrificing representation of higher level patterns. Lastly, a label length problem is common when using text in visualizations; differing lengths of names or labels may skew perception so that longer labels seem more important than shorter labels. This problem was encountered in the Titanic visualization with the varying lengths representations of passengers' names, and was corrected by only including a given name and a surname, the length of which could only vary so much.

All in all, this article has an interesting take on a somewhat less fashionable tool and puts forth the idea that text and typographic attributes can convey additional important information in data visualizations when used innovatively and correctly.

Reference: Banissi, Ebad, & Brath, Richard. (2016). Using Typography to Expand the Design Space of Data Visualization. She Ji: The Journal of Design, Economics and Innovation, 2(1), pp 59-87. https://www.sciencedirect.com/science/article/pii/S2405872616300107.

(Aischx, 2017)

## 3.10 This article explains 9 design principles which can be used for vizulation. These 9 design principles are:

https://www.idashboards.com/blog/2017/07/26/data-visualization-and-the-9-fundamental-design-principles/

1. Balance: A design is said to be balanced if key visual elements such as color, shape, texture, and negative space are uniformly distributed.
2. Emphasis: Draw viewers attention towards important data by using key visual elements.
3. Movement: Ideally movement should mimic the way people usually read, starting at the top of the page, moving across it, and then down. Movement can also be created by using complimentary colors to pull the user's attention across the page.
4. Pattern: Patterns are ideal for displaying similar sets of information, or for sets of data that equal in value. Disrupting the pattern can also be effective in drawing viewers attention; it naturally draws curiosity.
5. Repetition: Relationships between sets of data can be communicated by repeating chart types, shapes, or colors.
6. Proportion: If a person is portrayed next to a house, the house is going look bigger. In data visualization, proportion can indicate the importance of data sets, along with the actual relationship between numbers.
7. Rhythm: A design has proper rhythm when the design elements create movement that is pleasing to the eye. If the design is not able to do so, rearranging visual elements may help.
8. Variety: Variety in color, shape, and chart-type draws and keeps users engaged with data. Including more variety can increase information retention by the viewer. But when there is too much variety, important details can be overlooked.
9. Unity: Unity across design will happen naturally if all other design principles are implemented.

**Using Data Visualization to Find Insights in Data**

**Reference Link** http://datajournalismhandbook.org/1.0/en/understanding_data_7.html

This article is extracted from a book known as Data Journalism Handbook and this is one of the chapters of the book. The author starts the article by introducing a very simple idea that loading any dataset into a spreadsheet can also be a form of visualization as an invisible data becomes visible in a picture form into a table. Hence the focus should not be whether we need data visualization or not but should be on which form of data visualization is best in which situation.

The author then proceeds by stating that data visualization will not always unleash a readymade story on its own. Sometimes the insights are known before the visualization and sometimes an insight can be completely new. The author has given a process for finding insights in the following way:

Visualize Data-> Analyze -> Document Insights -> Transform Datasets ->Visualize Data

Each stage is explained in-depth further. Data Visualization can be done in many ways such as tables which are great for one dimensional data however they are bad for multi-dimensional data. Then he goes further to explain the situation where each type of visualization such as bar charts, maps, scatterplots, graphs, etc. are used. This gives a thorough understanding of when to use which type of visualization. Once we visualize the data we need to ask the following questions:

1 What can I see in this image? Is it what I expected? 2 Are there any interesting patterns? 3 What does this mean in the context of the data?

The basic question answer format gives an idea to the viewers about what kind of perspectives can we look at the data. Sometimes we discover something and sometimes we don't. But the author mentions that we

always learn something from the visualization. Once we document the data insights based on the above question we need to have the following points into consideration:

1 Why have I created this chart? 2 What have I done to the data to create it? 3 What does this chart tell me?

The above question answer format compels the viewers to think deeper about what exactly we are trying to find. Because many times the viewers are simply too overwhelmed with the size of data that they lose the basic idea. Hence this kind of approach help to stay focused. The author then mentions that based on the above insights we might have some idea about some interesting patterns. Since we already have an idea we might want to see it in more detail and hence we transform data in more details such as Zooming, Filtering, Outlier Removal. The author then explains how transformed data can help us to see a more detailed view of our insights.

Further the author gives a detailed explanation of which data visualization tool to use based on the situation. The entire process given above is explained in depth with the help of examples. The technical approach listed above is practical and can be implemented easily on our data visualization projects. I liked the author's approach because he has cleverly integrated the step-by-step process of finding insights with the technical way of handling datasets using tools such as Tableau, Python, etc. And the process can be repeated many times till we find the insights we are looking for.

Building advanced analytics application with TabPy

https://www.tableau.com/about/blog/2017/1/building-advanced-analytics-applications-tabpy-64916

?{great_viz, author = {{Bora Beran}}, year = {2017}, title = {Building advanced analytics applications with TabPy}, howpublished = {https://www.tableau.com/about/blog/2017/1/building-advanced-analytics-applications-tabp note = {Accessed: 2018-04-28} }

Imagine a scenario where we can just enter some x values in a dashboard form, and the visualization would predict the y variable!!! Here is a link that shows how to integrate and visualize data from Python in Tableau. This is especially relevant to all data science students, as this is one of the tools used for visualizing advanced analytics. The author here has given an example using data from Seattle's police department's 911 calls and he tries to identify criminal hotspots in the area. The author uses machine learning (spatial clustering) and creates a great interactive visualization, where you can click on the type of criminal activity and the graph will show various clusters. There are other examples and use cases that may be downloaded, and the scripts are also given by the author for anyone who is interested in trying it out.

`+ Theoretical background of data visualization`

- Contemporary research results

Some best practices for visualization:

http://www.dataplusscience.com/files/visual-analysis-guidebook.pdf

Here is free pdf to some best practices in visual analysis. It talks about the right charts to be used for various kinds of analysis. It is very relevant for data science students as we would be interested in presenting our analysis using simple and effective visualizations that tell the complete story.

Some of key areas for which the author highlights some best practices are for visualizing trends over time, comparison and ranking, correlation, distribution, geographical data etc.

The author gives examples on how simple graphs can also become more effective by just adding a few more elements or some simple adjustments.

I feel this is a great starting point to create effective charts and we may use these principles also when we start doing advanced analytics.

contributions

###Interactive Data Visualization

Interactive or Dynamic data visualization delivers today's complex sea of data in a graphically compelling and an easy-to-understand way. It enables direct actions on a plot to change elements and link between multiple plots. It enables users to accomplish traditional data exploration tasks by making charts interactive.

####Benefits of Interactive Data Visualization Software:

1. Absorb information in constructive ways: With the volume and velocity of data created everyday, dynamic data viz enables enhanced process optimization, insight discovery and decision making.
2. Visualize relationships and patterns: Helps in better understanding of correlations among operational data and business performance.
3. Identify and act on emerging trends faster: Helps decision makers to grasp shifts in behaviors and trends across multiple data sets much more quickly.
4. Manipulate and interact directly with data: Enables users to engage data more frequently.
5. Foster a new business language : Ability to tell a story through data that instantly relates the performance of a business and its assets.

(Internet, 2017)

There are multiple ways by which interactive data visualizations can be developed: 1. D3.js 2.Tableau 3.R shiny

####D3.js:

D3.js (or just D3 for Data-Driven Documents) is a JavaScript library for producing dynamic, interactive data visualizations in web browsers(From Wikipedia). It is highly functional, meaning you can reuse the code and add functions relevant to your project. Embedded within an HTML webpage, the JavaScript D3.js library uses pre-built JavaScript functions to select elements, create SVG objects, style them, or add transitions, dynamic effects or tooltips to them.

Some of the key advantages are: It is dynamic, free and open source and very flexible with all web technologies, the abiity to handle big data and the functional style allows to reuse the codes.

(Blog, 2017)

####Tableau:

Tableau is business intelligence (BI) and analytics platform created for the purposes of helping people see, understand, and make decisions with data. It is the industry leader in interactive data visualization tools, offering a broad range of maps, charts, graphs, and more graphical data presentations. It is a painless option when cost is not a concern and you do not need advanced and complex analysis.The application is very handy for quickly visualizing trends in data, connecting to a variety of data sources, and mapping cities/regions and their associated data.

The key advantages are: It provides non technical user the ability to build complex reports and dashboard with zero coding skills. Using drag-n-drop functionalities of Tableau, user can create a very interactive visuals within minutes. It can handle millions of rows of data with ease and users can make live to connections to different data sources like SQL etc.

(AbsentData, 2017)

####R Shiny :

R Shiny enables us to produce interactive data visualizations with a minimum knowledge of HTML, CSS, or Java using a simple web application framework that runs under the R statistical platform. Standalone apps can be hosted on a webpage or embedded in R Markdown documents and dashboards can be built using R shiny. It combines the computational power of R with the interactivity of the modern web.

The main advantages of using R Shiny are : Its flexibility of pulling in whatever package in R that you want to solve your problem, reaping the benefits of an open source ecosystem for R and Javascript visualization libraries, thereby allowing to create highly custom applications and enabling timely, high quality interac-

tive data experience without (or with much less) web development and without the limitations or cost of proprietary BI tools.

(Castañón, 2016)

# Chapter 4

# Avoiding Common Mistakes with Time Series

https://www.svds.com/avoiding-common-mistakes-with-time-series/

This article explains how time series data visualization can sometimes be deceptive. It first takes an example of two random time series data and plots them on a graph which gives an impression that the two are strongly correlated. But if we do some statistical testing the two do not show any relationship, this is an example of "correlation does not necessary mean causation". In another set of examples author has taken trending two random time series data and shown how even statistical tests can give a wrong interpretation. The article then explains using visualization how a general trended time series can be different than a more controlled and measured trending time series.

# Chapter 5

# Case Studies

- Description and replication of great examples of data visualization

http://flowingdata.com/2015/12/22/10-best-data-visualization-projects-of-2015/

The author picked top 10 projects for the best data visualization of 2015, for each pick, the author showed the project plot and also described the reason why he chose. So after reading this article, I have a basic understanding of what kind of characters should include in a good visualization project.

## 5.1 Description and replication of great examples of data visualization

reference:http://blog.visme.co/best-information-graphics-2016/#e030mFiF7wCpk7Ld.99

1. Connecting the Dots Behind the Election

https://www.nytimes.com/interactive/2015/05/17/us/elections/2016-presidential-campaigns-staff-connections-clinton-bush
html?_r=1

This article by the New York Times lists several different candidates and creates compelling visuals that link their campaigns to previous ones.

Each visual contains several different-sized dots that represent a specific campaign, administration, or other governmental organization related to the candidate's current campaign, which are then connected by arrows.

Hovering over a specific dot highlights the connections between the groups. The visual is a great way to put what would otherwise be a long slog through years of information into an easily accessible, easily viewable format so that voters can figure out where the candidates' experiences lie.

2. Spies in the Skies

https://www.buzzfeed.com/peteraldhous/spies-in-the-skies?utm_term=.so1GQ6ZGDo#.ec8kL3WkZe

The map is filled with red and blue lines (representing FBI and DHS aircraft, respectively) which illustrate the flight paths of the planes. When planes circle an area more than once, the circles become darker. The circles change in accordance to day and time, and individual cities can be typed into a search bar to see the flight patterns over them.

The visualization, rather creatively, almost looks like a hand-drawn map. While presenting a normally uncomfortable topic, this allows individuals to check things for themselves, hopefully providing some peace of mind.

3. Green Honey

http://muyueh.com/greenhoney/?es_p=1228877

The visualization spans a webpage. As you scroll down, the text changes, as do many colored dots that move over the white background. The dots are used to represent not only each colors' hue, but the numbers that fall into each category—for example, what colors are the most popular "base" colors for English and Chinese.

The continuous flow of this visualization helps really bring it together, allowing users to scroll through the information at their own pace, but also creating a seamless, creative work.

4. How People Like You Spend Their Time

http://flowingdata.com/2016/12/06/how-people-like-you-spend-their-time/

The visual lists several categories along one side of a graph—such as "personal care" and "work"—with a line illustrating the amount of time the average person in a certain demographic spends on each subject. Entering different statistics at the top—such as changing gender or age—causes the lines to shift to feature that demographic.

The simplicity of this visualization really helps the information get across and avoids bogging down the statistics. Sometimes, less is more.

5. Is it Better to Rent or Buy?

https://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html?_r=0

The calculator includes several sloping charts. Each chart includes a factor that'll affect how much you'll have to pay, such as the individual cost of your home and your mortgage rates. A movable scale along the bottom of each chart allows you to enter different data, changing the "cost of rent per month" on the side. If you can find a similar house to rent for that much per month or less, it's more cost effective to just rent the home.

This visualization is incredibly thorough and a useful tool for homeowners of any age and status.

6. What's really warming the world?

https://www.bloomberg.com/graphics/2015-whats-warming-the-world/

In this case study, it first claimed the background story and the analytical questions clearly. Then it analyzed each different factors separately using both verbal explanations and dynamic graphics to compare with the observed temperature movements, and then grouped related factors into Natural factors category or Human factors category. After that, it combined all the dynamic graphics into one and made the results more straightforward in terms of comparisons. In the end, the authors also provided more detailed methodology explanations with dataset sources to support the results shown above.

Overall, this case study is straightforward, easy to understand but also with enough information shown on each graphics.

7. The Strengths of Animated Data Visualization

http://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/?platform=hootsuite

The page linked above includes a great example of animated data visualization showing the time people spend on daily activities throughout the day. The plot is simple and easy to interpret, but it also includes a good number of variables including time, activity type, number of people doing each activity, and the order in which activities are done.

One of the plot's biggest strengths is that by using one dot to represent each person in the study and using animation, we can actually drill down to each individual and follow them throughout the day. The accumulation of dots for each particular activity also gives us an aggregate-level view of the same data, so we get both an individual and aggregate insights.

A drawback of the plot is that it is hard for our eyes to keep track of 1000 simultaneously moving dots. The author of the post addresses this by creating subsequent plots with stationary lines at key times of the day. This represents people's movements from one activity to another without overwhelming the reader.

Overall, this is an engaging, informative, and fun animated plot that has relevance and tells a story.

8. Case studies: **An Aging Population**

Aging population is always a hot topic in social economics and politics. I collect several different data visualizations that show aging population in the world. They are good examples to learn and apply to census data.

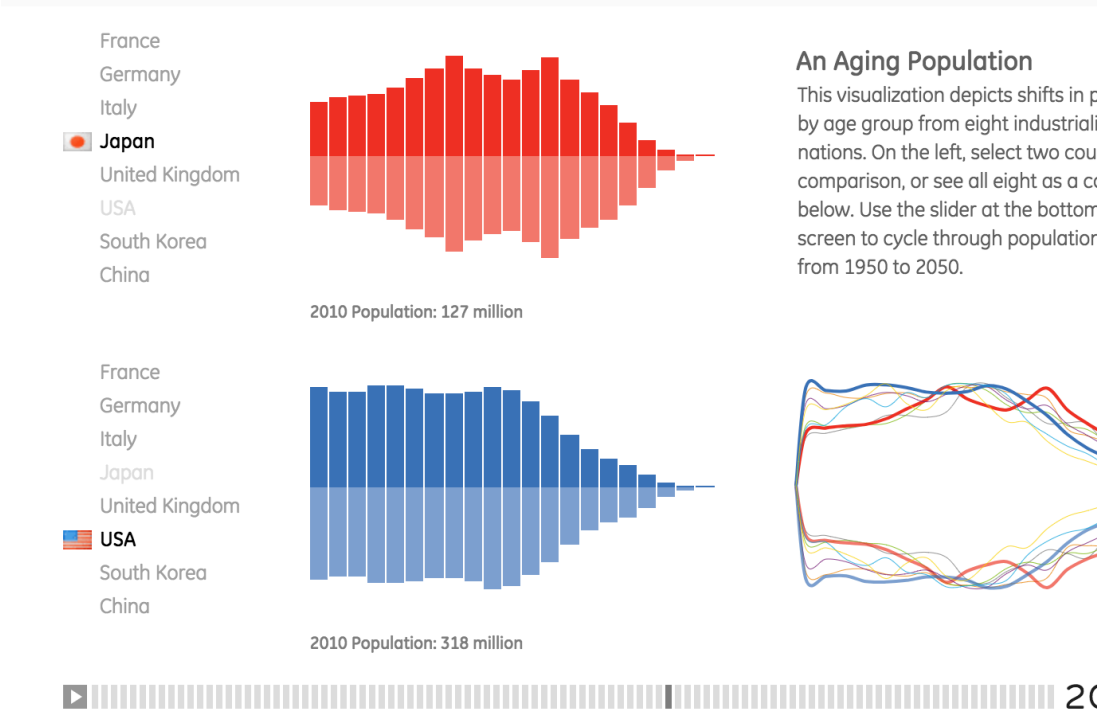
8.1 An Aging Nation: Projected Number of Children and Older Adults Source

This one includes bar chart and line graph to demonstrate the aging population compared with population of children. The good things about this visualization: simple to see and compare, color to differentiate the category, highlight the intersection point.

8.2 From Pyramid to Pillar: A Century of Change, Population of the U.S. Source

This is a **population pyramid**. "A **population pyramid** is a pair of back-to to histograms for each sex that displays the distribution of a population in all age groups and in gender".

It is a good candidate to compare changes in population distributions (sex, age, year). Also the shape of pyramid is used to interpret a population. To illustrate, A pyramid with a very wide base and a narrow top section suggests a population with both high fertility and death rates. It is a useful tool in the census data.




8.3 Animated pyramid Source

This is an animated and multiple population pyramids. It used to compare different patterns across countries. One additional benefit for the interactive population pyramid is that it shows the shape changes year by year, which is useful for countinous time-series comparison.

Similar projected with R code is provided for references: link

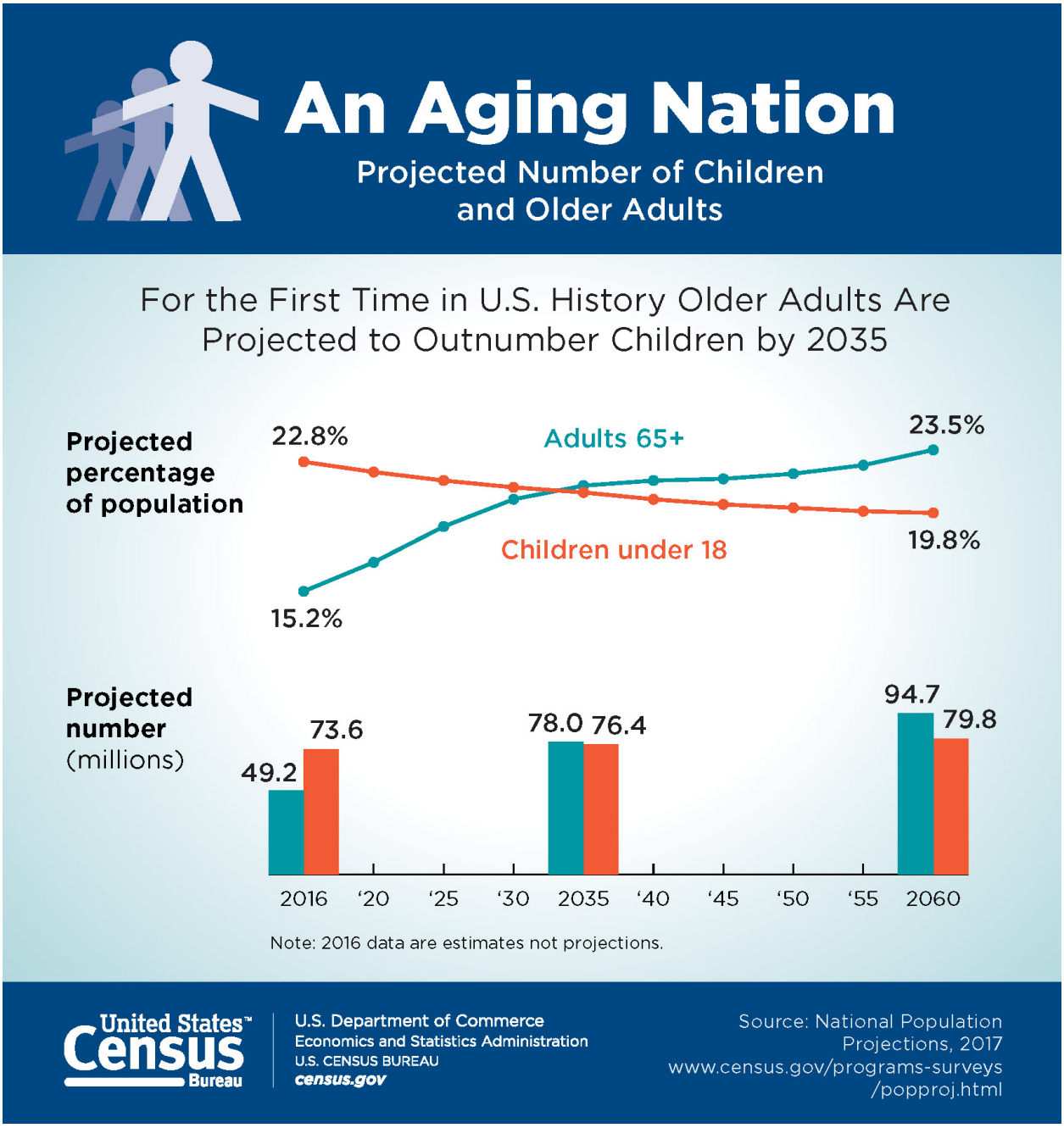

An Aging Nation
Projected Number of Children and Older Adults
For the First Time in U.S. History Older Adults Are Projected to Outnumber Children by 2035
Projected percentage of population
22.8%
Adults 65+
23.5%
Children under 18
19.8%
15.2%
Projected number (millions)
49.2
73.6
78.0
76.4
94.7
79.8
2016
'20
'25
'30
2035
'40
'45
'50
'55
2060
Note: 2016 data are estimates not projections.
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov
Source: National Population Projections, 2017
www.census.gov/programs-surveys/popproj.html

Figure 5.1:

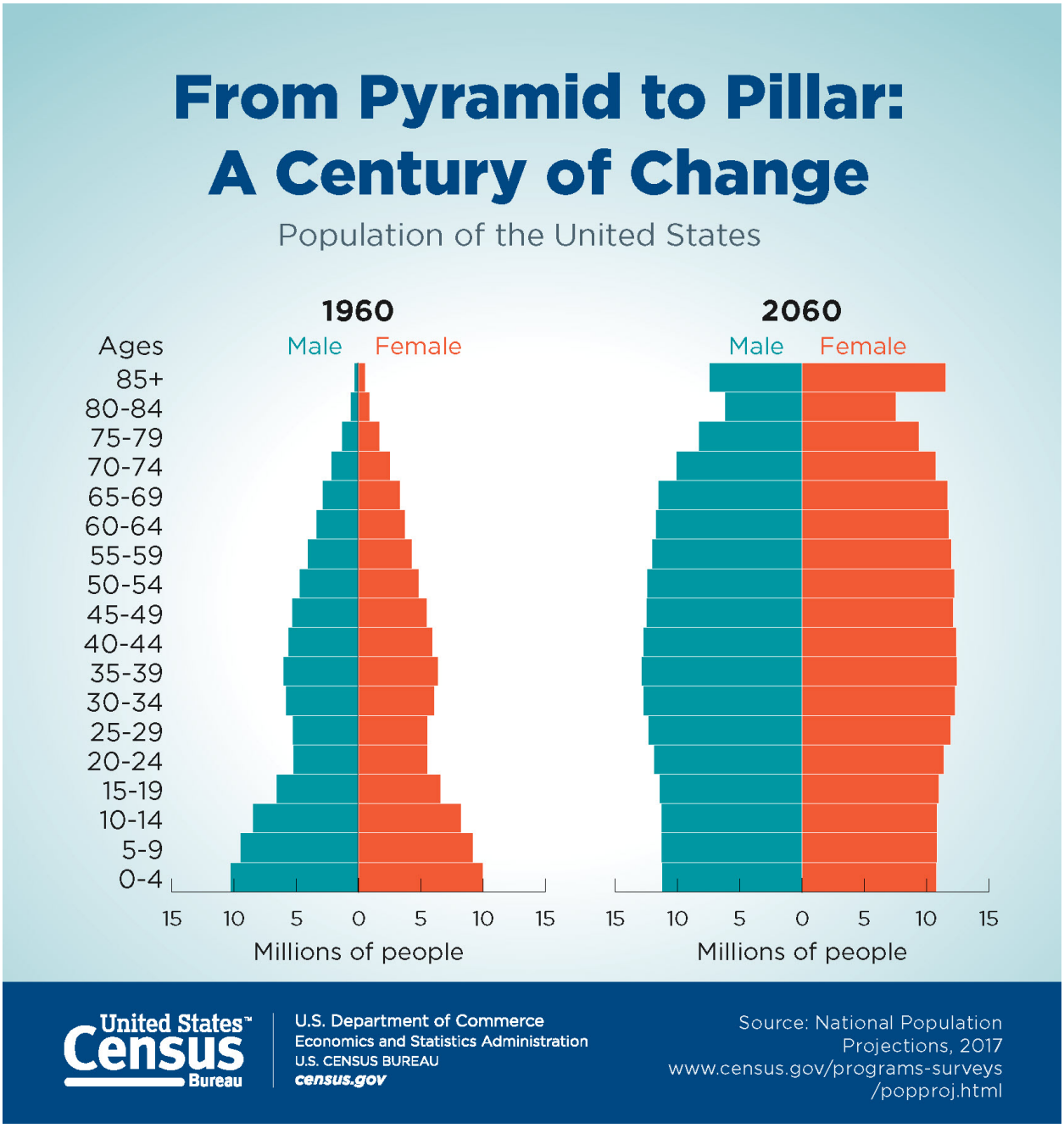

From Pyramid to Pillar:
A Century of Change
Population of the United States
1960
Male
Female
Ages
85+
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
15 10 5 0 5 10 15
Millions of people
2060
Male
Female
15 10 5 0 5 10 15
Millions of people
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov
Source: National Population Projections, 2017
www.census.gov/programs-surveys/popproj.html

Figure 5.2:

## 5.2 Deceptive data graphs examples

references: **Misleading Graphs: Real Life Examples http://www.statisticshowto.com/misleading-graphs/ **

Misleading graphs are sometimes deliberately misleading and sometimes it's just a case of people not understanding the data behind the graph they create. But some real life misleading graphs go above and beyond the classic types. Some are intended to mislead, others are intended to shock. The "classic" types of misleading graphs include cases where:

- **The Missing Baseline.**

For example, the Vertical scale is too big or too small, or skips numbers, or doesn't start at zero, like the graph below:

You might be thinking that the graph on the right shows The Times makes double the sales of The Daily Telegraph. But take a closer look at the scale and you'll see although The Times does make more sales, it's only beating the competition by about 10%.

- **The graph isn't labeled properly.**

Graghs can have the correct figures, but still can mislead you.

This one used a BIG HEADLINE makes you think that 5.3% of children get spinal cord injuries which is a pretty scary statistic for parents. But the real figure is about .0000003% (based on 2000 injuries per year out of a population of around 74,000,000).

And for the figure 1 used in this article: Misleading Graphs: Displaying a Change in One Variable Using Area or Volume https://www.forbes.com/sites/naomirobbins/2012/02/28/misleading-graphs-displaying-a-change-in-one-variable-#696674551781, the label for the smaller triangle in this graph says $26.4 while the label for the larger triangle says $114.6. $114.6 is 4.34 times $26.4. It certainly looks to me as if more than 4.34 smaller triangles will fit in the larger triangle. It is the altitudes of the triangles that are proportional to the numbers in the labels.

- **Data is left out.**

Only include part of the data like the following graph which using temperatures of the first half of the year to prove it was rising dramatically.

For more examples of misleading graphs or deceptive graphs you can read the following articles for more inspirations:

- bar charts without zero & evenly spaced tick marks for uneven intervals: https://www.forbes.com/sites/naomirobbins/2011/11/17/whats-wrong-with-this-graph/#502ab1a42a33
- graphs not drawn to scale: https://www.forbes.com/sites/naomirobbins/2012/02/16/misleading-graphs-figures-not-dra#351dcf9c15ef
- **Treating correlation as causation.**

Even if the labels and data in your graph is correct, it does not mean that the conclusion is logically correct. A correlation between X and Y does not automatically indicate that the change in one variable is caused by the change in the values of the other one, whereas the causation means that one event is the result of the occurrence of the other event. From the graph, we should bear in mind that it only presents the correlation between ice cream sold and murders, rather than causation.

Some Interesting Visualizations: https://blog.hubspot.com/marketing/great-data-visualization-examples http://blog.visme.co/data-visualizations-current-events/ Visualization is like art. It speaks where words fail. There are phenomenas like the Syrian war, the number flights during Thanksgiving in the USA, the understanding of depths for developing perspective about the range of the issue, the controversy of '#OscarsSoWhite', etc. on which we can write bundles of paragraphs, but they might still have scope for ambiguity. The links show some intricate visualizations of the topics like those mentioned above,

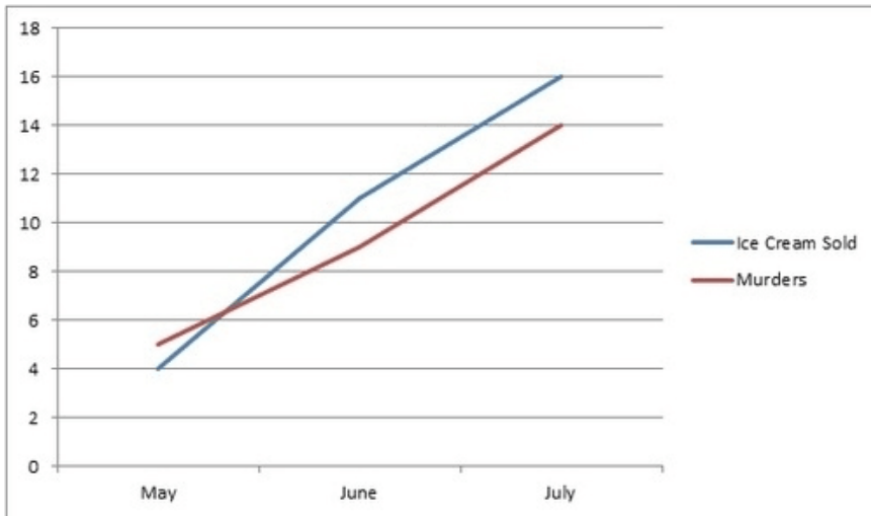



Figure 5.3: A strange correlation between ice cream sales and murders (Source: [@harlin-coorelation])

and speak volumes without requiring paragraphs to explain what is going on within these visualizations. According to me, it is really interesting to see that almost anything in this world can be explained by visualizations. Visualizations are not just limited to businesses and their analytics. Wars, rescue operations, etc. can also be visualized to get a clear idea of all the details of the issues. 1. Picking up from one of the charts shown in the above mentioned links, the visualization of 'A guide to Who is Fighting Whom in Syria' is one of the most interesting charts in the list. The visualization and its report can be seen at http://www.slate.com/blogs/the_slatest/2015/10/06/syrian_conflict_relationships_explained.html

This visualization makes an extremely complicated topic like the Syrian War easily understandable. It consists of 3 different emojis in three different colours, with each (colour+facial expression) combination showing the relationship between the various groups involved in the Syrian War. When you click on each of the emoji, a small dialogue box pops up which explains the relationship between the various countries and rebel groups involved in the war. This is not only easy to understand, but it is also pleasing to the eyes.

2. The second visualization 'Adding up the White Oscars Winners' can be seen here (https://www.bloomberg.com/graphics/2016-oscar-winners/) in an article by Bloomberg. The writes of this article developed the attributes of the future winners of Oscars by taking up the attributes of the past winners. It is extremely interesting to see how the article shows the features of the Best Actress, Actor, movies, etc. in a simple and captivating visual. The visualization is interactive and we can click on each attribute like 'Hair Color', 'Eye Color', etc. to see what are the features of the actors and actresses who are more likely to win the Oscars.

Similarly, the visualization gives information about the different aspects of movies that are more likely to win, like 'Length' ,'Month' ,'Budget', etc.

## 5.3 Application of Data Visualization

**Data Preprocessing**

We use data visualization for outliar detection in the dataset. Different methods for outlier detection in functional data have been developed during the years. Among them, several rely on different notions of functional depth , on robust principal components, or on random projections of infinite-dimensional data into R. Also, some distributional approaches have been considered (Gervini, 2009). In functional data analysis, we observe curves defined over a given real interval and shape outliers may be defined as those

curves that exhibit a different shape from the rest of the sample. Whereas magnitude outliers, that is, curves that lie outside the range of the majority of the data, are in general easy to identify, shape outliers are often masked among the rest of the curves and thus difficult to detect.

(Unwin, 2008)

8. Young voters, class and turnout: how Britain voted in 2017

Reference: https://www.theguardian.com/politics/datablog/ng-interactive/2017/jun/20/young-voters-class-and-turnout-ho

The article's goal is to convey the change in party votes in the 2017 UK general election compared to votes in 2015. The change in party votes was shown with regards to three demographic factors: age, class, and ethnicity. For each factor, there are four graphs (one per political party), each illustrated in their party's standard color. The change in percent of votes is shown as an arrow where the arrow's shaft is the length of the difference from 2015 to 2017 while the x-axis is the demographic factor split into different bins. What makes this a good visualization is that it is very easy to read and interpret. The color-coding of the arrows and party name makes it easy to pick out the different parties and the arrow lengths highlight just how large of a change happened. For example, in the Age section, it is easy to see the pattern between the Labour party gaining many voters ages 18 to 44 and the Conservative party gaining voters ages 45 and up.

8. Vizwiz blog: casestudies about how to improve your visualizations

http://www.vizwiz.com/

This is a blog about Tableau based data visualization. The author is Andy Kriebel who is a famous Tableau Zen Master. I would like to recommend this blog because it is not only practical, but also full of insights.

My favorite part of this blog is so called "Makeover Monday", which will develop a new visualization based on an original one. For example, the author re-designed "The Seasonality of Confirmed Malaria Cases in Zambia Southern Province" by pointing out "what works well", "what could be improved" and also his goals for the new visualization (ref: http://www.vizwiz.com/2018/04/malaria.html) That's how you can learn all the insight and reason behind a good visualization.

Besides, this blog also includes great tips and showcases for Tableau.

7. Uber: Crafting Data-Driven Maps

Map visualization is very important for companies like Uber that needs to track metrics using geo space points. In this article, the designer from Uber talks about the challenges of design such visualization and their solutions. While a lot of the problems are related to the large scale of the data, there are some insights on using scatter plots and hex bins, adding trip lines and making custom tools to help make decisions. The visualization in this article is beneficial for developing geo spatial graphics.

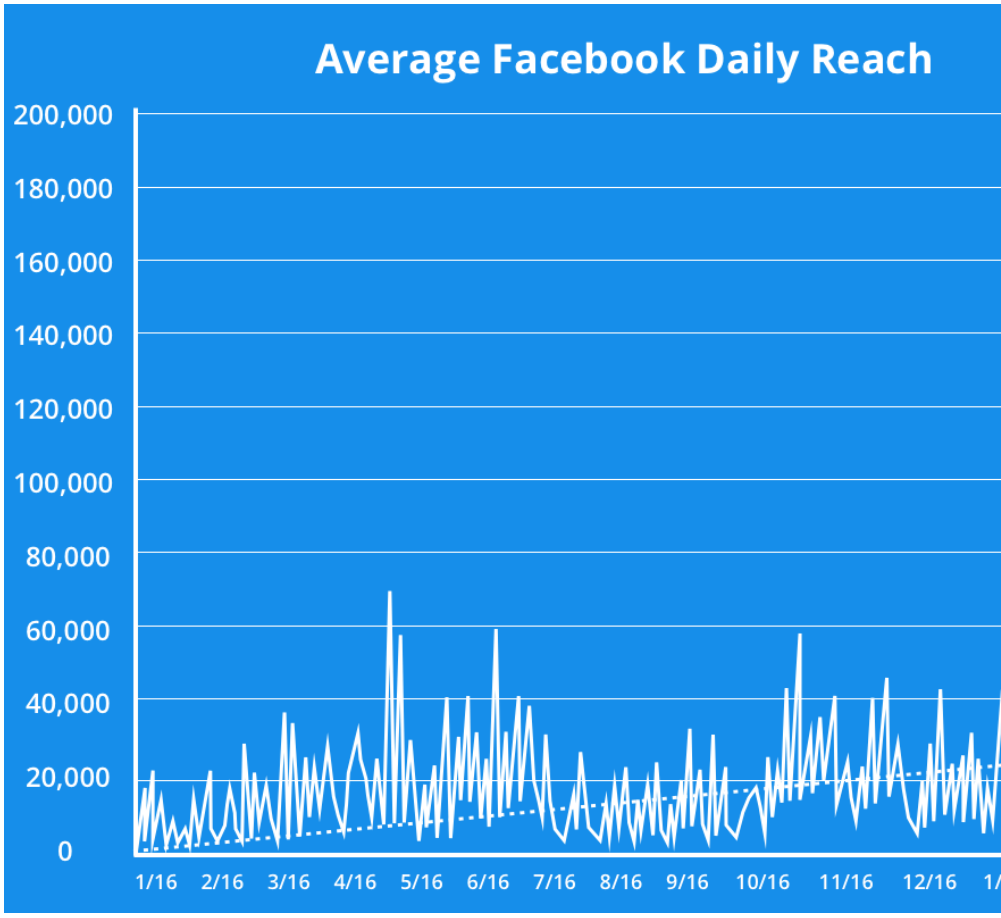
Reference: https://medium.com/uber-design/crafting-data-driven-maps-b0835b620554

Kissmetrics blog: visualization of metrics

Kissmetrics blog is a place where people talk about analytics, marketing and testing through narratives and metrics visualization. Metrics are important in real-life world especially when developing/promoting products. Visualization of metrics are also essential so that stakeholders can monitor performance, identify problems and deep dive into potential issues.

A good example from the Kiss metrics blog is about Facebook's Organic Reach. One important point in the blog discussed whether the Facebook's organic reach is decreasing drastically. The general trend shows that there is a huge decline in Facebook's page organic reach.

The following graphs show that the engagement is actually increasing, meaning while the quantity of content




is decreasing, the quality is increasing.

This resonates with what we have learnt at class in terms of how different perspectives of interpreting data can lead to different conclusions.

Reference:https://blog.kissmetrics.com/is-facebook-organic-reach-dead/

### 5.3.1 15 Data Visualizations That Will Blow Your Mind

**Reference**

Allison Stadd,January 21, 2015.15 Data Visualizations That Will Blow Your Mind,Udacity. https://blog.udacity.com/2015/01/15-data-visualizations-will-blow-mind.html

If a picture is worth a thousand words, a data visualization is worth at least a million.

As inspiration for your own work with data, check out these 15 data visualizations that will wow you. Taken together, this roundup is an at-a-glance representation of the range of uses data analysis has, from pop culture to public good.

1. Every Satellite Orbiting Earth

- https://qz.com/296941/interactive-graphic-every-active-satellite-orbiting-earth/

This interactive graph, built using a database from the Union of Concerned Scientists, displays the trajectories of the 1,300 active satellites orbiting the Earth as you read this. Each satellite is represented by a circular icon, color-coded by country and sized according to launch mass.

2. Simpson's Paradox

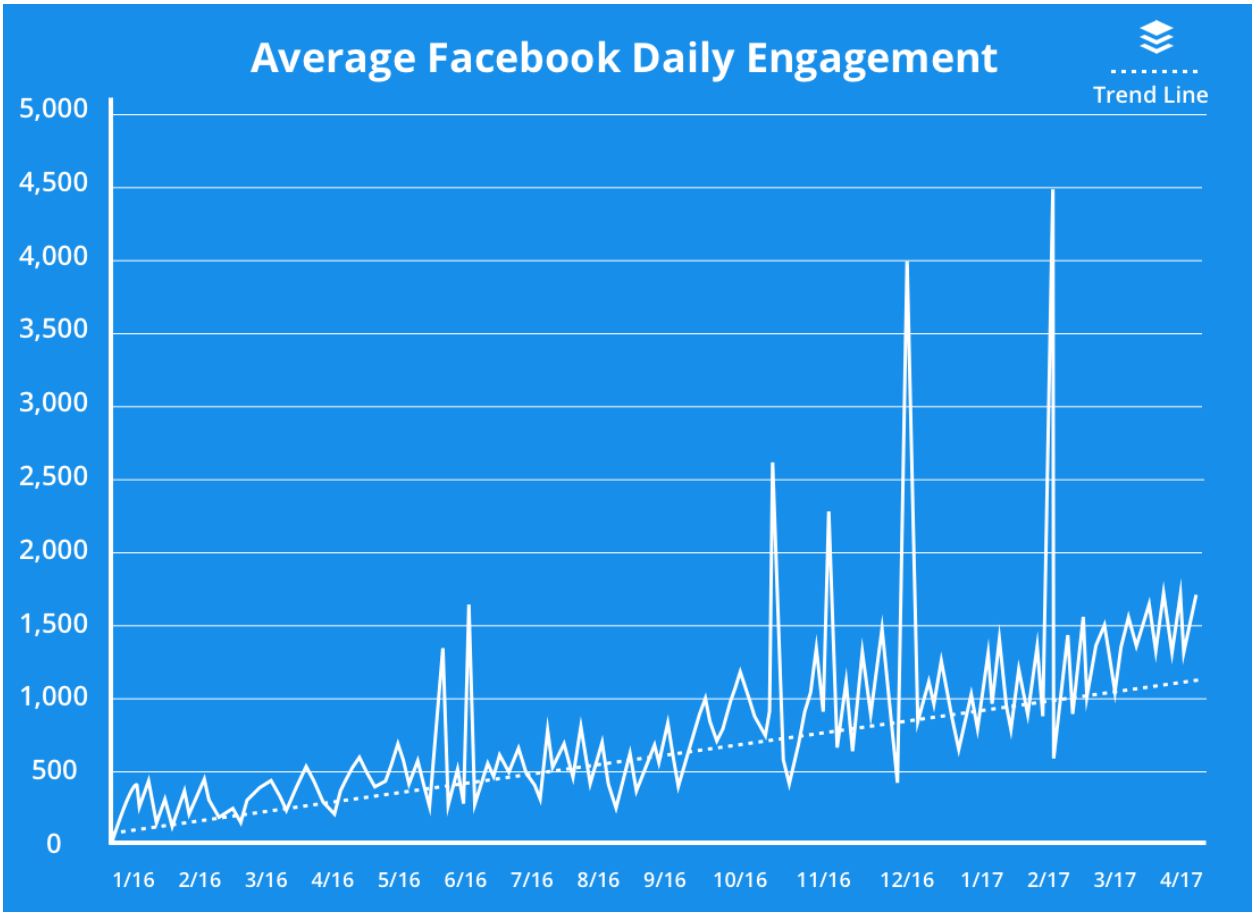

Average Facebook Daily Engagement
Trend Line
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
1/16
2/16
3/16
4/16
5/16
6/16
7/16
8/16
9/16
10/16
11/16
12/16
1/17
2/17
3/17
4/17

Figure 5.4:

- http://vudlab.com/simpsons/

The Visualizing Urban Data Idealab (VUDlab) out of the University of California-Berkeley put together this visual look at data that disproves the claim in a 1973 suit that charged the school with sex discrimination. Though the graduate schools had accepted 44% of male applicants but only 35% of female applicants, researchers later uncovered that if the data were properly pooled, there was actually a small but statistically significant bias in favor of women. That's called a Simpson's Paradox.

3. Charles Minard's Visualization of Napoleon's 1812 March

- https://www.edwardtufte.com/tufte/minard

This classic lithograph dates back to 1869, displaying the number of men in Napoleon's 1812 Russian army, their movements, and the temperatures they encountered along their way. It's been called one of the "best statistical drawings ever created." The work is an important reminder that the fundamentals of data visualization lie in a nuanced understanding of the many dimensions of data. Tools like D3.js and HTML are no good without a firm grasp of your dataset and sharp communication skills.

4. Hans Rosling's 200 Countries, 200 Years, 4 Minutes

- https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo

Global health data expert Hans Rosling's famous statistical documentary The Joy of Stats aired on BBC in 2010, but it's still turning heads. One segment in particular is pretty mind-blowing. In "200 Countries, 200 Years, 4 Minutes," Rosling uses augmented reality to explore public health data in 200 countries over 200 years using 120,000 numbers, in just four minutes.

5. Renting vs. Buying

- https://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html

Mike Bostock, New York Times graphics department editor and inventor of D3.js, built a complex interactive data calculator that offers a cost/benefit analysis for prospective homebuyers. Along with his colleagues Shan Charter and Archie Tse, Bostock tapped into everything from home price and mortgage-interest tax deduction to property tax rate and inflation to help you determine whether to rent or buy a home.

6. Music Timeline

- https://research.google.com/bigpicture/music/

Google's Music Timeline illustrates a variety of music genres waxing and waning in popularity from 2010 to present day, based on how many Google Play Music users have an artist or album in their library, and other data such as album release dates.

7. State of the Union 2014 Minute by Minute on Twitter

- http://twitter.github.io/interactive/sotu2014/#p1

Twitter's data team assembled an impressive interactive data hub that depicts how Twitter users across the globe reacted to each paragraph of President Obama's 2014 State of the Union address. You can slice and dice the data by topic hashtag (for example, #budget, #defense, or #education) and state. Pretty powerful.

8. NYC Street Trees

- https://www.cloudred.com/labprojects/nyctrees/#about

Using data from NYC Open Data, this interactive visualization shows the variety and quantity of street trees planted across the five New York City boroughs.

9. Millennial Generation Diversity

- http://money.cnn.com/interactive/economy/diversity-millennials-boomers/

CNNMoney's interactive chart showing the size and diversity of the millennial generation compared to baby boomers was built using U.S. Census Data. It turns dry numbers into an intriguing story, illustrating the racial makeup of different age groups from 1913 to present.

10. Goldilocks Exoplanets

- https://news.nationalgeographic.com/news/2014/04/140417-exoplanet-interactive/

Using data from the Planetary Habitability Laboratory at the University of Puerto Rico, the interactive graph plots planetary mass, atmospheric pressure, and temperature to determine what exoplanets might be home, or have been home at one point, to living beings.

11. Washington Wizards' Shooting Stars

- http://www.washingtonpost.com/wp-srv/special/sports/wizards-shooting-stars/

This detailed data visualization demonstrates D.C.'s basketball team's shooting success during the 2013 season. Using stats released by the NBA, the visualization lets you examine data for each of 15 players. See how successful each person was at a variety of types of shots from a range of spots on the court, compared with others in the league.

12. U.S. Migration Patterns

- https://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html?abt=0002&abg=0

The New York Times data team mapped out Americans' moving patterns from 1900 to present, and the results are fascinating to play around with. You can see where people living in each state were born, and to what states people move from others.

13. Selfie City

- http://selfiecity.net/

Selfie City, a detailed multi-component visual exploration of 3,200 selfies from five major cities around the world, offers a close look at the demographics and trends of selfies. The team behind the project collected and filtered the data using Instagram and Mechanical Turk. Explore the differences between selfies snapped in, say, New York and Berlin, as well as those between men and women across the world.

14. The American Workday

- https://www.npr.org/sections/money/2014/08/27/343415569/whos-in-the-office-the-american-workday-in-one-graph?/templates/story/story_php=

NPR tapped into American Time Use Survey data to ascertain the share of workers in a wide range of industries who are at work at any given time. The chart overlays the traditional 9 AM-5 PM standard over the graph for a reference point, helping you draw interesting conclusions.

15. Global Carbon Emissions

- https://www.theguardian.com/environment/ng-interactive/2014/dec/01/carbon-emissions-past-present-and-future-in

This data visualization, based on data from the World Resource Institute's Climate Analysis Indicators Tool and the Intergovernmental Panel on Climate Change, shows how national CO  emissions have transformed over the last 150 years and what the future might hold. Explore emissions by country for a range of different scenarios.

# Chapter 6

# Patterns

- Reusable solutions to everyday data visualization questions
- Applied by multiple members of the course

## 6.1 Why pie chart is bad: a comparison with bar chart

Using pie chart is usually considered as a bad idea when it comes to data visualization. But why? Here, we explore some cons of using pie chart to convey information and compare its effectiveness to bar chart (Hickey, 2013) (Henry, 2017) (Quach, 2016).

1. Some information may look nearly identical in pie chart. But if the data is presented with bar charts, the story is different. See figure 6.1 and 6.2 for examples.

2. It is difficult to compare the slices of a circle to figure out the distinctions in size between each pie slice, especially when there are a lot of categories. See figure 6.3 for example.

3. Pie chart is easy to be manipulated (e.g. using a 3D pie chart). See figure 6.4 for example.

4. Pie chart may be useful when comparing 2 different categories with different amounts of information. Specifically, it does a better job to distinguish two parts with a 25:75 split or one that is not 50:50 as people are sensitive to a right angle or a dividing line that is not straight. But this could be simply done by showing two numbers! See figure **??** and **??** for examples.

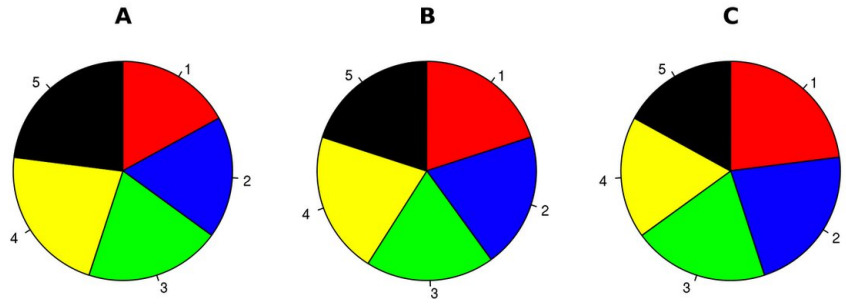



Figure 6.1: Are there any differences among the pollings at points A, B and C? (Source: [@hickey-pie-worst])

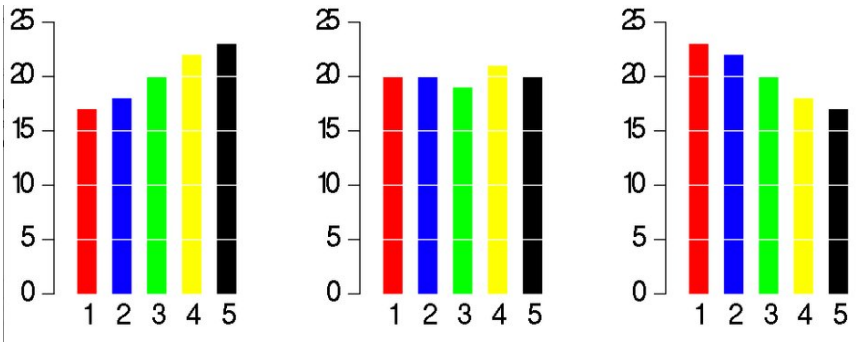



Figure 6.2: The differences can be clearly told from the bar charts. (Source: [@hickey-pie-worst])

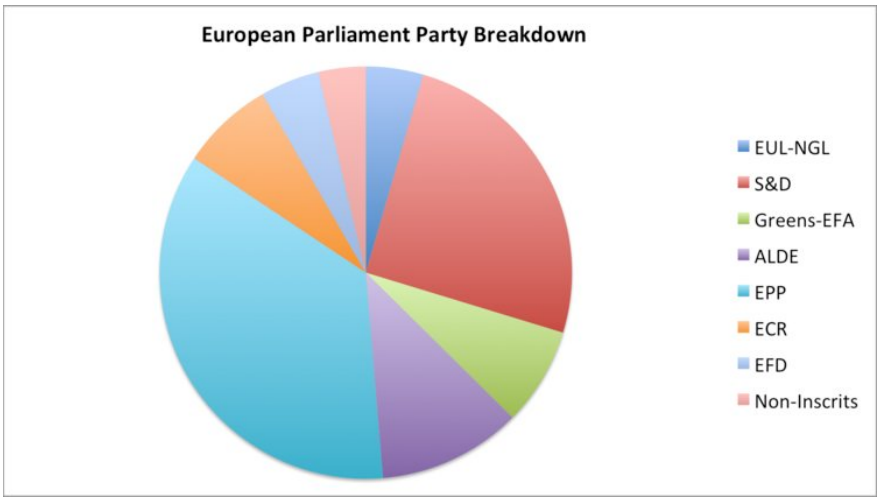



Figure 6.3: It is hard to compare the size of the slides. (Source: [@hickey-pie-worst])

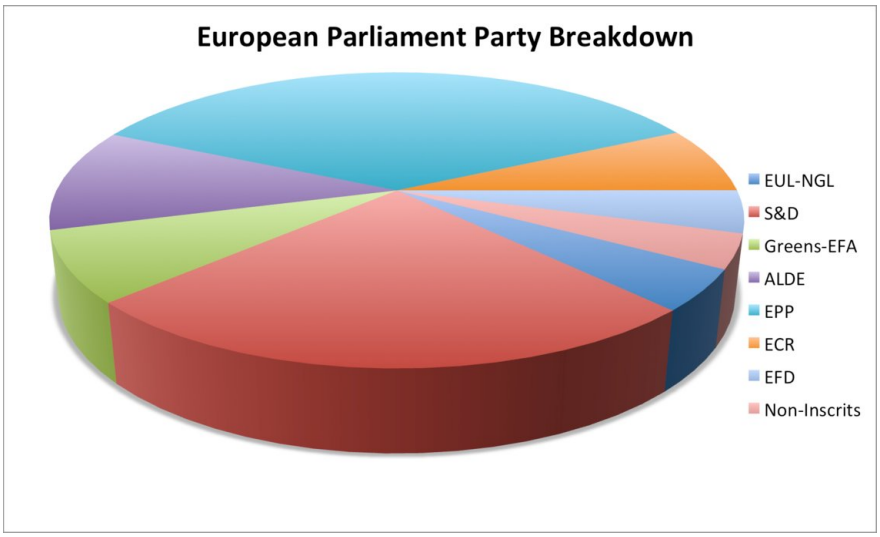



Figure 6.4: S and D (red) appears to be roughly even with EPP (teal) in a 3D pie chart. (Source: [@hickey-pie-worst])

## 6.2 Chose the right baseline in data visualization

Baseline is very important to data visualization. If baseline is different, the meanning will change a lot. Now here is a case study to show the importance of baseline and how to use it in different ways.

Here I use the same method for a new dataset to .

```
# Create the data.
a <-rep(c(2010,2011,2012,2013,2014,2015),each = 4)
b <- seq(1:24)
c <- c(64.9,65.33,71.67,79.17,68.78,69.83,78.61,92.68,89.28,90.43,97.96,106.96,100.66,107.53,117.06,119
data <- as.data.frame(cbind(a,b,c))
colnames(data) <-c("year","quater","sales")
```

1. Regular quaterly sales. We can see sales decreased a lot around 2014. **The baseline here is historical sales.**

```
# Regular time series for sales
par(cex.axis=0.7)
data.ts <- ts(data$sales, start=c(2010, 1), frequency=4)
plot(data.ts, xlab="", ylab="", main="sales per quater", las=1, bty="n")
```

2. Quaterly and yearly change sales. **The baseline here is zero and look at the percentage changes.**

```
 # Quaterly change
curr <- as.numeric(data$sales[-1])
prev <- as.numeric(data$sales[1:(length(data$sales)-1)])
quaChange <- 100 * round( (curr-prev) / prev, 2 )
barCols <- sapply(quaChange,
    function(x) {
        if (x < 0) {
            return("#2cbd25")
        } else {
            return("gray")
        }
    })
#monChange.ts <- ts(monChange, start=c(1976, 2), frequency=12)
barplot(quaChange, border=NA, space=0, las=1, col=barCols, main="% change, quaterly")
```

```
# Year-over-year change
curr <- as.numeric(data$sales[-(1:4)])
prev <- as.numeric(data$sales[1:(length(data$sales)-4)])
annChange <- 100 * round( (curr-prev) / prev, 2 )
barCols <- sapply(annChange,
    function(x) {
        if (x < 0) {
            return("#2cbd25")
        } else {
            return("gray")
        }
    })
barplot(annChange, border=NA, space=0, las=1, col=barCols, main="% change, annual")
```

From this plot, it is very clear that the magnitude of drops in sales for some quaters.

3. The sales difference compare to now. **The baseline here is the current sales.**

```
# Relative to current 2015
curr <- as.numeric(data$sales[length(data$sales)])
salesDiff <- as.numeric(data$sales) - curr
barCols.diff <- sapply(salesDiff,
    function(x) {
        if (x < 0) {
            return("gray")
        } else {
            return("black")
        }
    }
)
barplot(salesDiff, border=NA, space=0, las=1, col=barCols.diff, main="Sales difference from last quater
```

4. Sales difference compared to the first quater. ** The baseline here is the first quater sales.**

```
# Relative to first quater
ori <- as.numeric(data$sales[1])
salesDiff <- as.numeric(data$sales) - ori
barCols.diff <- sapply(salesDiff,
    function(x) {
        if (x < 0) {
            return("gray")
        } else {
            return("black")
        }
    }
)
barplot(salesDiff, border=NA, space=0, las=1, col=barCols.diff, main="Sales difference from first quate
```

5. The difference between quater sales and mean. ** The baseline is mean now.**

```
# difference from the mean
mean <- mean(as.numeric(data$sales))
salesDiff <- as.numeric(data$sales) - mean
barCols.diff <- sapply(salesDiff,
    function(x) {
        if (x < 0) {
            return("gray")
        } else {
            return("black")
        }
    }
)
barplot(salesDiff, border=NA, space=0, las=1, col=barCols.diff, main="Sales difference from mean")
```

So before we start to plot, we should decide the baseline we want to use. Different baseline will lead to totally different graphs.

Reference: https://flowingdata.com/2013/11/26/the-baseline/

## 6.3 Tips to improve Data Visualization

### 6.3.1 1.Comparison

Include a zero baseline if possibleAlthough a line chart does not have to start at a zero baseline, it should be included if it gives more context for comparison. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances; Always choose the most efficient visualization; Watch your placement You may have two nice stacked bar charts that are meant to let your reader compare points, but if they're placed too far apart to "get" the comparison, you've already lost; Tell the whole story Maybe you had a 30% sales increase in Q4. Exciting! But what's more exciting? Showing that you've actually had a 100% sales increase since Q1. ### 2.Copy Don't over explain If the copy already mentions a fact, the subhead, callout, and chart header don't have to reiterate it; Keep chart and graph headers simple and to the point There's no need to get clever, verbose, or pun-tastic. Keep any descriptive text above the chart brief and directly related to the chart underneath. Remember: Focus on the quickest path to comprehension; Use callouts wisely Callouts are not there to fill space. They should be used intentionally to highlight relevant information or provide additional context; Don't use distracting fonts or elements Sometimes you do need to emphasize a point. If so, only use bold or italic text to emphasize a point—and don't use them both at the same time. ### 3.Color Use a single color to represent the same type of data; Watch out for positive and negative numbers Don't use red for positive numbers or green for negative numbers. Those color associations are so strong it will automatically flip the meaning in the viewer's mind; Make sure there is sufficient contrast between colors; Avoid patterns Stripes and polka dots sound fun, but they can be incredibly distracting. If you are trying to differentiate, say, on a map, use different saturations of the same color. On that note, only use solid-colored lines (not dashes); Select colors appropriately; Don't use more than 6 colors in a single layout. ### 4.Ordering Order data intuitively There should be a logical hierarchy. Order categories alphabetically, sequentially, or by value; Order consistently; Order evenly Use natural increments on your axes (0, 5, 10, 15, 20) instead of awkward or uneven increments (0, 3, 5, 16, 50). ### 5.Audience perspective Let the users lead;Know your audience,Designers should consider the way users prefer to understand information, even in choosing basic analytic approaches. For users to feel comfortable adopting and sharing insights from analytics, they must be able to explain and defend the data. ### 6.Use layers to tell a story While style is one form of customization, layering unique data sets on a single visualization can tell a richer narrative and connect users to the data without getting too crowded. On a map, this can be as simple as zooming in and out, but it can also involve drill-downs (choosing a data point and expanding it to show more detail), links and other shortcuts. ### 7.Keep it simple Analytic results shouldn't be presented to 10 decimal places when the user doesn't need that level of precision to make a decision or understand a concept. Effective visual interfaces avoid 3-D effects or ornate gauge designs (a.k.a. "chart junk") when simple numbers, maps or graphs will do.

References: https://www.columnfivemedia.com/25-tips-to-upgrade-your-data-visualization-design

## 6.4 Tips for Tableau

Running totals

Common Baseline

Weighted averages

Moving average

Grouping by aggregates

Different years comparison

Appending excel sheets

Bar chart totals

Fixed axis when re-drawing charts

Auto-fitting screen behavior depending on data selection

References: http://cdn2.hubspot.net/hubfs/257922/Docs/BlueGranite_whitepaper_10useful.pdf

# Chapter 7

# Tips to improve Data Visualization

1.Comparison

Include a zero baseline if possibleAlthough a line chart does not have to start at a zero baseline, it should be included if it gives more context for comparison. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances; Always choose the most efficient visualization; Watch your placement You may have two nice stacked bar charts that are meant to let your reader compare points, but if they're placed too far apart to "get" the comparison, you've already lost; Tell the whole story Maybe you had a 30% sales increase in Q4. Exciting! But what's more exciting? Showing that you've actually had a 100% sales increase since Q1.

2.Copy

Don't over explain If the copy already mentions a fact, the subhead, callout, and chart header don't have to reiterate it; Keep chart and graph headers simple and to the point There's no need to get clever, verbose, or pun-tastic. Keep any descriptive text above the chart brief and directly related to the chart underneath. Remember: Focus on the quickest path to comprehension; Use callouts wisely Callouts are not there to fill space. They should be used intentionally to highlight relevant information or provide additional context; Don't use distracting fonts or elements Sometimes you do need to emphasize a point. If so, only use bold or italic text to emphasize a point—and don't use them both at the same time.

3.Color

Use a single color to represent the same type of data; Watch out for positive and negative numbers Don't use red for positive numbers or green for negative numbers. Those color associations are so strong it will automatically flip the meaning in the viewer's mind; Make sure there is sufficient contrast between colors; Avoid patterns Stripes and polka dots sound fun, but they can be incredibly distracting. If you are trying to differentiate, say, on a map, use different saturations of the same color. On that note, only use solid-colored lines (not dashes); Select colors appropriately; Don't use more than 6 colors in a single layout.

4.Ordering

Order data intuitively There should be a logical hierarchy. Order categories alphabetically, sequentially, or by value; Order consistently; Order evenly Use natural increments on your axes (0, 5, 10, 15, 20) instead of awkward or uneven increments (0, 3, 5, 16, 50).

5.Audience perspective

Let the users lead;Know your audience,Designers should consider the way users prefer to understand information, even in choosing basic analytic approaches. For users to feel comfortable adopting and sharing insights from analytics, they must be able to explain and defend the data.

6.Use layers to tell a story

While style is one form of customization, layering unique data sets on a single visualization can tell a richer narrative and connect users to the data without getting too crowded. On a map, this can be as simple as zooming in and out, but it can also involve drill-downs (choosing a data point and expanding it to show more detail), links and other shortcuts.

7.Keep it simple

Analytic results shouldn't be presented to 10 decimal places when the user doesn't need that level of precision to make a decision or understand a concept. Effective visual interfaces avoid 3-D effects or ornate gauge designs (a.k.a. "chart junk") when simple numbers, maps or graphs will do.

References: https://www.columnfivemedia.com/25-tips-to-upgrade-your-data-visualization-design http://www.govtech.com/pcio/10-Tips-for-Data-Visualization.html

# Chapter 8

# Tips for Tableau

1. Running totals
2. Common Baseline
3. Weighted averages
4. Moving average
5. Grouping by aggregates
6. Different years comparison
7. Appending excel sheets
8. Bar chart totals
9. Fixed axis when re-drawing charts
10. Auto-fitting screen behavior depending on data selection References: http://cdn2.hubspot.net/hubfs/257922/Docs/BlueGranite_whitepaper_10useful.pdf

# Chapter 9

# Ethics

- Implications of (good and bad) data visualization
  - The role of data visualization in politics, society, and business

Tableau: Viz of the Day

Tableau has a gallery called Viz of the Day (https://public.tableau.com/en-us/s/gallery ) that displays great data visualization examples created by Tableau. It is cool to see how people are using all kinds of data to create informative yet fun data visuals. Data being used is also attached so we can try to mimic what other people did as well.

One interesting example I found is Describe Artists with Emoji (https://public.tableau.com/en-us/s/gallery/what-emoji-say-about-music?gallery=featured). Using the data from Spotify, the author listed the 10 most distinctive emoji used in the playlists related to popular artists. The table being used in this visual is very straight forward to link artist to the emojis and is very easy to compare among artists. When you hover over the emoji, further information is presented.

**1. Data visualization in political and social sciences** - (Reference: https://github.com/mschermann/data_viz_reader/files/1933699/Zinovyev_Data_Visualization.pdf)

The basic objective of data visualization is to provide an efficient graphical display for summarizing and reasoning about quantitative information. And during the last decades, political science has accumulated a large corpus of various kinds of data, which makes it gradually become a more quantitative scientific field and requires using quantitative information in the analysis and reasoning.

Data visualization plays several important roles in it: 1) helps create informative illustrations of the data, recapitulating large amount of quantitative information on a diagram; 2) helps formulate new or supporting existing hypotheses from quantitative data; 3) guides a statistical analysis of data and checks its validity.

Some useful visualization methods are: 1) *Statistical graphics and infographics*; 2) *Geographical information systems (GIS)*; 3) *Graph visualization or network maps*; 4) *Data cartography.*

**2. Role of data visualization in business** -(Reference:https://www.iotforall.com/data-visualization-strategy-for-business

According to an Experian report, 95% of U.S. organizations say that they use data to power business opportunities, and another 84 percent believe data is an integral part of forming a business strategy. Visualization helps data impact business in following ways:

1) *Cleansing*

The simplest way to explain the importance of visualization is to look at visualization as the means to making sense of data. Even the most basic, widely-used data visualization tools that combine simple pie charts and bar graphs help people comprehend large amounts of information fast and easily, compared to paper reports and spreadsheets.

In other words, visualization is the initial filter for the quality of data streams. Combining data from various sources, visualization tools perform preliminary standardization, shape data in a unified way and create easy-to-verify visual objects. As a result, these tools become indispensable for data cleansing and vetting and help companies prepare quality assets to derive valuable insights.

2)*Extracting*

Known versatile tools for data visualization and analytics – Elastic Stack, Tableau, Highcharts, and more complex database solutions like Hadoop, Amazon AWS and Teradata, have wide applications in business, from monitoring performance to improving customer experience on mobile tools. New generation of data visualization based on AR and VR technology, however, provides formerly infeasible advantages in terms of identifying patterns and drawing insights from various data streams.

Building 3D data visualization spaces, companies can create an intuitive environment that helps data scientists grasp and analyze more data streams at the same time, observe data points from multiple dimensions, identify previously unavailable dependencies and manipulate data by naturally moving objects, zooming, and focusing on more granulated areas. Moreover, these tools allow us to expand the capabilities of data visualization by creating collaborative 3D environments for teams. As a result, new technology helps extract more valuable insights from the same volume of data.

3)*Strategizing*

As the amount of data grows, it becomes harder to catch up with it. Therefore, data strategy becomes the necessary part of the success in applying data to business. Then how data visualization become an important tool in your strategic kit? First, it helps you cleanse your data. Secondly, it allows you to identify and extract meaningful information from it. Finally, data visualization tools enable continuous real-time monitoring of how your strategy and now data-driven decisions influence performance and business outcomes. In other words, these tools visualize not only the data, but also the results, and help correct and optimize strategy on the go.

Data visualization is one of the initial steps made to derive value from data. It's also one of the most important steps, as it determines how efficiently analysts can work with data assets, what insights they are able to extract and how their data strategy will develop over time.

Therefore, the quality and capabilities of data visualization directly influence how data impacts your business strategy and what benefits data applications can bring to the companies and their industries._

**Implications of (good and bad) data visualization**

Raw data is often meaningless or their meaning is not easily concluded. When people face a large set of measurements they are unable or unwilling to spend the time required to process it. Our modern living contributes to an ever-growing pool of "big data" and our ability to collect this type of information becomes easier and easier. Thus filtering, visualization, and interpretation of data become increasingly important.

We should understand what to do with data, but first we should understand why their presentation in graphical format is so powerful.

- EASY RECALL - People can process images more quickly than words.As data are transformed into imagery, the readability and cognition of the content greatly improve. While people can only seem to remember just 10% of what they hear and only 20% of what they read, retention jumps up to 80% when they see visual ingormation and do some modifications in them.
- PROVIDES WINDOW FOR PERSPECTIVE- With infographics you can pack a lot of information into a small space. Colors, shape, movement, contrast in scale and weight, and even sound can be used to denote different aspects of the data allowing for multi-layered understanding. Below is an example for a good graph: Reference: (Mullis, 2012a)
- ENABLES QUALITATIVE ANALYSIS- Color, shape, sounds, and size can make evident relationships within data very intuitive. When data points are represented as images or components of an entire

scene, readers are able to see the big picture and understand how the information fits within a larger context.

- INCREASE iN USER PARTICIPATION- Interactive infographics can substantially increase the amount of time someone will spend with the content.

Because of their impact, infographics are widely used nowadays. A quick google will produce a huge array of great examples — as well as poor ones. Because while people recognize the value of information graphic design, and a number of tools are available today that make the creation of them possible for the layperson, it doesn't mean that they're all successful or even necessary.

In the example below, the information would be better presented does not easily answer the simple question: How many airplane seats are left empty each year? It could have been more clear with the numbers and comparisons. Reference: (Mullis, 2012b)

Some useful visualization methods are: 1) *Statistical graphics and infographics*; 2) *Geographical information systems (GIS)*; 3) *Graph visualization or network maps*; 4) *Data cartography.*

**Misrepresentation through Data Visualization** - (Reference: https://venngage.com/blog/misleading-graphs/)

While the ideal purpose of data visualization is to improve others' understanding of the data presented, visualization can also be used to mislead. Some of the main methods of doing so are omitting baselines, axis manipulation, omitting data, and going against graphing convention.

Omitting baselines is used to imply a greater difference between two categories, such as in poll results comparing political parties. Axis manipulation by increasing the highest value on the y-axis affects the visibility of a slope, making data with an otherwise visible trend appear flat. Omitting selected data points or narrowing the window of a graph is used to hide an overall trend, such as a graph of a stock only showing a current trend and hiding previous bubbles. Graphs can also be designed to subvert convention so that at first glance the graph is conveying the opposite message, for example, by using the reader's associations of colors and temperature to create a graph where hot is blue and cold is red.

**A basis for why we should pursue ethical data visualization** Reference: Cairo, Alberto. "Ethical Infographics: In data visualization, journalism meets engineering." The IRE Journal, Spring 2014. https://www.scribd.com/document/230474170/Ethical-Info-Graphics-In-data-visualization-journalism-meets-engineering Cairo, Alberto. The Functional Art weblog. 19 June 2016. http://www.thefunctionalart.com/2014/06/infographics-data-and-visualization.html

Alberto Cairo addresses into the ethical 'why' of data visualization in this article, while still grounding the discussion in straightforward analysis of what to do and what not to do. He emphasizes that the effectiveness of the communicative display is as important as the information itself. This makes intuitive sense because useful information is rendered utterly useless if no one can understand it.

Cairo sees data visualization as a harmonization of journalism and engineering. From these two disciplines, he takes the journalist ethos of truth-telling and honesty and combines this with an engineering focus on efficacy and efficiency. The result is a data visualization that contains accurate and relevant information which is clearly and concisely conveyed. Cairo describes himself as a "rule utilitarian" and uses this to explain why it is ethical or, in his words, "morally right," to create graphics in this way. Here, it very useful to review his blogpost introducing the article.

Essentially, the goal is to create the most good while doing the least harm. As such, conveying truthful and honest relevant information increases a persons understanding. Increased understanding and knowledge positively correlates with personal well-being.

So, the information presented must be accurate and relevant. Cairo briefly addresses guidelines for this which are applicable in all information gathering fields: beware of selection bias when choosing preexisting datasets, validate the data, and include important context. False or irrelevant information doesn't improve anyone's decision-making capacity, so it cannot enhance well-being.

Even if the information is both accurate and relevant, moral engineering pitfalls may remain. To avoid the unethical trap of inscrutable (or misleading) graphics, Cairo exhorts us to take an evidenced based approach when possible. The purpose of the graphic dictates the form it takes; aesthetic preferences should never override clarity.

Again, since the ethical purpose is to improve well-being through understanding, a graphic which is confusing or misleading is unethical, regardless of intent, since it actually creates misunderstanding for the audience. While it can be a bit jarring to think of an poorly designed graphic as "morally wrong", it is important to think of the unintended consequences of visuals which have a powerful impact on their viewers.

## 9.1 Importance of ethics in visualization

Ref:

https://backspace.com/notes/2016/01/ethics-in-data-visualization.php

Over the years, researchers and lawyers have come up with rules and practices for proper data collection and utilization, with particular attention on human subject research.

Consent of the subjects to use their data, evaluating any risk with use or collection of data or protecting anonymity of data are some of the rules that must be considered for ethical research methods. Under U.S.law for research institutions receiving federal funding, ethical aspects must be considered. These rules continue to evolve.

Research have found that even if viewers do not support an idea, data presented in charts can persuade viewers on the subject matter. It means that visualization can also be used for deception and there are lot of techniques that can produce dangerous visualization.Techniques such as truncated axis (where the y-axis does not start at zero) or using area to represent quantity (for instance comparing the size of two adjacent circles) were found to lead to wrong conclusions.

Misleading, incomprehensible, or incredible data visualization can jeopardize people's trust, goodwill, or faith in research and advocacy on vital human rights issues. Its ethical responsibility to create visualizations to give correct and faithful representation of data and subjects.

# Chapter 10

# Conclusion

Reflection, Key Learnings, Outlook

# References


**1. 3 Expert Data Visualization Tips for Grabbing Readers' Attention**

URL: https://towardsdatascience.com/3-expert-data-visualization-tips-for-grabbing-readers-attention-206d8c4621bf

*Summary*: This article found on Medium explores three important aspects to focus on when creating a data visualization. The importance of each aspect is explained along with helpful questions to ask and to help you evaluate your visualization to ensure it caters to your audience. Although it primarily focuses on the appearance of visuals, it also discusses the psychology of the reader as they're looking at a data visual, which offers a unique and useful perspective.

Here is an outline of each of the 3 aspects: 1. Know what you really want to say. We want to share patterns, trends, anomalies, etc. with others through data visuals but we must find the right things to represent.

2. Design. Visuals should be kept as simple as possible without leaving out key points. This makes sense because then the audience can focus on what's really important.

3. Labeling. This section of the article shows a nice comparison of before and after removing labels from a chart, and the 'after' chart looks much cleaner and easier to interpret.

I think often when working with data, we tend to gravitate toward including more information in a visual, so an important takeaway for me is that less is more, and not everything we want to show has to be crammed into one big-picture visual.

**2. Choose best colors for cartography visualization in a professional manner**

URL: http://colorbrewer.org

*Summary*: It has been carefully designed to be a diagnostic tool for evaluating the robustness of individual color schemes. Full use of this tool will benefit your map designs because colors (even very similar colors) are easy to differentiate when they appear in a nicely ordered sequence (such as a legend). The task of differentiating the colors, however, becomes much harder when the patterns on the map are complex, such as in the lower left corner of the diagnostic map.

It will automatically recommend the color schemes in the following aspects:

1: Can you easily distinguish every color in the random section of the map (the lower left)? If you have a ten-class map, you should be able to see clearly ten unique colors.

2: Within each large band of color on the map, we placed several polygons filled with each map color ('outliers'). For example, if you have a seven-class map, there will be six outlier colors per band, demonstrating the appearance of all map colors with each as a surrounding color. Can you see each outlier clearly? Do all pairs of outliers in the band look different? If not, perhaps you should choose a different scheme or fewer classes.

3: You can also change the settings to colorblind-friendly on this site.

**3. Visualization Tools: An introduction to tools for creating infographics, timelines and other data visualizations.**

URL:https://guides.library.harvard.edu/c.php?g=310952&p=2073191

This website lists lots of tools to do different type of visualizations, check it out.

**4. Visual Capitalist**

URL:http://www.visualcapitalist.com/category/politics/

This company/website creates visual contents in the field of business and marketing.

**5. Misleading Graph**

As a student to learn how to be a good data scientis or business analytics professional, it is important to learn how to read the chart and interpret the statistic. Graphs can be one of the best ways to present statistical information, but they can also be one of the most misleading, even when they are completely accurate. Here, I would like to share how to detect misleading graphs.Furthermore, we can learn how to improve our data visualization skills.

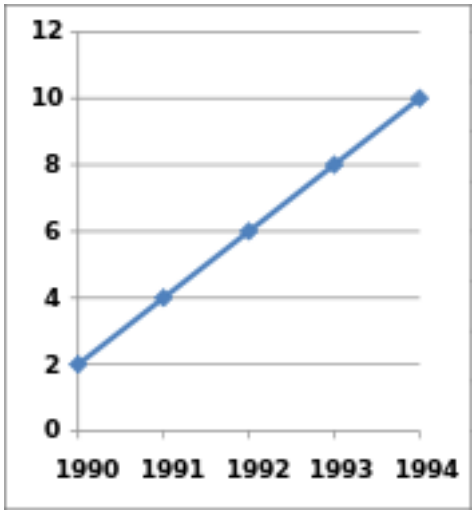
1. Omitting Baselines

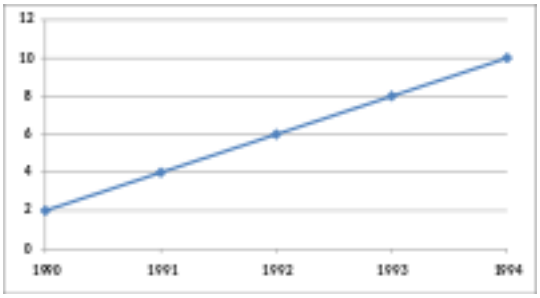
In the data visulization terms, we call it truncated graph. A truncated graph (also known as a torn graph) has a y axis that does not start at 0. These graphs can create the impression of important change where there is relatively little change.Truncated graphs are useful in illustrating small differences.[16] Graphs may also be truncated to save space. Commercial software such as MS Excel will tend to truncate graphs by default if the values are all within a narrow range. Truncating graphs make the readers to change their judement for something that is not significant looks like a huge differece.


A example of using good data in a misleading graph to fool readers comes from Fox News.

2. Axis Manipulation

Another trick of misleading graph is axis change: Changing thy y-axis maximum afftect how the graph look like. A higher maximum will make the grpha to appear less volatiliy ,less strrp than a lower maximum. The other way of axis change is changing the ratio of a graph's dimensions. This way will affect how the graph appears. We demostrate chaning the ratio of graph dimension for



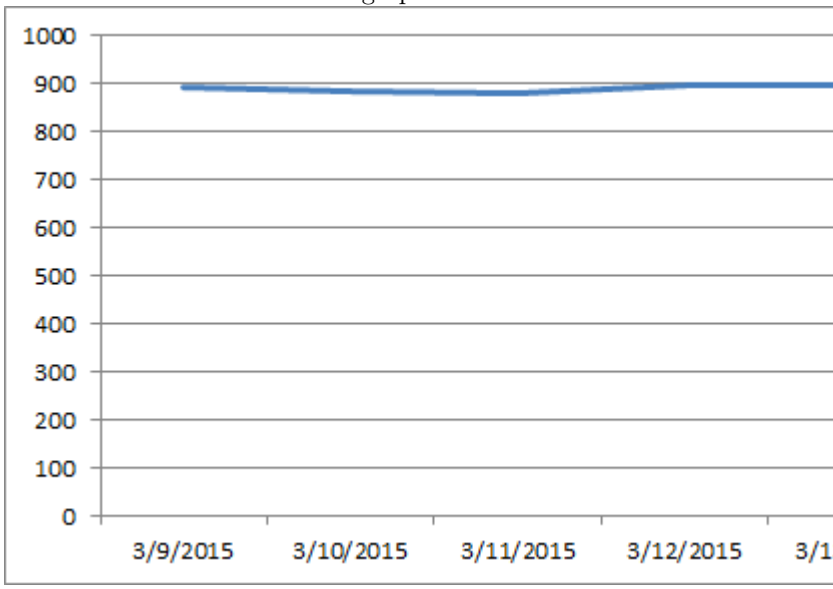



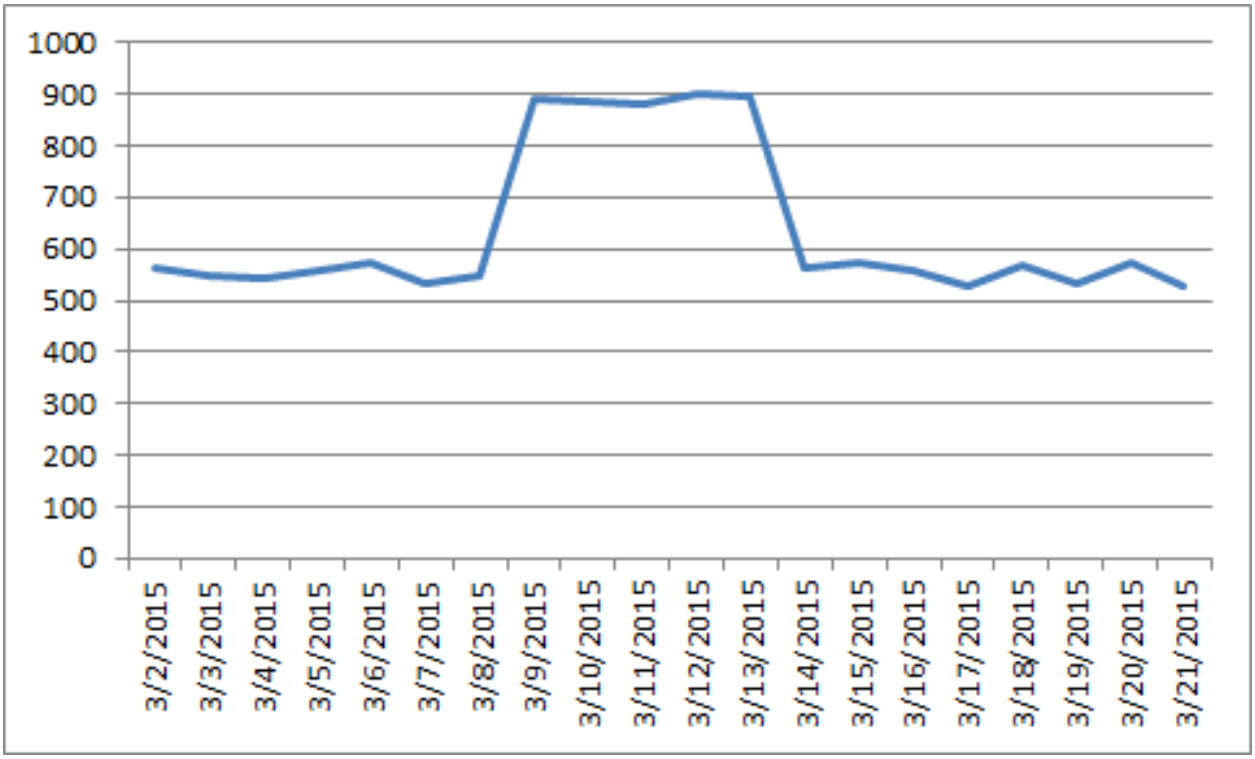
below graphs.


Axis manipulation is the opposite of truncating data, because they include the axis and baselines but change them so much that they lose meaning. This type of graph manipulation can be used to push a false narrative.

3.Cherry Picking Data This is to pick the data that shows a typical viewpoint. For example, we know house price in Bay area kept increasing since 2011. However, for those house agencies who want convince buyers that house prices has decreased , they might select some areas in typical months that house prices happend to decreased.

It is not technically wrong but it is definitely misleading.This is often called improper extraction or tactic omitting data, when only a certain chunk of data is included.This is more common in graphs that have time as one of their axis. Here is the graph to show what it is.






Reference: How Writers Use Misleading Graphs To Manipulate You BY RYAN MCCREADY, AUG 10, 2017 https://venngage.com/blog/misleading-graphs/ Misleading graph, wikipedia https://en.wikipedia.org/wiki/Misleading_graph#Truncated_graph Data Analysis: Displaying Data - Deception with Graphs https://web.archive.org/web/20030402093134/http://www.sao.state.tx.us/Resources/Manuals/Method/data/12DECEPD.pdf

4. The Year in Visual Stories and Graphics: New York Times

http://www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/index.html#dataviz

Every year, New York Times will select a collection of the year's best storyteller visualizations, which includes different forms of state-of-the-art news visualizations. Personally, I think it can be a good inspiration when we feel like "I don't know how to be creative on this".

# Bibliography


A great visualizer (1982). A ficticious web page title. http://great_viz_org/. Accessed: 2018-04-26.

AbsentData (2017). Advantages and disadvantages of tableau. https://www.absentdata.com/advantages-and-disadvantages-of-tableau/. Accessed: 2018-04-28.

Aischx, G. (2017). Using data visualization to find insights in data. http://datajournalismhandbook.org/1.0/en/understanding_data_7.html.

Andalde, S. (2014). Misleading graphs: Real life examples. Accessed: 2018-04-26.

Blog, C. S. (2017). D3.js for dynamic data reporting. https://hackernoon.com/d3-js-the-perfect-dynamic-platform-to-build-amazing-data-visualizations-ebe930f0648f. Accessed: 2018-04-28.

Castañón, J. (2016). Shiny: a data scientist's best friend. https://medium.com/ibm-data-science-experience/shiny-a-data-scientists-best-friend-883274c9d047. Accessed: 2018-04-28.

Halper, D. (2012). Over 100 million now receiving federal welfare. Accessed: 2018-04-26.

Henry, K. (2017). In defense of pie charts, and why you shouldn't use them. https://medium.com/@KristinHenry/in-defense-of-pie-charts-and-why-you-shouldnt-use-them-df2e8ccb5f76.

Hickey, W. (2013). The worst chart in the world. http://www.businessinsider.com/pie-charts-are-the-worst-2013-6.

Internet (2017). Reasons to have dynamic data reporting. http://www.convergesolution.com/blog/blog-post/5-reasons-why-you-should-have-dynamic-data-reporting-solution-datavisualization/. Accessed: 2018-04-28.

Mullis, L. (2012a). The impact of data visualization. Accessed: 2018-04-28.

Mullis, L. (2012b). The impact of data visualization. Accessed: 2018-04-28.

Quach, A. (2016). Why pie charts often suck: And how we did better. https://medium.com/the-mission/to-pie-charts-3b1f57bcb34a.

Unwin, A. (2008). The applications of data visualization. Accessed: 2018-04-28.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.7.