A Reader on Data Visualization

MSIS 2629 Spring 2018

2018-04-26

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},
```

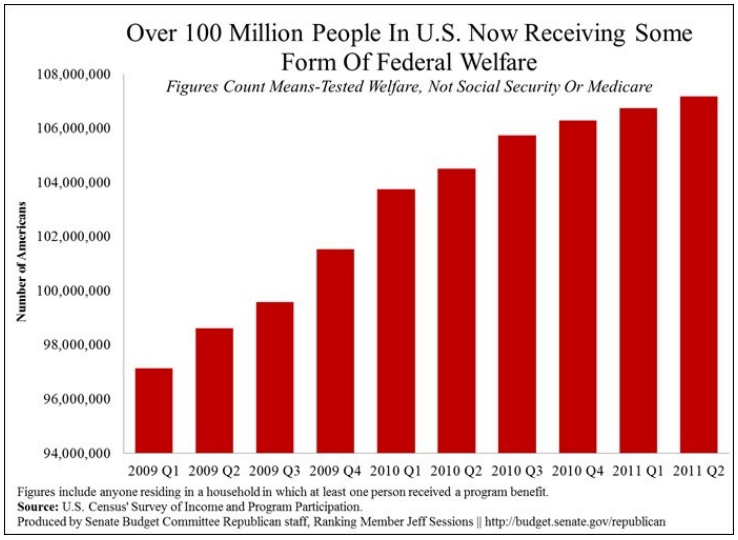

Figure 1.1: An Example of a deceptive visualization

```
  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.

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

https://infogram.com/page/data-visualization

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

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

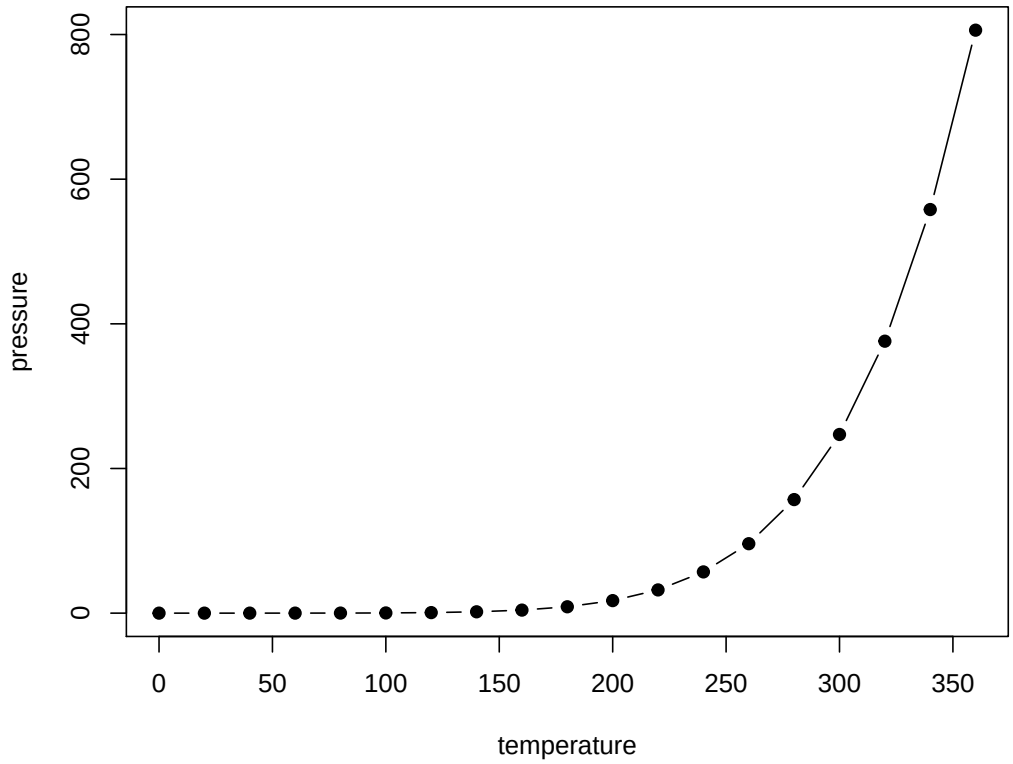

Figure 2.1: Here is a nice figure!

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

Chapter 3

Fundamentals

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 acquainted 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

- Theoretical background of data visualization

Practitioners Guide to Best Practices in Data Visualization

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

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.

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 has can have 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.

- Theoretical background of data visualization
  - 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 relationship.

  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.

- Theoretical background of data visualization

  1. 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.

- Theoretical background of data visualization

A Brief History of Data Visualization

Reference

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 modern times, 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.

- Contemporary research results
- Contemporary research results reference—fundamentals–example-USDATA

Contemporary research results

- Next Steps for Data Visualization Research
- references: https://medium.com/@uwdata/next-steps-for-data-visualization-research-3ef5e1a5e349

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.
- Contemporary research results

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

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/

- Theoretical background of data visualization

Definions of Date Deception and Graphic Integrity

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.

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 be directly proportional to the numerical quantities measured.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out 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, deflated and standardized units of monetary measurement are nearly always better than nominal units.
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

6.Graphics must not quote data out of context.

Chapter 4

Case Studies

- Description and replication of great examples of data visualization ### 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.

4.1 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

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.

Chapter 5

Patterns

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

5.1 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.
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.
3. Pie chart is easy to be manipulated (e.g. using a 3D pie chart).
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!

5.2 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 .

```
# Load the data.
#setwd("/Users/boxiao/Desktop")
data<-read.csv("galaxy_sales.csv",header = TRUE)
data <- rbind(data[1:19,],c("Q4 '14",80),data[20:22,],c("da", 83))
data$year <- rep(c(2010,2011,2012,2013,2014,2015),each = 4)
data$qua <- seq(1:24)
data$Quarter <-NULL
```

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/

Chapter 6

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.*

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.

Chapter 7

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.

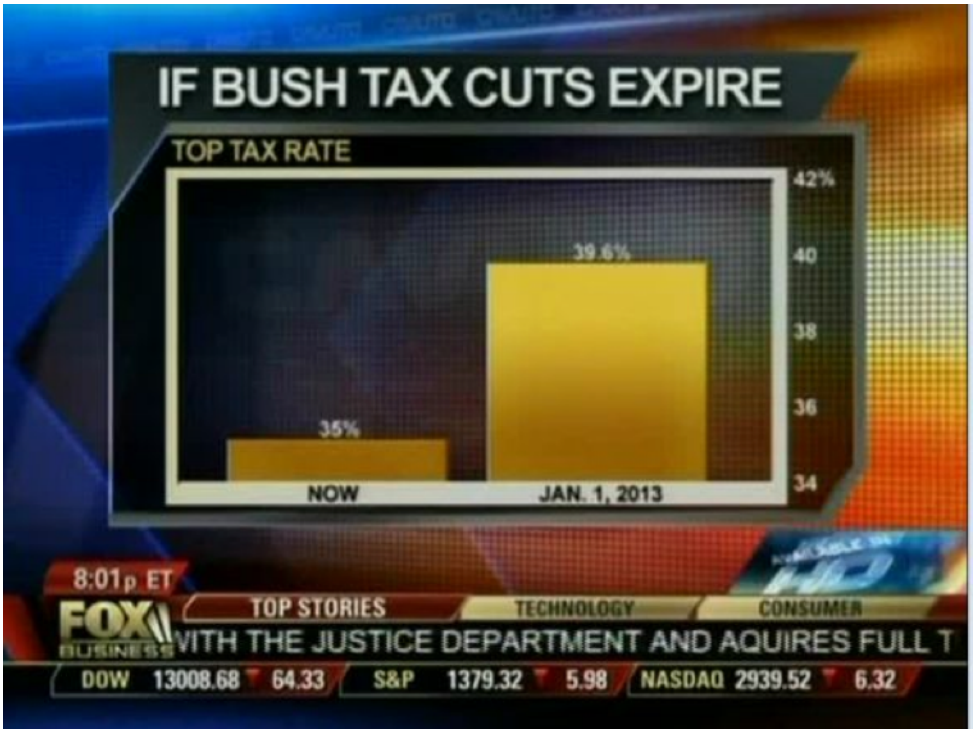

Figure 7.1:

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.

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

Chapter 8

Fundamentals

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/

Bibliography

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

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

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.

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

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.