A Reader on Data Visualization

MSIS 2629 Spring 2019

2019-05-20

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 explains how we organize ourselves.

1.1 References

EVERY reference 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 (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},
```

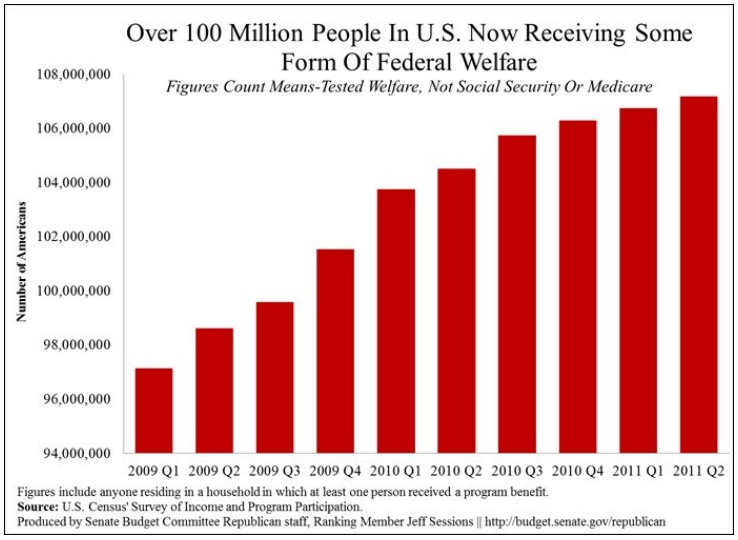


Figure 1.1: An Example of a deceptive visualization

```
  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}
```

1.3 Basic Guidelines

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 **??**. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table **??**.

```
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).

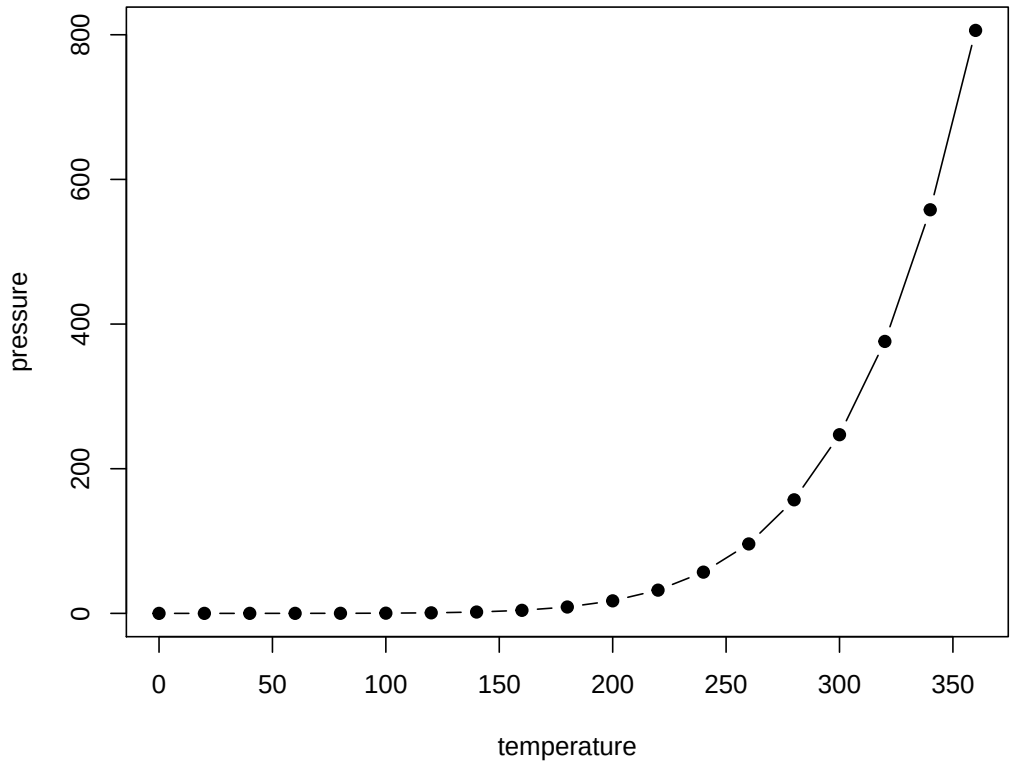


Figure 1.2: Here is a nice figure!

Table 1.1: 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 2

Introduction

2.1 What is Data Visualization?

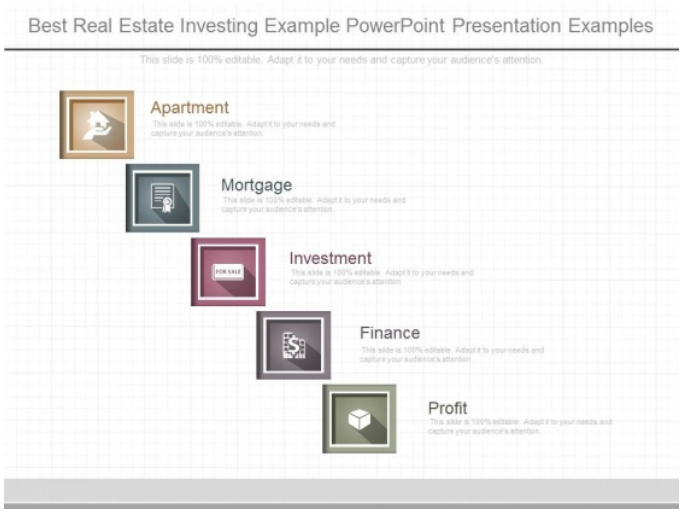
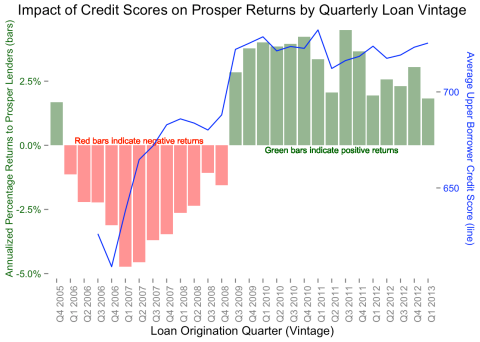
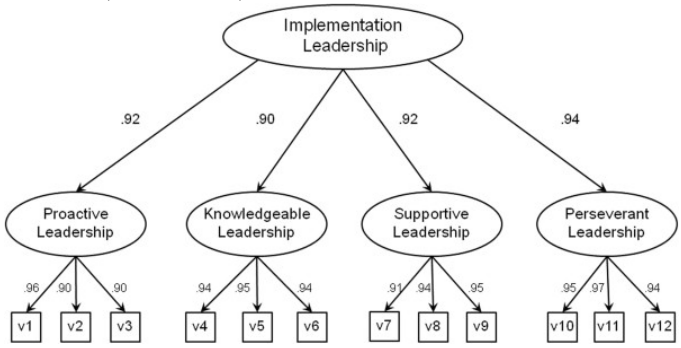
Data visualization is a graphic representation that expresses the significance of data. It reveals insights and patterns that are not immediately visible in the raw data. It is an art through which information, numbers, and measurements can be made more understandable. According to (Friedman 2008):

> The primary goal of data visualization is to communicate information clearly and effectively through graphical means. It does not mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects more intuitively.

The main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way.

"Data is the new oil" may be a cliche, but it is true. Like oil, data in its raw, unrefined form is pretty worthless. To unlock its value, data needs to be refined, analyzed and understood (Disney 2017). More and more organizations are seeing potential in their data connections, but how do you allow non-experts to analyze data at scale and extract potentially complex insights? One answer is through interactive graph visualization.

Information visualization is the art of representing data so that it is easy to understand and manipulate, thus making the information useful. Visualization can make sense of information by helping to find relationships in the data and support (or disproving) ideas about the data. (Disney 2017) shares some examples and common uses of information visualization, such as:

Benefit	Example
Presentation: to explain or persuade	 Source: ("Best Real Estate Investing Example Powerpoint Presentation Examples,PPT:SG-14716000770769" 2019)
Exploratory Analysis: to identify relationships or unusual cases in the data	 Source: (smith 2019)
Confirmatory Analysis: to confirm our understanding and analysis of the data	 Source: (Aarons 2014)

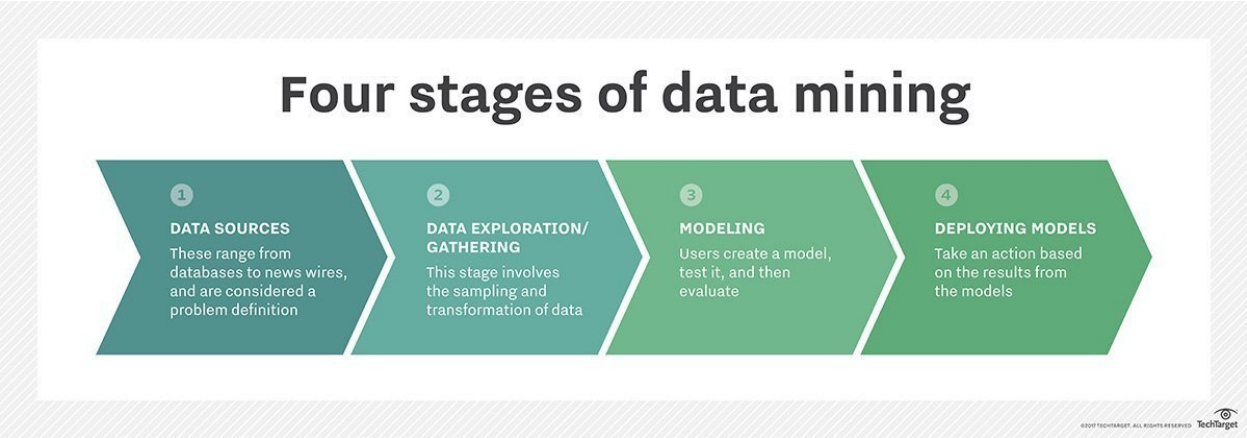
Why data visualization is such a powerful tool:

- Intuitive: Presenting a graph as a node-link structure instantly makes sense, even to people who have never worked with graphs before.
- Fast: It is fast because our brains are great at identifying patterns, but only when data is presented in a tangible format. Armed with visualization, we can spot trends and outliers very effectively.
- Flexible: The world is densely connected, so as long as there is an interesting relationship in your data somewhere, you will find value in graph visualization.
- Insightful: Exploring graph data interactively allows users to gain more in-depth knowledge, understand

the context and ask more questions, compared to static visualization or raw data.

2.2 Importance of Data Exploration

"Data exploration tasks are those of examining data without having an a priori understanding of what patterns, information, or knowledge it might contain." (Jeff Baker 2009) While the majority of research has been on the output of and creation of visualizations, but the important aspect of these tasks is to first understand the data that is being presented. As mentioned in (Jeff Baker 2009), the main aspect of exploration is the understanding of perception in comparing two measures in a data source. The contrast between object A and object B shows an ability to compare by saying A is greater than B or the opposite occurs. As a brief introduction, data exploration is the process of finding the best way to pull out an outcome that a specific audience can perceive. By understanding the different types of perceptions and studying data based off of these concepts, data exploration is pivotal in the final visualization process due to taking the audience of the viz into account first and foremost. Here are some data exploration methods for references: Companies can conduct data exploration via a combination of automated and manual methods. Analysts commonly use automated tools such as data visualization software for data exploration because these tools allow users to quickly and simply view most of the relevant features of a data set. From this step, users can identify variables that are likely to have interesting observations.



By displaying data graphically – for example, through scatter plots, density plots or bar charts – users can see if two or more variables correlate and determine if they are good candidates for further analysis, which may include:

Univariate analysis: The analysis of one variable. Bivariate analysis: The analysis of two variables to determine their relationship. Multivariate analysis: The analysis of multiple outcome variables. Principal components analysis: The analysis and conversion of possibly correlated variables into a smaller number of uncorrelated variables. Manual data exploration methods may include filtering and drilling down into data in Excel spreadsheets or writing scripts to analyze raw data sets.

2.3 Importance of Data Visualization

According to the World Economic Forum, the world produces 2.5 quintillion bytes of data every day, and 90% of all data has been created in the last two years. With so much data, it's become increasingly difficult to manage and make sense of it all. It would be impossible for any single person to wade through data line-by-line and see distinct patterns and make observations. Data proliferation can be managed as part of the data science process, which includes data visualization("Data Visualization," n.d.). Data visualization is the presentation of data in a pictorial or graphical format. It enables decision-makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you

can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed("History of Data Visualization," n.d.).

Data visualization has become an indispensable part of the business world and an ever increasing part of managing our daily life. Effective data visualization should be informative, efficient, appealing, and in some cases interactive and predictive. Pittenturf explains basic criteria that a data visualization should satisfy to be effective (Pittenturf 2018):

Criteria	**Description**
Informative	The visualization should be able to convey the desired information from the data to the reader.
Efficient	The visualization should not be ambiguous.
Appealing	The visualization should be captivating and visually pleasing.
Interactive and Predictive (Optional)	The visualizations can contain variables and filters with which the users may interact to predict results of different scenarios.

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

Another early adopter of graph visualization techniques was the financial services industry. Fraud detection is about finding unusual connections between accounts, transactions, insurance policies, devices, etc. There is great value in visualizing that data as a graph. Known fraud detection is primarily automated with rule scoring and pattern matching. Visualization lets you review edge-cases and outliers more quickly. Speed is important because sometimes analysts only have seconds to approve or deny a transaction. In those cases, visualizations are simple, small and with limited interaction. To get a clear overview quickly, analysts need effective layouts. Other functionalities, like expanding and filtering help fraud analysts to see context on demand.

Three things are consistent across both graph visualization use cases:

They involve highly connected data (apparently).

That highly connected data conceals risk insight.

That insight is needed to power quick and confident decision making.

"The only new thing in the world is the history you do not know." Harry S Truman

Given the recent explosion in data availability and visualization tools, it would be natural to assume that statistical graphics and data visualizations are relatively modern developments. However, data visualization is not a modern product it has developed over time to incorporate the tools we use today and the trends we foresee. The graphic representation of quantitative information has deep roots that reach into the histories of the earliest map-making and visual depictions, and up to thematic cartography, statistics, medicine, and other fields. Developments in technologies (printing, reproduction) mathematical theory and practice, and empirical observation and recording, and those developments enabled the broader use of graphics and new advances in form and content. It is essential to gain some understanding of the background of data visualization to help us in the proper application and execution of current visualization concepts. The

following section provides an overview of the intellectual history of data visualization from medieval to recent times, as well as describes and illustrates some significant advances along the way (Friendly 2006).

Time	Phase	Description
Pre-17th Century	Early Maps and Diagrams	Data visualization has come a long way. Before the 17th century, data visualization already existed. Though displayed in other formats such as maps, the content is much similar to today's visualizations, which mostly presented geologic, economic, and medical data. The earliest seeds of visualization arose in geometric diagrams, in tables of the positions of stars and other celestial bodies, and in the making of maps to aid in navigation and exploration.
1600-1699	Measurement and Theory	Among the most important problems of the 17th century were those concerned with the physical measurement of time, distance, and space for astronomy, surveying, map making, navigation and territorial expansion. This century also saw considerable new growth in theory as well as the dawn of practical application.
1700-1799	New Graphic Forms	With some rudiments of statistical theory, data of interest and importance, and the idea of graphic representation somewhat established the 18th century witnessed the expansion of these aspects to new domains and new graphic forms.
1800-1850	Beginnings of Modern Graphics	With the foundation provided by the previous innovations of design and technique, the first half of the 19th century witnessed explosive growth in statistical graphics and thematic mapping at a rate which would not equal until modern times.
1850–1900	The Golden Age of Statistical Graphics	By the mid-1800s, all the conditions for the rapid growth of visualization had generated a "perfect storm" for data graphics. Official state statistical offices were established throughout Europe, in recognition of the growing importance of numerical information for social planning, industrialization, commerce, and transportation.
1900-1950	The Modern Dark Ages	If the late 1800s were the "golden age" of statistical graphics and thematic cartography, the early 1900s can be called the "modern dark ages" of visualization. There were few graphical innovations, and by the mid-1930s, the enthusiasm for visualization which characterized the late 1800s had been supplanted by the rise of quantification and formal, often statistical, models in the social sciences.
1950–1975	Rebirth of Data Visualization	Still under the influence of the formal and numerical zeitgeist from the mid-1930s on, data visualization began to rise from dormancy in the mid-1960s
1975–present	High-D, Interactive and Dynamic Data Visualization	During the last quarter of the 20th century, data visualization has blossomed into a mature, vibrant and multidisciplinary area of research, as seen in this handbook, and software tools for a wide range of visualization methods and data types are available for every computer.

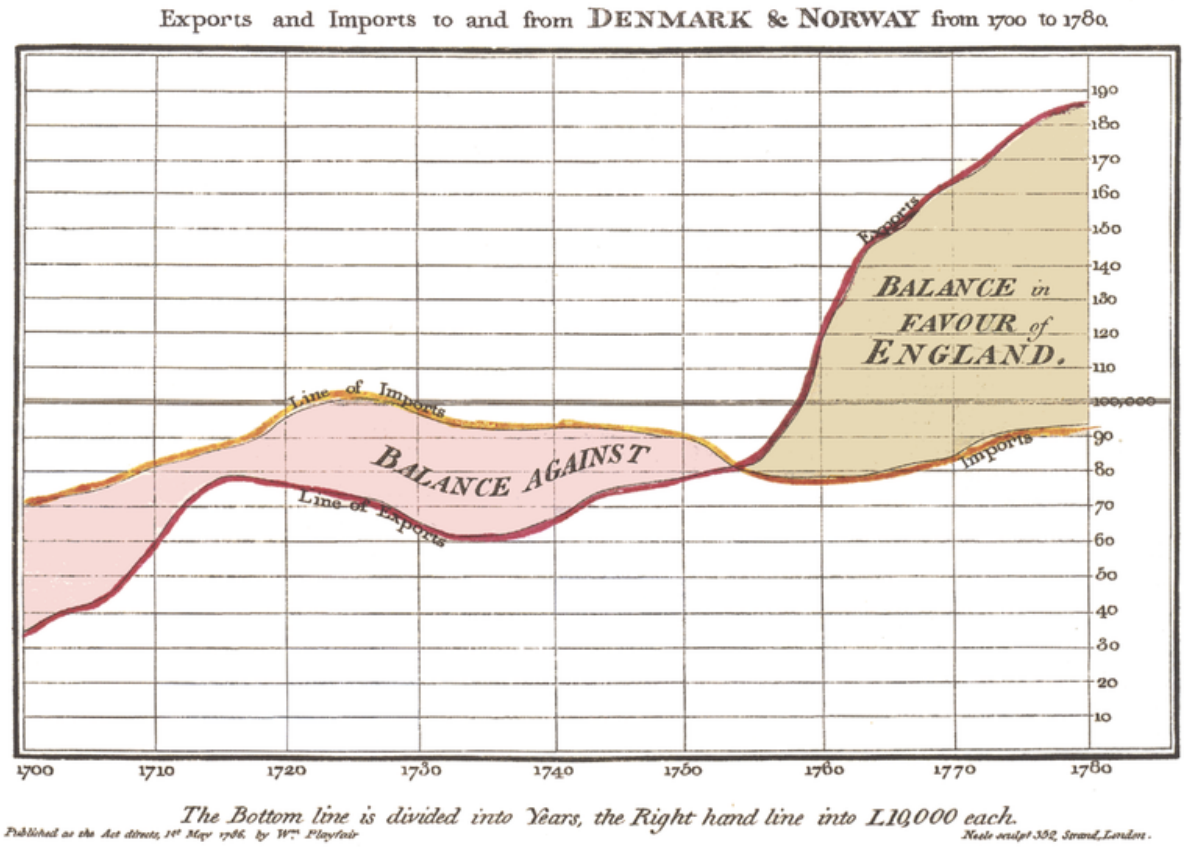
2.3.1 Key Figures in the History of Data Visualization

The idea of visualizing data is old: After all, that's what a map is—a representation of geographic information—and we've had maps for about 8,000 years. But it was rare to graph anything other than geography. (n.d.)(https://www.smithsonianmag.com/history/surprising-history-infographic-180959563/) The history of data visualization is full of incredible stories marked by significant events, led by a few key players. The article (Infogram 2016) introduces some of the fantastic men and women who paved the way

by combining art, science, and statistics. One of them is Charles Joseph Minard, whose most famous work is the map of Napoleons Russian campaign of 1812 which could use as a data product for Data Visualization. Below we have some visualizes with their famous works and other stories in the article (Infogram 2016).

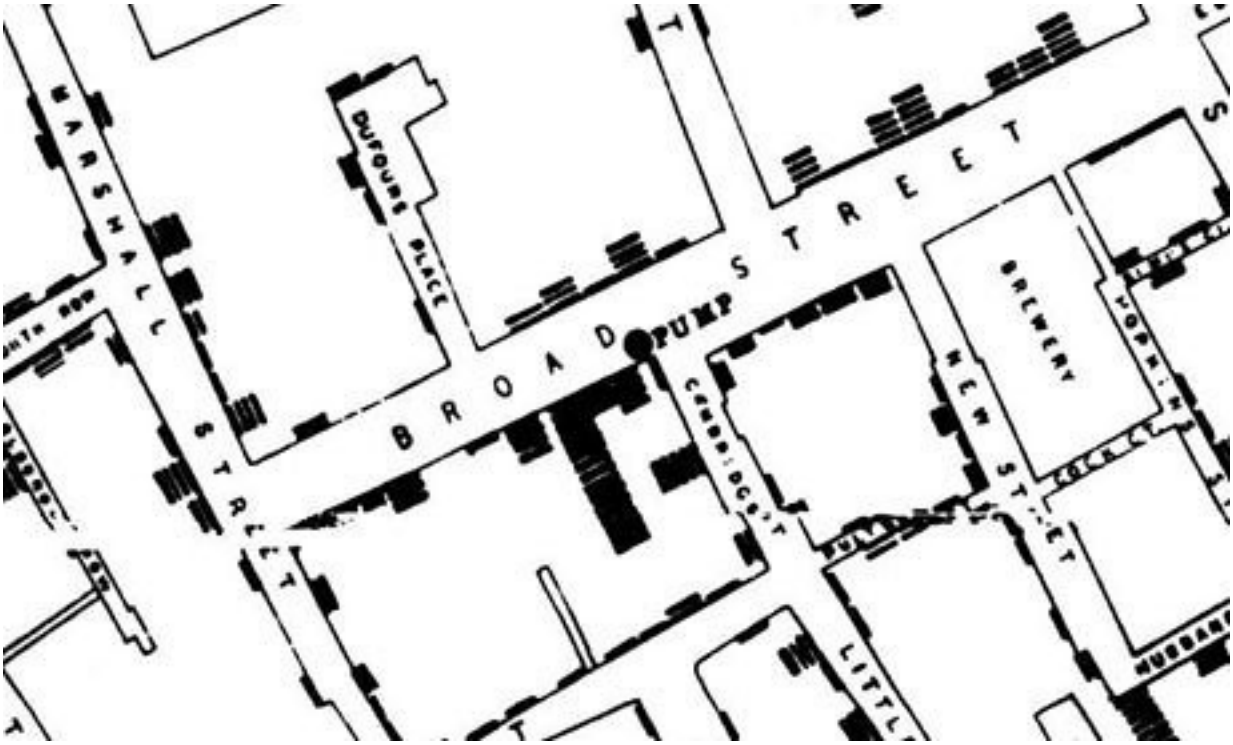
2.3.1.1 William Playfair (1759–1823)

William Playfair is considered the father of statistical graphics, having invented the line and bar chart are used 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.



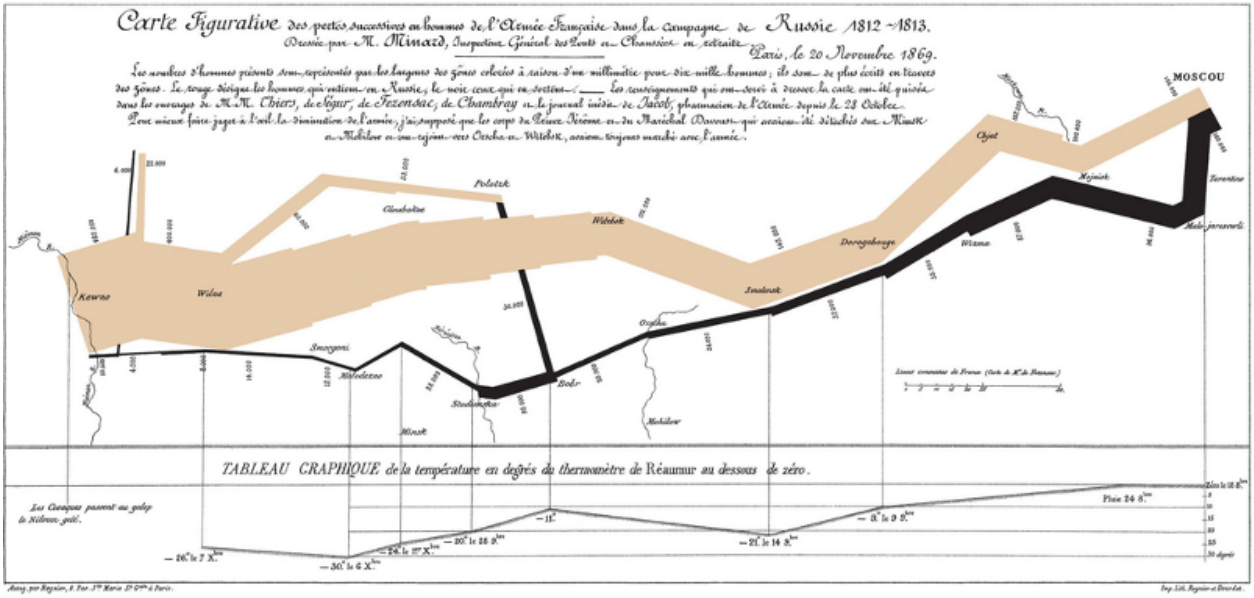
2.3.1.2 John Snow (1813–1858)

In 1854, a cholera epidemic spread quickly through Soho in London. The Broad Street area had seen over 600 dead, and the surviving residents and business owners had primarily 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 an evident concentration around the water pump on Broad Street, which helped find the cause of the infection.



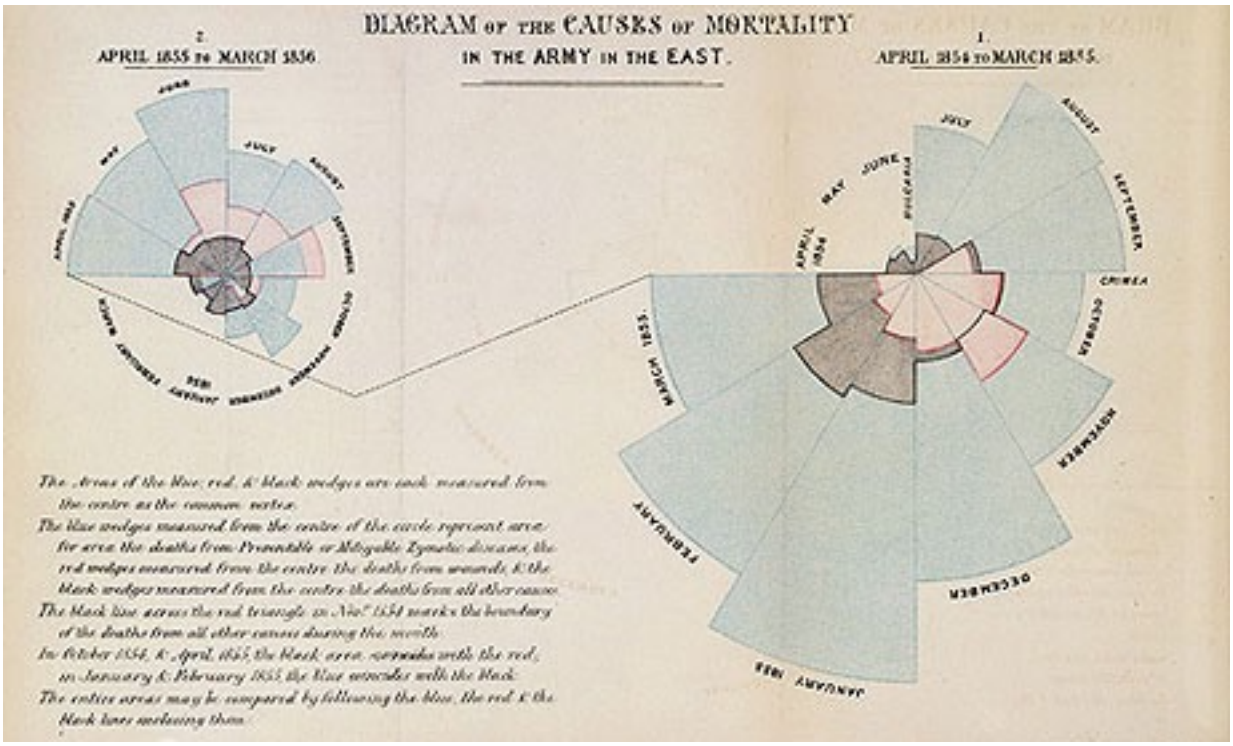
2.3.1.3 Charles Joseph Minard (1781–1870)

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 illustrating the dramatic loss of his army over the advance on Moscow and the following retreat. 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 has been called one of the "best statistical drawings ever created." The work is an essential 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 proper without a firm grasp of your dataset and sharp communication skills. It represents the earliest beginning of data journalism.



2.3.1.4 Florence Nightingale (1820–1910)

Florence Nightingale is famous for her work as a nurse during the Crimean War, but she was also a data journalist. She realized soldiers were dying from poor sanitation and malnutrition, so she kept meticulous records of the death tolls in the hospitals and visualized the data. Her coxcomb or rose diagrams helped her fight for better hospital conditions and ultimately save lives.



(Source:(Infogram 2016))

2.3.1.5 Edmond Halley (1656–1742)

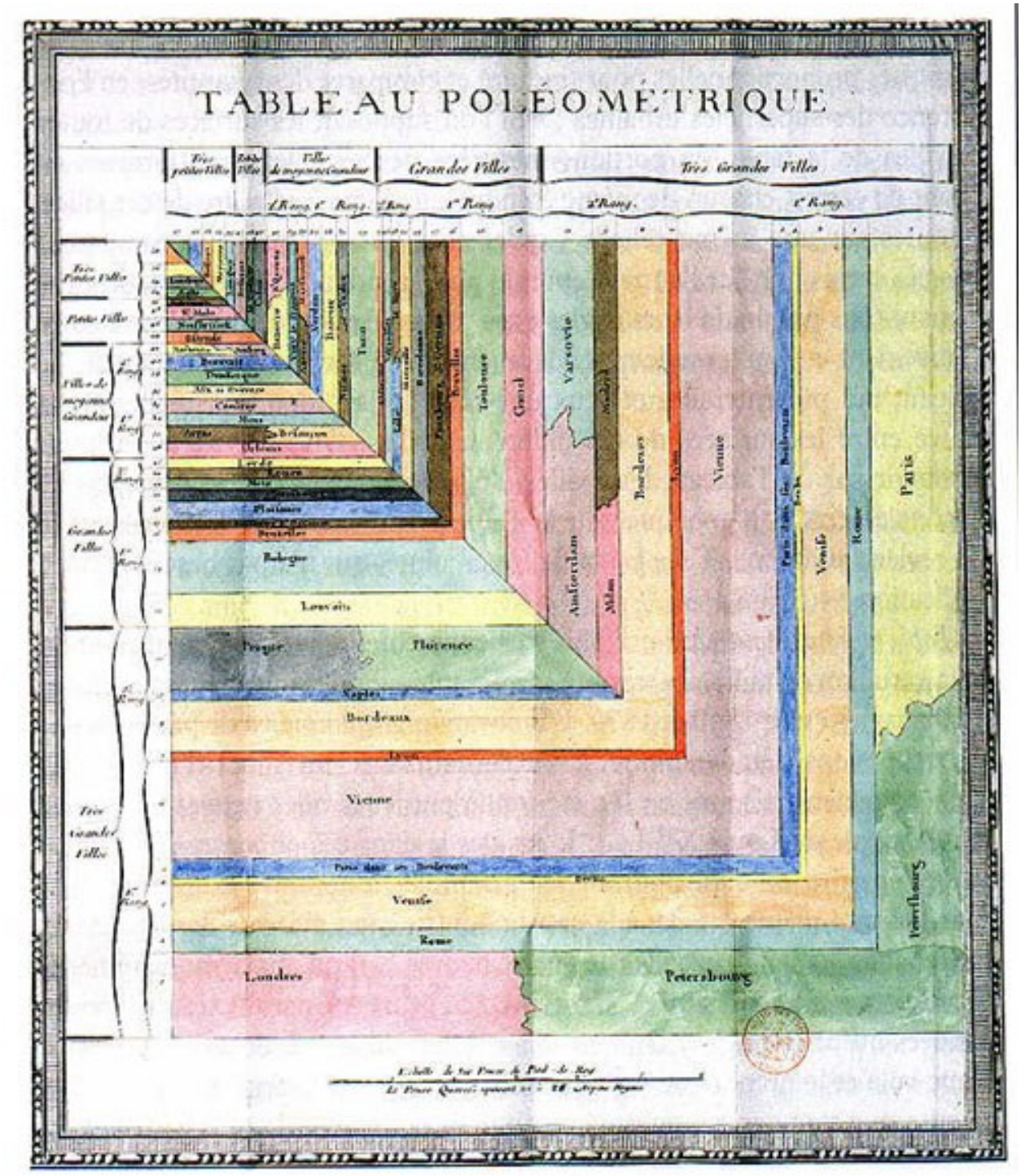
Edmond Halley was an English astronomer, geophysicist, mathematician, meteorologist, and physicist who is best known for computing the orbit of Halley's Comet. According to the BBC, Halley developed the use of contour lines on maps to connect and describe areas that display differences in atmospheric conditions from place to place. These lines are now commonly used to describe meteorological variation common to us from weather reports.

(Source:(Infogram 2016))

2.3.1.6 Charles de Fourcroy (1766–1824)

Charles de Fourcroy was a French mathematician and scholar. He produced a visual analysis of the work of French civil engineers and a comparison of the demographics of European cities.In 1782 he published Tableau Poléometrique, a treatise on engineering and civil construction. His use of geometric shapes predates the

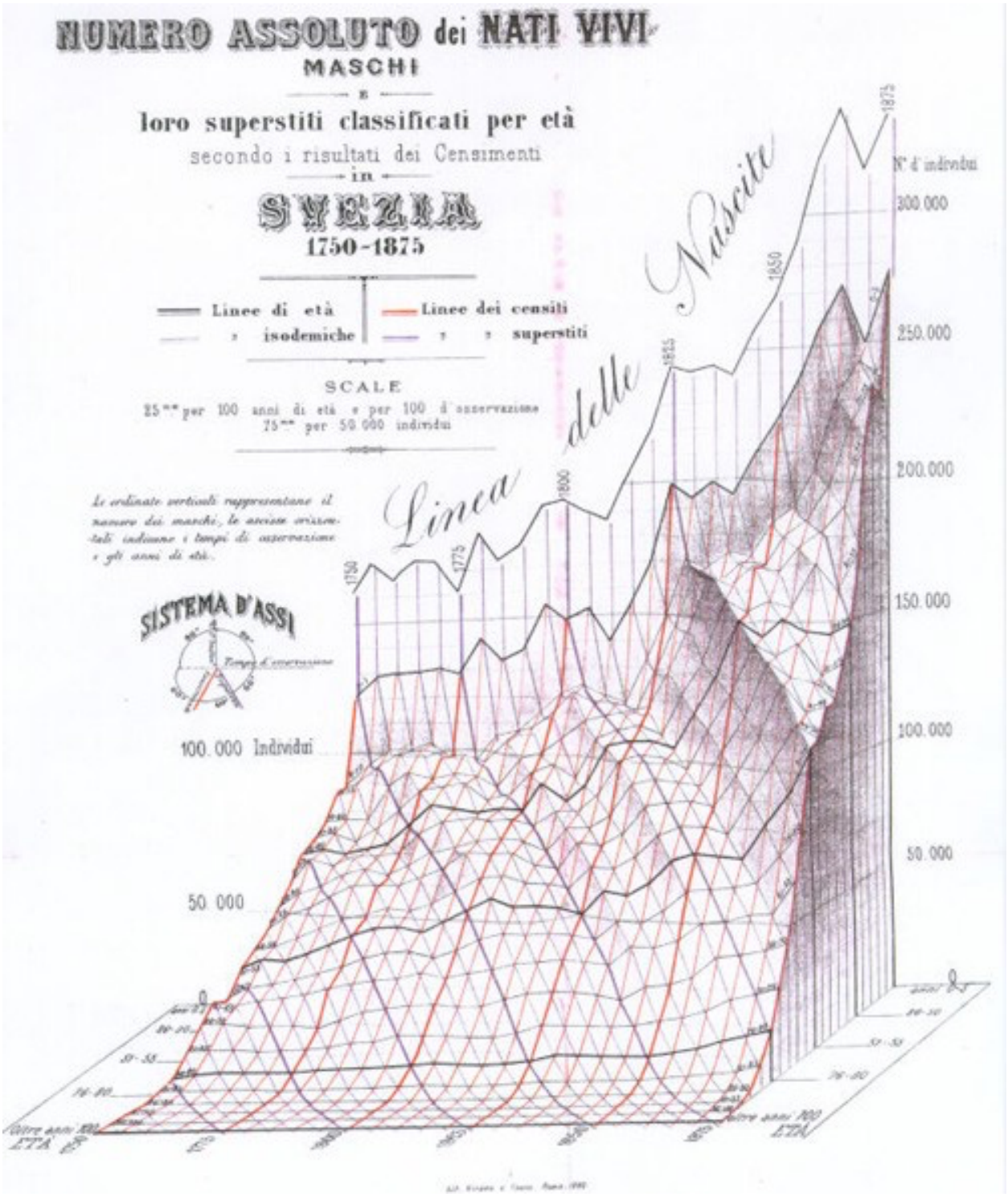
modern treemap, which is widely used today to display hierarchical data.



(Source:(Infogram 2016))

2.3.1.7 Luigi Perozzo (1856–1916)

Luigi Perozzo was an Italian mathematician and statistician who stood out for being the first to introduce 3D graphical representations, showing the relationships between three variables on the same graph. Perozzo published one of the first 3D representations of data showing the age group of the Swedish population between the 18th and 19th centuries.



(Source:(Infogram 2016))

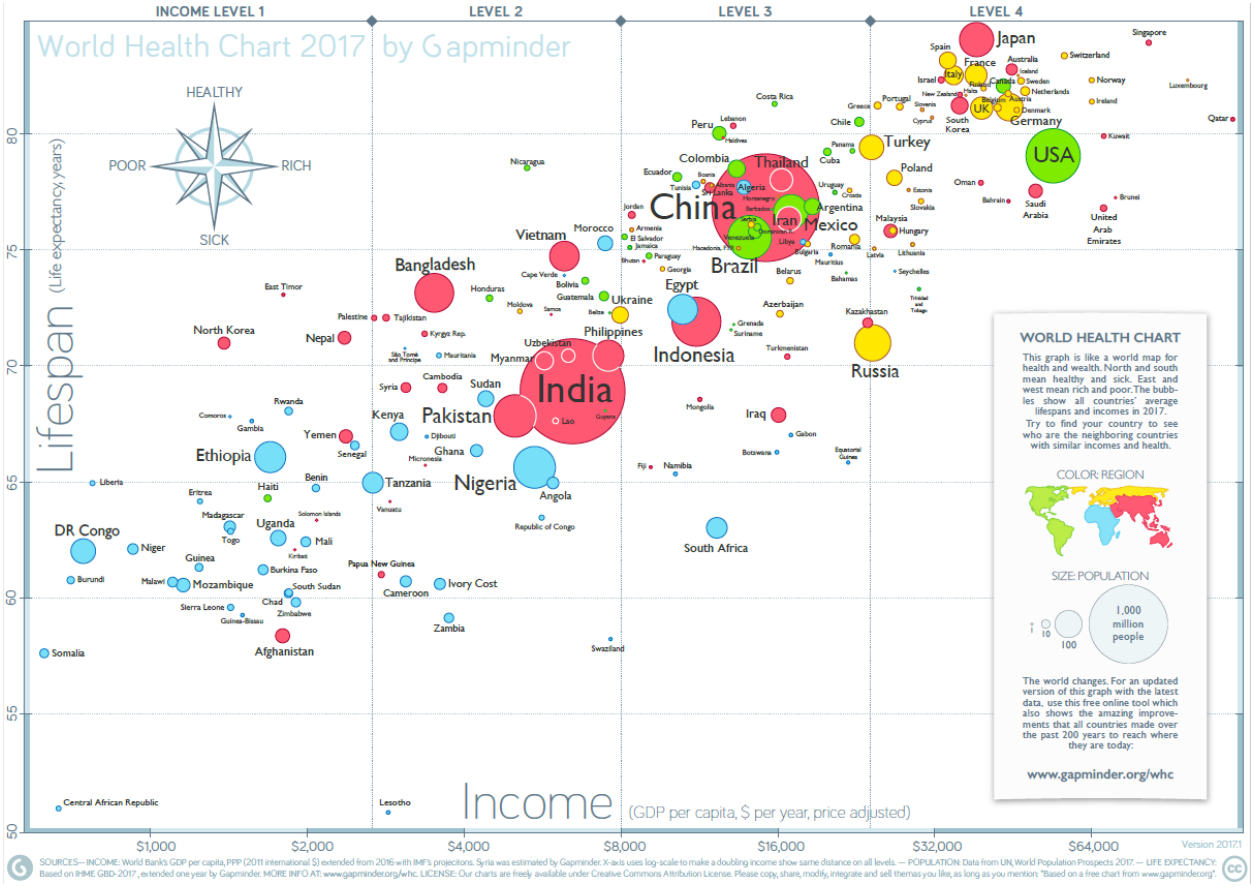
2.4 Contemporary Visualists

2.4.1 Hans Rosling

Hans Rosling was a Swedish physician, academic, statistician, and public speaker. He was the Professor of International Health at Karolinska Institute and was the co-founder and chairman of the Gapminder Foundation, which developed the Trendalyzer software system.

Rosling was born in Uppsala, Sweden, on the 27th of July 1948. From 1967 to 1974 Rosling studied statistics and medicine at Uppsala University, and in 1972 he studied public health at St. John's Medical College, Bangalore, India. He became a licensed physician in 1976 and from 1979 to 1981 he served as District Medical Officer in Nacala in northern Mozambique. On 21 August 1981, Rosling began investigating an outbreak of konzo, a paralytic disease first described in the Democratic Republic of the Congo. His investigations earned him a Ph.D. at Uppsala University in 1986.

Rosling took his interest in global health and developed stunning visualizations about it using statistical methods and data from the UN. He was a noted TED speaker and one of his most interesting TED talks is *"Asia's Rise: How and When"* (Rosling 2009). In this video, Hans shows trends of the Western countries vs. Developing countries like India and China and makes predictions using stunning visualizations like the Bubble chart. He also predicts the exact date on which India and China will move ahead of USA as strong economic forces.



In this chart Rosling describes the relative standings of countries in terms of Income versus Lifespan. He

started his analysis from year 1856 and this is where various countries stand till now. With the current growth trends, Rosling predicted that India and China will once again be the global leaders in Income and Healthcare by the year 2048.

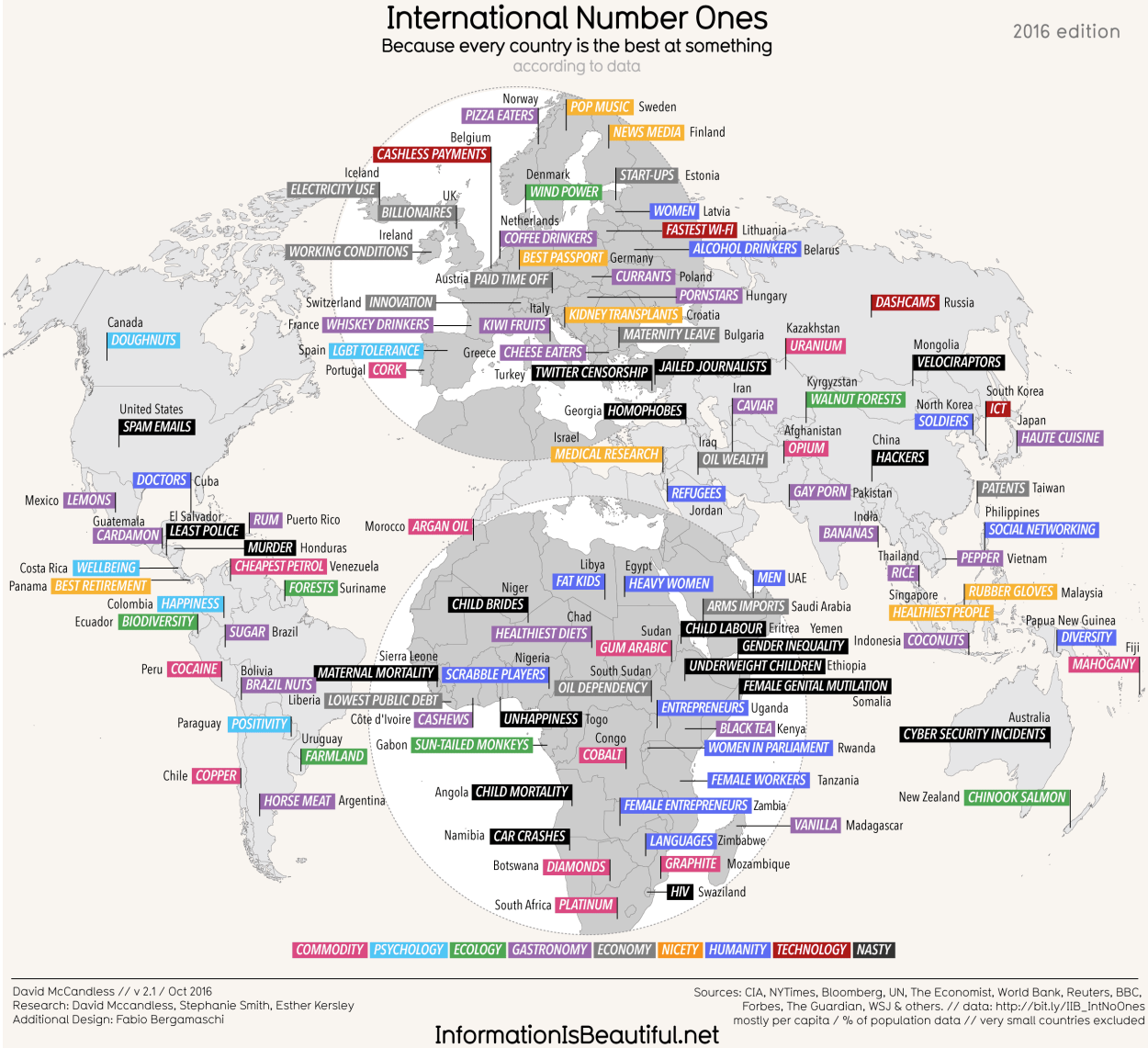
Hans was the co-founder and developer of the Gapminder Foundation which is committed to promoting global sustainable development through the understanding of statistics and data describing issues at the local, national, and global level. One of the most important goals of the Gapminder foundation is to end ignorance in the world by developing fact-based visualizations to present the world as it is (Ruan et al. 2017).

As part of Gapminder, Hans Rosling has developed many interactive and intuitive visualizations. Gapminder provides free teaching material to dismantle misconceptions and promote a fact-based worldview (Rosling, n.d.).

2.4.2 David McCandless

David McCandless is a British data-journalist and his blog *Information is Beautiful* (McCandless 2018) hosts some of the most visually stunning graphs, charts, and maps on a wide range of topics like science, food, dogs and countries. One such chart, *International Number Ones: Because every country is good at something (according to data),* is a captivating work that displays something each country is the best at (McCandless 2016). The visualizations on this website are updated and revised whenever new data is available.

One such chart, *International Number Ones: Because every country is good at something (according to data),* is a captivating work that displays something each country is the best at (McCandless 2016). The visualizations on this website are updated and revised whenever new data is available.



2.5 How to tell stories using data visualization' and 'exploratory data visualization' theory example

2.5.1 A cool website: MIT Media Lab

(Deloitte, Datawheel, and Hidalgo 2016) MIT Media Lab in collaboration with Deloitte has created a new visualization tool, that aggregates US government open source data and mines information to generate trends and stories about cities, jobs, industries, etc. Just looking at any of the open data sources would give us an idea about the vastness (breadth and depth) of the available data. It is impressive that they have brought it all together on a single platform in a convenient format.

We think of a topic, and it is possible it is there! The benefits are a better understanding of our consumers, talent pool, jobs, climate, which helps improve our decision-making ability. The best part is that the data is also available for download so we can replicate the visuals, redesign and tell our own stories with this data. There are also other similar websites, that has some good visualizations on census data: (Bureau 2018a)

2.5.2 Standing out categorization of information on the website.

What stands out is the categorization of information on the website which enables the following:

- Easy browsing of various categories of information available at a single glance
- An easy search on any topic of interest and get more in-depth information
- Logical construction of information using data and visuals under each category
- Comparative analysis of cities
- Variety of exploratory visualizations to learn from
- Most important - Storytelling through data, such as the evolution of the American Worker, how poverty is bad for people's health, how men still dominate in the highest paying industries, and opioid addiction damage.

2.5.3 Automatic visualization vs. context-specific visualization

Automatic visualization is a bad idea, generally speaking. Some parts of visualization indeed should be automatic, such as standard chart types and recurring geometries. Pieces of visualization, such as annotation and axis construction can be automatic.

However, full-on automation where insight fountains out from any dataset is farfetched at this point because this requires automatic analysis. Analytics is usually context-specific and requires more than boilerplate statistics. The most interesting visualizations are context-specific. (Flowingdata 2017)

2.5.4 Data analytics software

Data analytics software can analyze vast amounts of data and incredible speeds, but how can it explain the results of that analysis? Today, the only means of doing this is with graphics. Data visualization still leaves room for interpretation. However, technology is catching up. Narrative generation software can run as a plug-in to the dashboards. Tools like Savvy install on a server and allow every dashboard user to get a written summary of demand. This software is fully plugged and play, it takes seconds to install, and it is easy to use. It runs with Microsoft Excel, Qlik Sense 3.0, and is available as an API. It is an example of how automation is making our working lives more comfortable by automating repetitive tasks and allowing us to leverage existing data reserves fully. (Manning 2016)

2.5.5 Data visualization can't explain data, leaving room for interpretation.

According to Alysson Ferreira, a UI Engineer, there is an increasing need to understand the latest trends quickly and efficiently, which means there's also a need for significant sources of trustworthy information. (Ferreira 2017)

This is where data visualization comes in. Data visualization is the art of displaying information by combining the beauty of imagery with the conciseness of statistics, which allows us to organize complex data into useful graphical representations. In simple terms, data visualization is the art of translating complex data into meaningful information. (Towler 2015)

Plug in any data set into a magic box and it spits out a beautiful visualization you can show all of your co-workers, friends, and family. That is the promise of many startups, but it does not quite work that way. The problem is that graphics alone don't fully explain data, and so we are inundated with queries: why did the numbers fall in whatever month? Data visualization cannot explain dayta, leaving room for interpretation.

Although simple visualizations such as standard chart types (bar chart, line chart, etc.) are already automated to a certain extent in Microsoft Office tools and other software available in the market, full on automation where insight fountains out from any data set are far-fetched at this point because this requires

automated analysis. The automated analysis here means that the tool or algorithm has to understand the context and also select the best visualization.

2.5.6 One great tool: D3.js.

The focus in today's world has been on open source tools and technologies and these tools although being free for the most part to require more effort to integrate to the current visualization workflow seamlessly. As mentioned in one of the articles about D3.js:

D3.js is one of the first data visualization tools that comes to mind when talking about free, open-source alternatives. It is a JavaScript-based library for creating web visualization and displays the results on the web page. However, with high power comes great responsibility. **D3.js** is one of the first data visualization tools that comes to mind when talking about free, open-source alternatives. It is a JavaScript-based library for creating web visualization and displays the results on the web page. However, with high power comes great responsibility.

Ultimately, the focus should be on our goals rather than our tools.

2.6 Additional Resources for Aspiring Data Visualists

Data visualization domain is vast and an aspiring visualizes, a person who can link between storytelling and data-experience design can further broaden his/her knowledge through:

2.6.1 Tableau Community

(Tableau Software 2018a) helps you to explore Tableau further :

- It will help us enhance our learning
- Get answers for most of your doubts In tableau
- Post new questions and crowd source answers
- Attend events, seminars and join conferences conducted locally/ globally
- Give back to the community once you become an expert in that field

There are very active Tableau social media groups (Tableau Software 2018b):

- Tableau Enthusiasts: LinkedIn Group (19K members)
- Tableau Software Fans & Friends: LinkedIn Group (45k members)

2.6.2 Blogs

Here are some blogs recommended by Tableau (Tableau Software 2018c):

Blog	Description	Link
Storytelling with Data	This blog provides information about the fundamentals of data visualization and how to make data a critcal component of your story.	http://www.storytellingwithdata.com/
Information is Beautiful	This blog was founded by David McCandless, the author of two bestselling infographics books, and provides a variety of visualizations, all of which are continuously revised and updated with the most recent data.	https://informationisbeautiful.net/

Blog	Description	Link
Visualizing Data	This blog provides a space for data visualizers to share news and thoughts about the field as well as offers diverse content about current and cutting-edge techniques, discussion of both practical and theoretical topics.	http://www.visualisingdata.com/
Junk Charts	This blog critiques a variety of graphics, providing insights about what works and what doesn't in each visualization, and how to improve them.	http://junkcharts.typepad.com/
The Pudding	This blog explores complex and contested issues through visual essays.	https://pudding.cool/
The Atlas	This blog provides visualizations on a plethora of topics.	https://www.theatlas.com/
Graphic Detail	This blog is the hub of The Economist's data journalism; it provides examples of charts, maps, and infographics (all of which are often interactive).	https://www.economist.com/blogs/graphicdetail
Tableau Blog	The Tableau blog is a source of data viz trends, issues important to the Tableau community, and updates about Tableau products.	https://www.tableau.com/about/blog
Michael Sandberg's Data Visualization Blog	Michael Sandberg's blog discusses a wide variety of data visualization examples. In addition, the blog covers other topics such as Infographics, Data Science, Business Intelligence, Data Ethics, Storytelling and much more.	https://datavizblog.com/
Beautiful Data Blog	This blog features discussion of all things data and has a wide range of categories to browse through.	http://beautifuldata.net/
insert reference	Michael Sandberg's blog discusses a wide variety of data visualization examples	https://datavizblog.com/
insert reference	This blog features discussion of all things data	http://beautifuldata.net/

2.6.3 Podcasts

Here are some podcasts recommended by Tableau (source: https://www.tableau.com/about/blog/2015/8/want-more-data-your-ears-here-are-6-podcasts-worth-listening-43300) and other worth listening resources:

Podcast	Description	Link
Policy Viz	The host of the podcast chats with guests primarily about data visualization, open data, big data, and technology to help people communicate data better. Episodes are all 20 minutes or shorter, and it's mostly the guests speaking, not the host.	https://policyviz.com/podcast/
Tableau Wanna Be Podcastl	The hosts are respected and passionate members of the Tableau community. The podcast is about all things related to Tableau, including interviews with other members of the Tableau community, Tableau employees, and in-depth discussions about visual analytics.	https://soundcloud.com/tableau-wannabe-podcast
What's The Point	It talks about the many differnt ways data affect our lives, with interviewing people from many different walks of life (for example, astronomy, publishing, web tracking). It's a short show each time, and the guests are the stars.	https://fivethirtyeight.com/tag/whats-the-point/
Data Stories	This long-running podcast is about data visualisation in general and has consistently drawn in some very high-calibre guests. Each episode is about an hour, which suits longer commutes.	http://datastori.es/
Storytelling with Data	This podcast is an additional resource to the Storytelling with Data blog, which covers topics related to data storytelling, better presentations, and all things data viz.	http://www.storytellingwithdata.com/podcast

2.6.4 Useful Links on Data Visualization Resources, Trends, and Tutorials

Resource	Description	Link
(Catalogue 2018)	You can find different types of plots used in data visualization	https://datavizcatalogue.c
(Kosara 2018a)	Robert Kosara's website which contains recent developments happening in visualization and are likely to have an impact.	
(Research 2018)	About Robert Kosara and his research papers.	
(Kosara 2018b)	Robert Kosara's twitter handle.	
(FlowingData 2018)	Website which offers courses, tutorials and happenings in viz.	http://flowingdata.com/
(Infogram 2018)	An infogram helps a user making different types of plots and learning the art of visualization. Engaging infographics, reports, charts, dashboards and maps can be easily created in minutes with it.	
insert reference	Improving data visualisation for the public sector (does this link work?)	http://www.improving-visualisation.org/
insert reference	This resource is a data visualization gallery of weekly explorations of United States Census data	https://www.census.gov/dataviz/
(Joerg Blumtritt, n.d.,)	This blog features discussion of all things data	http://beautifuldata.net/

Resource	Description	Link
(Sandberg, n.d.)	Michael Sandberg's blog discusses a wide variety of data visualization examples	https://datavizblog.com/
(NA, n.d.)	This resource is a data visualization gallery of weekly explorations of United States Census data	https://www.census.gov/dataviz/
(Agency 2018)	This resource provides a series of interactive data visualizations using FEMA data	https://www.fema.gov/data-visualization
insert reference	This blog features discussion of all things data	http://beautifuldata.net/
insert reference	Improving data visualisation for the public sector (does this link work?)	http://www.improving-visualisation.org/
insert reference	Michael Sandberg's blog discusses a wide variety of data visualization examples	https://datavizblog.com/
insert reference	This resource is a data visualization gallery of weekly explorations of United States Census data	https://www.census.gov/dataviz/
insert reference	This resource provides a series of interactive data visualizations using FEMA data	https://www.fema.gov/data-visualization
insert reference	The History of Data Visualization Dashboard Insight, Dashboard Insight, 2013	http://www.dashboardinsight.com/news/news-articles/the-history-of-data-visuali aspx
insert reference	Current research: Deceptive visualizations, Infogram, 2016	https://medium.com/@Infogram/study-asks-how-deceptive-

Resource	Description	Link
insert reference	Agata Kwapien in Data Visualization, 2015	https://www.datapine.com/blog/misleading-data-visualizat
insert reference	A Brief History of Data Visualization, York University, Michael Friendly, 2006	http://www.datavis.ca/papers/hbook.pdf
insert reference	Data Visualization and the 9 Fundamental Design Principles, Melissa Anderson, 2017	https://www.idashboards.com/blog/2017/07/26/data-visualization-and-the
insert reference	A Practitioner Guide to Best Practices in Data Visualization.Interfaces 47(6):473-488, Jeffrey D. Camm, Michael J. Fry, Jeffrey Shaffer, 2017	https://doi.org/10.1287/inte.2017.0916
insert reference	The 7 Best Data Visualization Tools In 2017	https://www.forbes.com/sites/bernardmarr/2017/07/20/the-7-best-data-visualizati#3a12b8ea6c30
insert reference	The Data Visualisation Catalogue	https://datavizcatalogue.com

Chapter 3

Fundamentals

This chapter covers foundational design principles and both general and more specific best practices, as well as explores popular visualization tools and some special topics relevant to the field of data visualization, and concludes with a discussion of what's next for the field.

3.1 Design Principles

The role of data visualization in communicating the complex insights hidden inside data is vital. This is becoming more and more important since the audience for data visualizations is also expanding along with the size of data. Data visualizations are now consumed by people from all sorts of professional backgrounds. For the same reason, the ease of consumption is now a hot topic. While data scientists and analysts have an eye for digging out the key insights from even complex visualizations, a top business stakeholder or an average person might not be able to do the same.

And this is what makes effective data visualization the need of the hour. Communicating the data effectively is an art. However, many data scientists lag behind when it comes to the design and aesthetic aspects of visualizing data.

Here are some of the key design principles for creating beautiful and effective data visualizations for everyone.

(Source: (Koshy 2018))

3.1.1 Melissa Anderson's Principles of Design

The following principles are from (Anderson 2017).

Criteria	Description
Balance	A design is said to be balanced if key visual elements such as color, shape, texture, and negative space are uniformly distributed. Balance doesn't mean that each side of the visualization needs perfect symmetry, but it is important to have the elements of the dashboard/visualization distributed evenly. And it is important to remember the non-data elements, such as a logo, title, caption, etc. that can affect the balance of the display.

Criteria	Description
Emphasis	Draw viewers' attention towards important data by using key visual elements. Emphasis is the component that is most related to when reading the nine principles of design. It is the key to be conscious of what is drawing the viewer's attention to the 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.
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 complementary colors to pull the user's attention across the page or with use of animation
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.
Repetition	Relationships between sets of data can be communicated by repeating chart types, shapes, or colors.
Proportion	If a person is portrayed next to a house, the house is going to look bigger. In data visualization, the proportion can indicate the importance of datasets, along with the actual relationship between numbers. 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. Proportion can be misused intentionally as well as unintentionally, since images are easier to interpret than data by humans. 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.
Rhythm	A design has proper rhythm when the design elements create the movement that is pleasing to the eye. If the design is not able to do so, rearranging visual elements may help.
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. 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.
Unity	Unity across design will happen naturally if all other design principles are implemented.

3.1.2 Gestalt Principles of Design

Data is simply a collection of many individual elements (i.e., observations, typically represented as rows in a data table). In data viz, our goal is usually to group these elements together in a meaningful way to highlight patterns and anomalies. Described this way, it makes sense that the following principles by Gestalt are a good set of guidelines to assemble different elements into groups (FusionCharts 2012).

Principle	Description
Proximity	White space can be used to group elements together and separate others
Similarity	Objects that look similar are instinctively grouped together in our minds
Enclosure	Helps distinguish between groups
Symmetry	Objects should not be out of balance, or missing, or wrong. If an object is asymmetrical, the viewer will waste time trying to find the problem instead of concentrating on the instruction.
Closure	We tend to complete shapes and paths even if part of them is missing
Continuity	We tend to continue shapes beyond their ending points (similar to closure)
Connection	Helps group elements together
Figure and ground	We typically notice only one of several main visual aspects of a graph; what we do notice becomes the figure, and everything else becomes the "background". This one is especially interesting because it is not as obvious as some of the others, but is really important in matching a data viz design to its purpose.

3.1.3 Tufte's Principles of Design

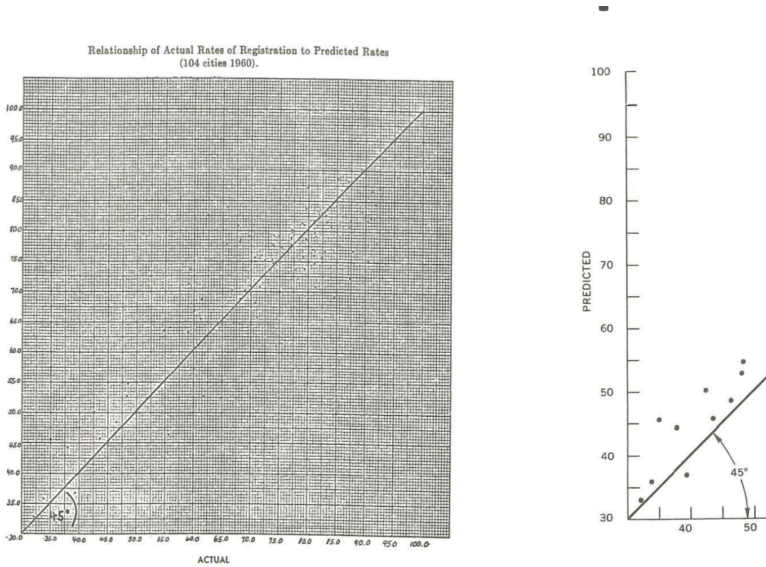
A graph should be impressive and can obtain audience's attention. How can we achieve this? We must consider several aspects: **efficiency, complexity, structure, density and beauty**. We also should consider the audience whether they will be confused about the design.

3.1.3.1 Principle 1: Maximizing the data-ink ratio, within reason

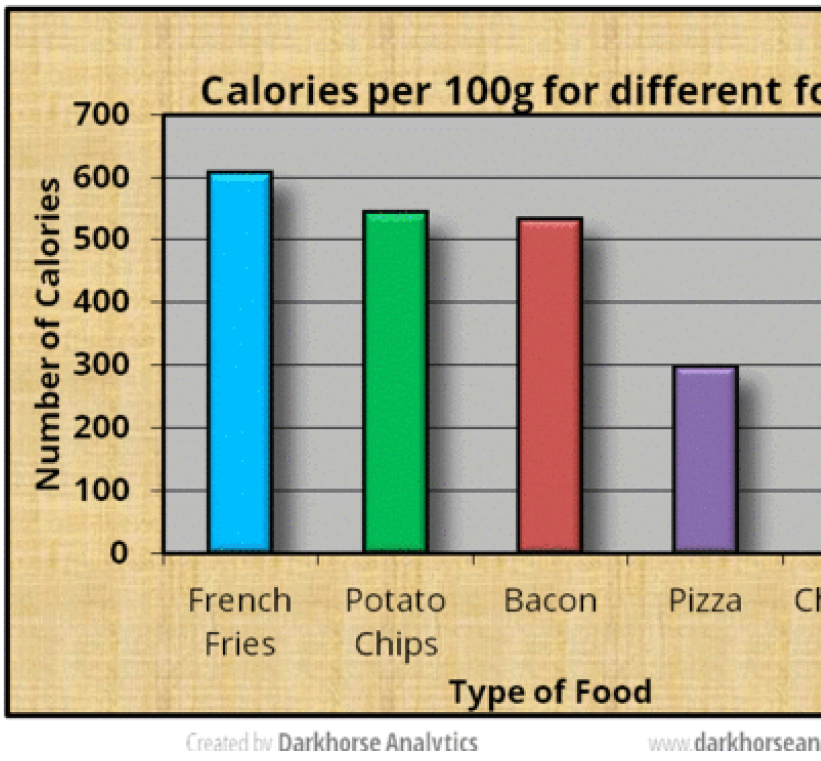
Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented. It is also the proportion of graphic's ink devoted to the non-redundant display of data-information.

$$Data\ Ink\ Ratio = \frac{Data\ Ink}{Total\ Ink}$$

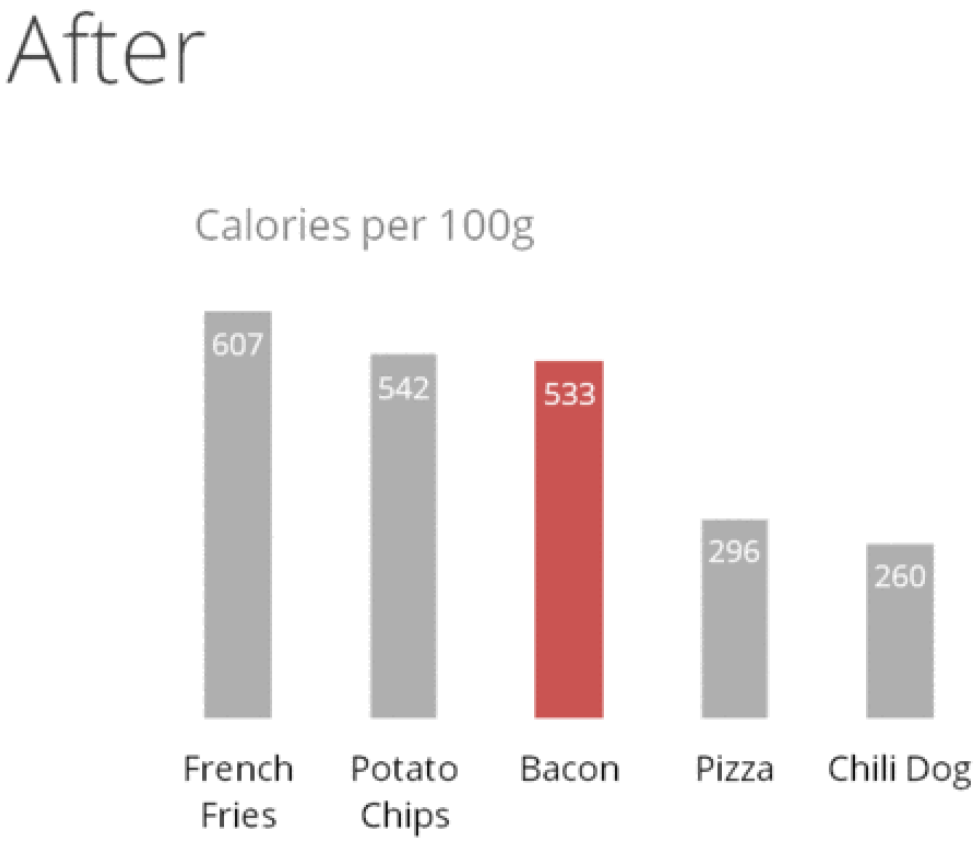
This basic idea is illustrated in the following visualizations.



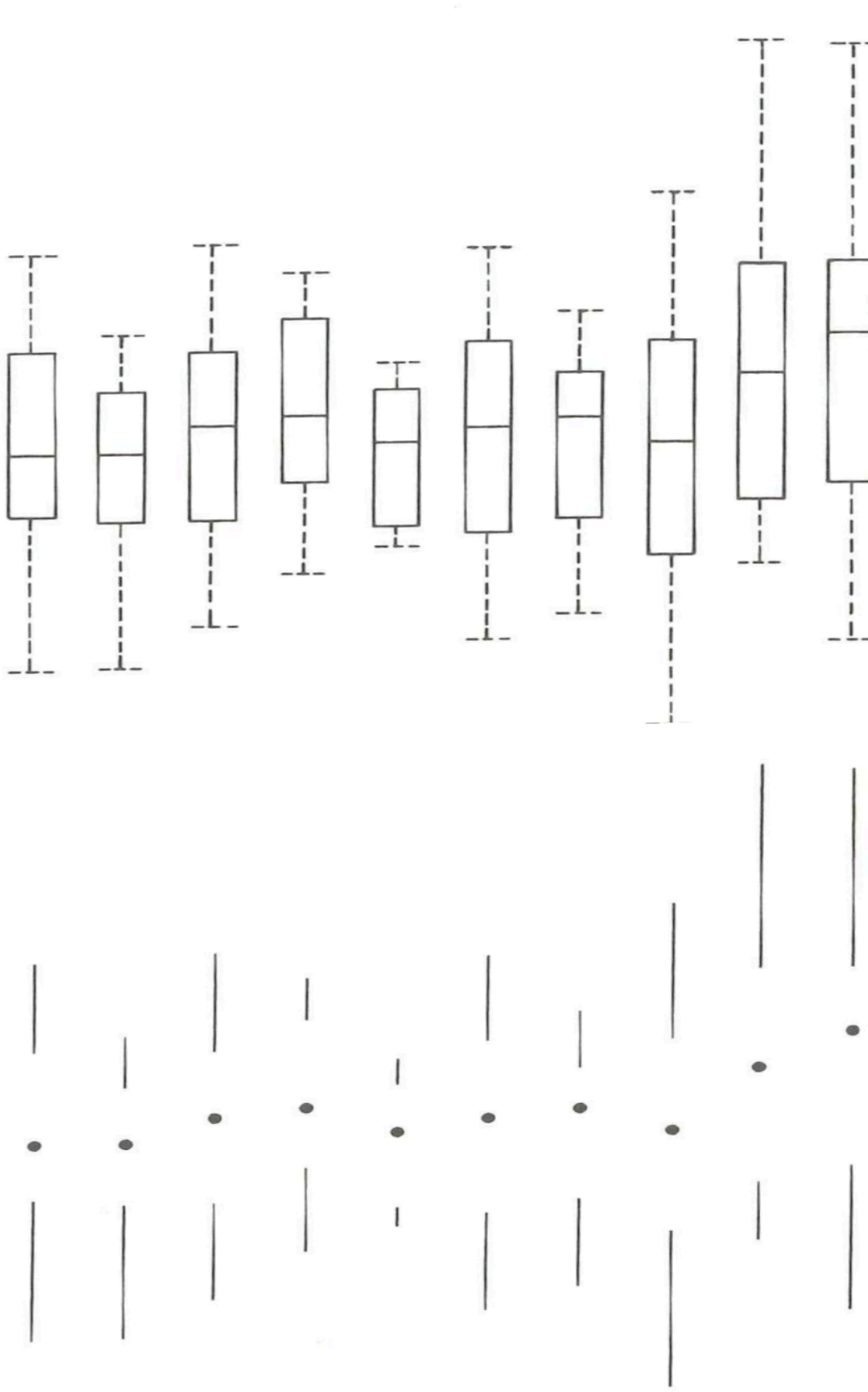
Erase non-data-ink and redundant data-ink.
(Source:(Tufte 1986))



Erase non-data-ink and redundant data-ink.
(Source: (Plotly 2017))



(Source: (Plotly 2017))



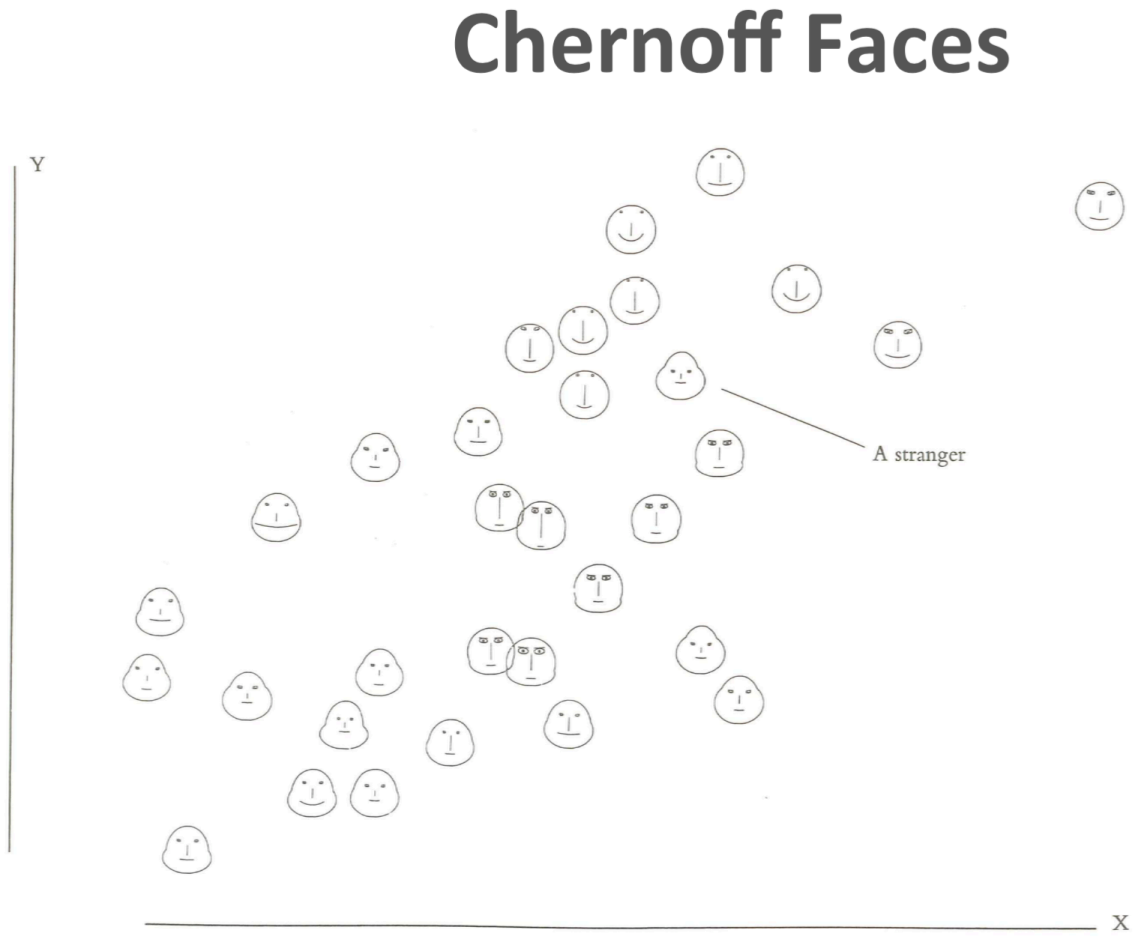
Always revise and edit

(Source: (Tufte 1986))

kwright76_submit When creating visualizaitons, information is best displayed by adhering to the data to pixel ratio rule. This rule states that a visulaization should contain as much data as possible while also using as little pixels as possible. Through a comprehensive editing and testing process, any visualization can continually be improved upon. The main stakeholder of any visualization is the audience and their ability to understand what the visualization is trying to get across. Although if an audience member is not able to understand the visualization, there is nothing lost, but for those that do understand it, something is gained. It is a great feat for an audience member to be able to understand a statistical graphic because it is the most frequently made mistake in underestimating an audience. New designs, although may appear odd, have not been seen before and can be successful visualizations. Therefore, visualizations that produce a lot of data with the space provided with detailed statistics and are able to be understood by a wide array of audiences are produced with a well put together revising and testing system. contributions

3.1.3.2 Principle 2: Mobilize every graphical element, perhaps several times over, to show the data.

The danger of multifunctioning elements is that they tend to generate graphical puzzles, with encodings that can only be broken by their inventor. Thus design techniques for enhancing graphical clarity in the face of complexity must be developed along with multifunctioning elements. In other words, we should try to make all present graphical elements data encoding elements. We must make every graphical element effective (See the following example).

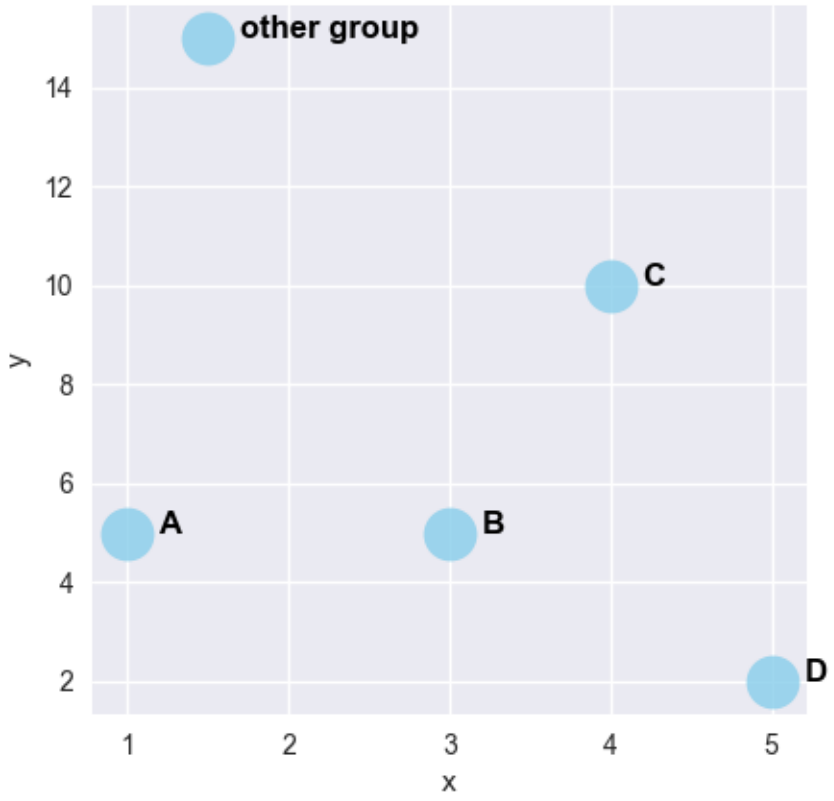


(Source: (Tufte 1986))

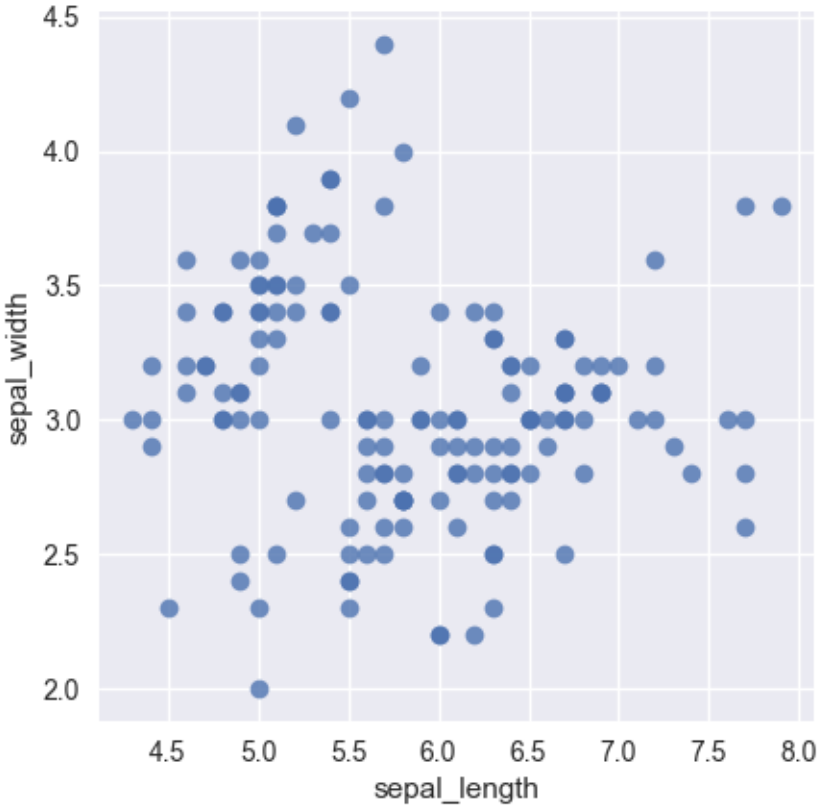
3.1.3.3 Principle 3: Maximize data density and the size of the data matrix, within reason.

High preformation graphics should be designed with special care. As the volume of data increases, data measures must shrink (smaller dots for scatters, thinner lines for busy time-series).

$$Data\ Density = \frac{Entries\ in\ the\ Data\ Matrix}{Area\ of\ Chart}$$



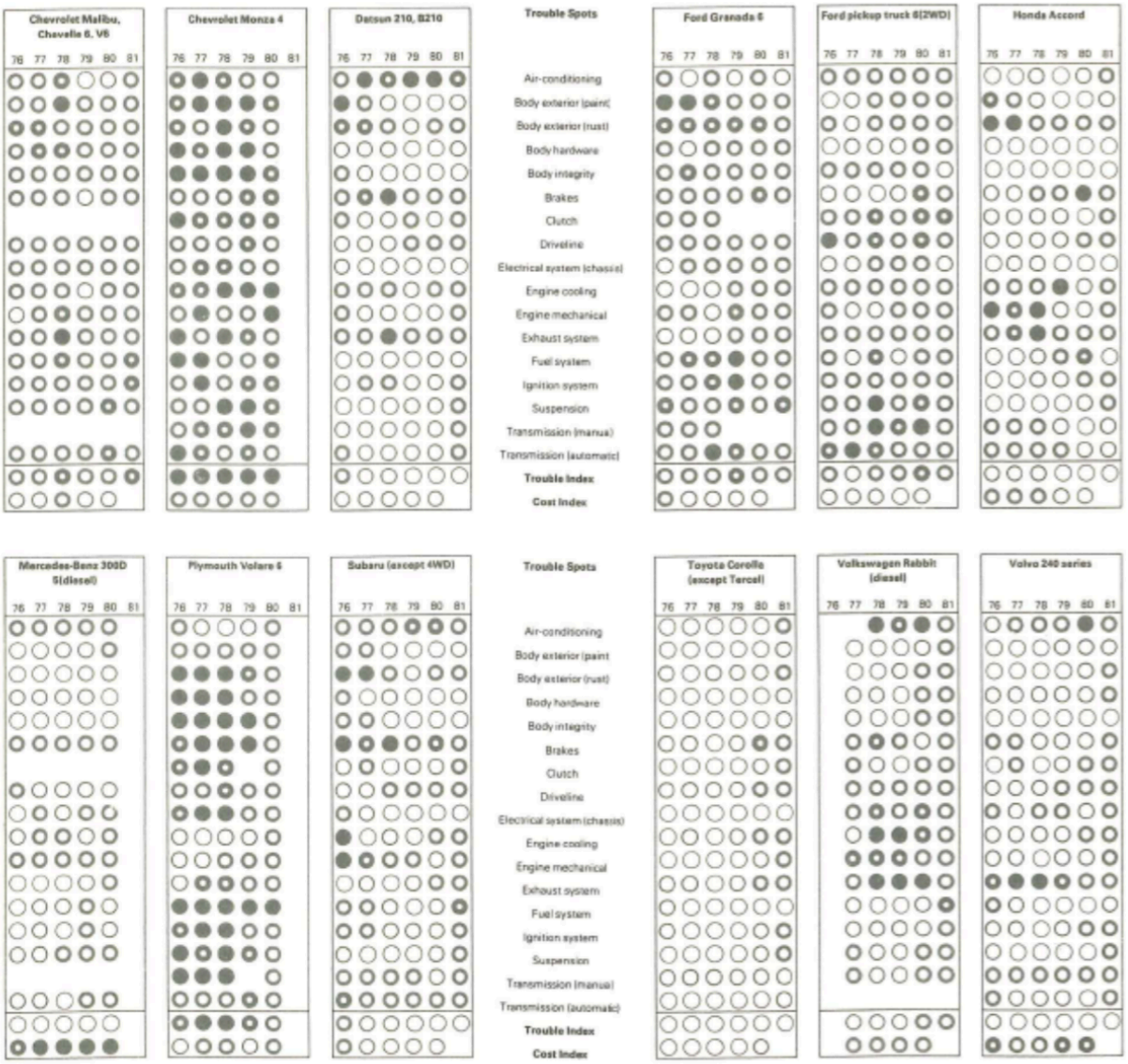
(Source: (gallery, n.d.))



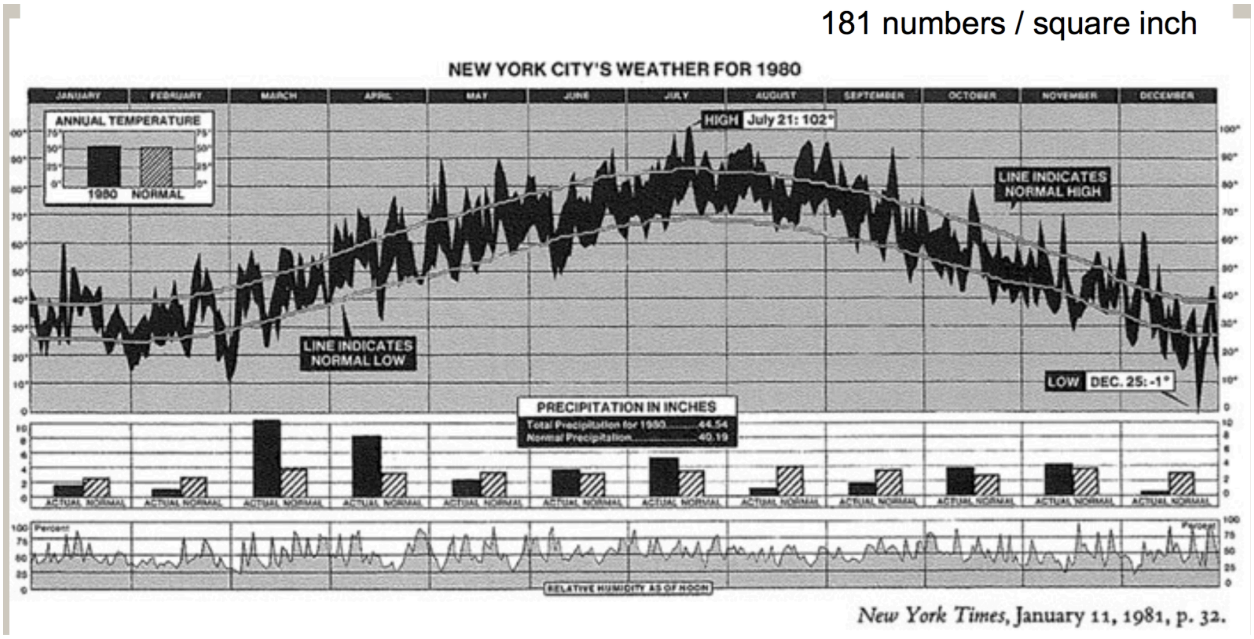
(Source: (gallery, n.d.))

3.1.3.4 Principle 4: Escape flatland - small multiples, parallel sequencing.

Data is multivariate doesn't necessarily mean 3D projection. How can we enhance multivariate data on inherently 2D surfaces? We can use small multiple graphs or parallel sequencing skill.



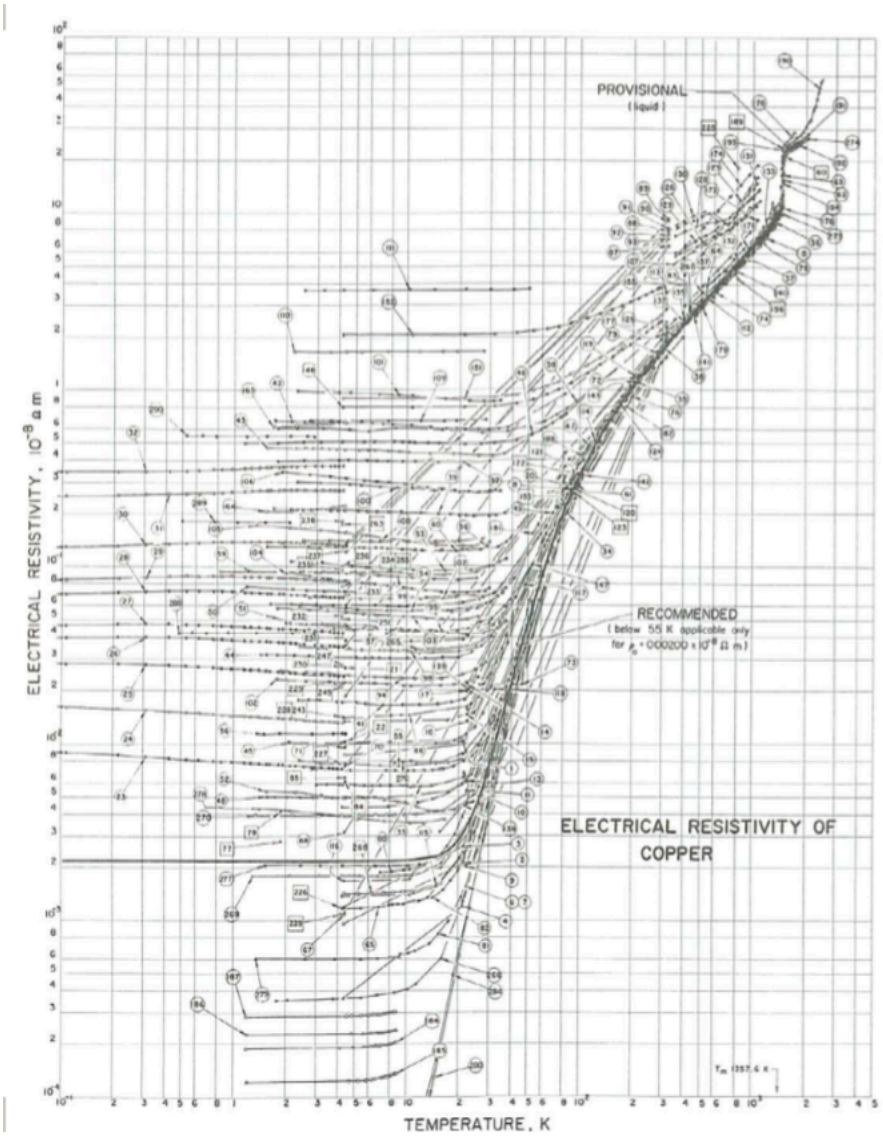
(Source: (Tufte 1986))



(Source: (Tufte 1986))

3.1.3.5 Principle 5: Provide the user with an overview and details on demand.

A carefully designed view can show a macrostructure (overview) as well as microstructure (detail) in one space.



(Source: (Tufte 1986))

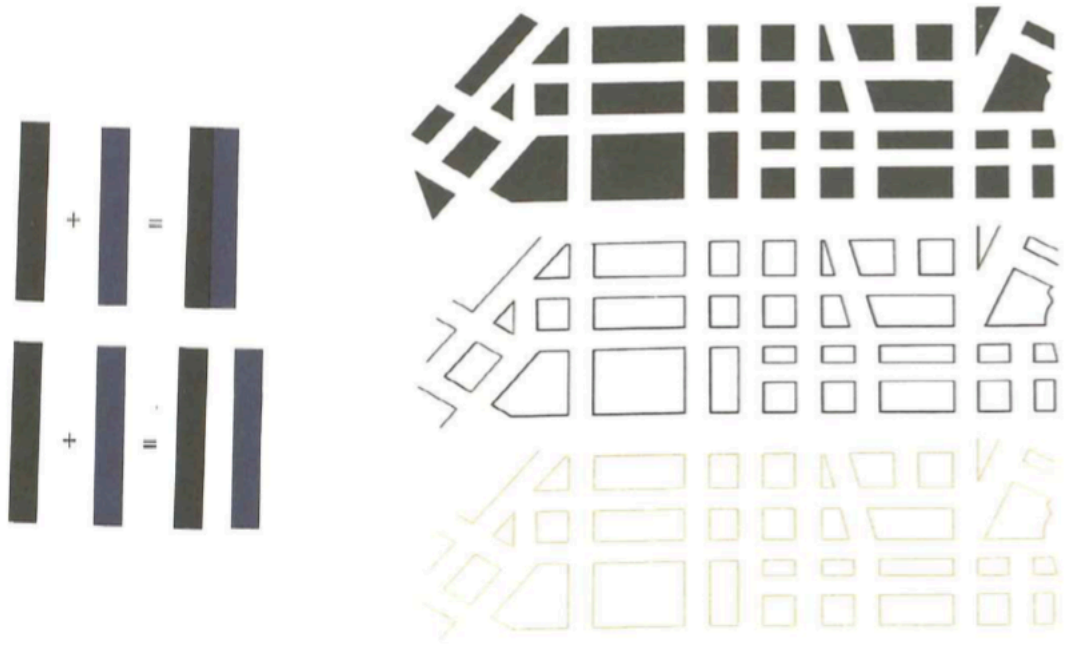
3.1.3.6 Principle 6: Utilize Layering & Separation.

1. Supported by Gestalt laws (The principles of grouping):

- Grouping with colors
- Using Color to separate
- 1 + 1 = 3 (clutter)

2. Graphics are almost always going to improve as they go through editing, revision, and testing against different design options.

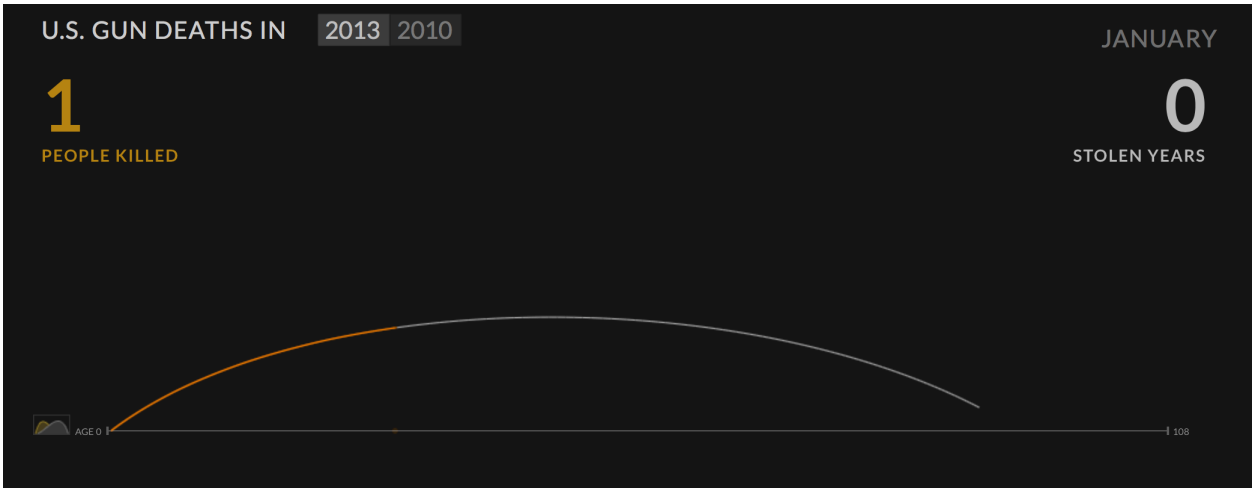
3. Try to figure out whether the audience looking at the new designs be confused? Nothing is lost to those puzzled by the frame of dashes, and something is gained by those who do understand. We should always be aware of the audience for whom we are making the charts. Furthermore, it is always not safe to assume that if you understand the statistical graphics, your readers will too.



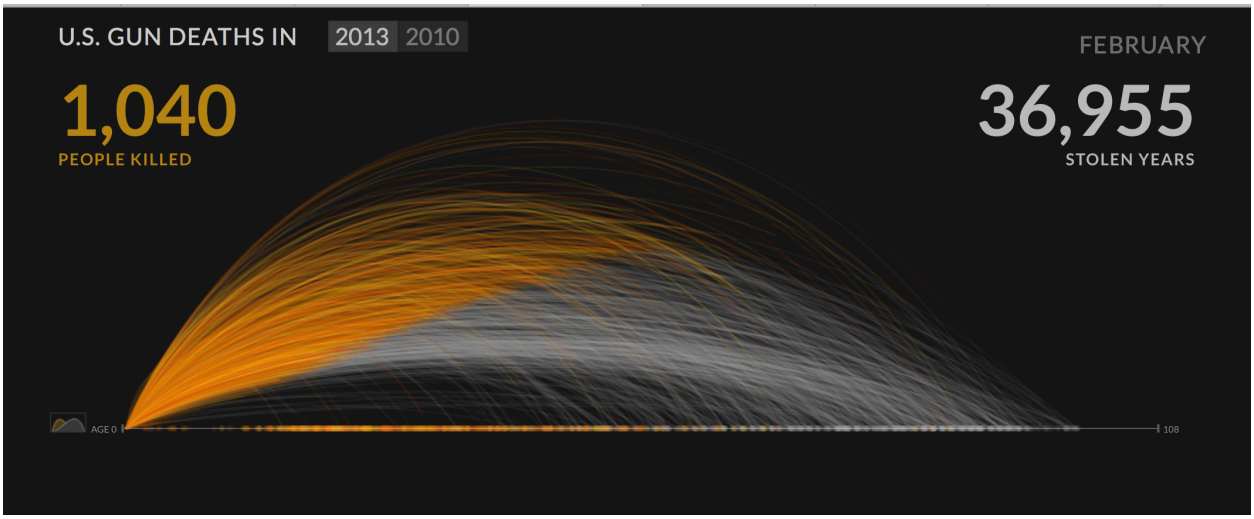
(Source: (Tufte 1986))

3.1.3.7 Principle 7: Utilize narratives of space and time.

Tell a story of position and chronology through visual elements.



(Source: (Periscope 2018))



(Source: (Periscope 2018))

3.1.4 Common Mistakes to Avoid:

Mistake	Description
Starting off with too much complexity	It's tempting to provide highly detailed, real-time dashboards. Instead of spending a lot of time working through the first iteration, however, it's better to work through several short cycles of prototyping, testing and adjusting.
Using metrics no one understands	Dashboards should use metrics or concepts that a broader audience understands. Esoteric jargon and metrics will not help get a message across.
Cluttering the dashboard with unimportant graphics and unintelligible widgets	Dashboards should be simple in visual appeal, rather than flashy or over-designed; rapidly and easily informing the audience of the primary message of the dashboard should be the priority, and clutter will only detract from that.
Waiting for complex technology and big BI deployment projects	Implementations of some of traditional business intelligence tools often take a much longer time than originally anticipated. Waiting for a traditional BI project to materialize may mean delays. A dashboarding solution takes a long time to implement and is a repetitive, iterative process with incremental improvements.
Underestimating the time or resources to create and maintain the dashboard	Even though a dashboard is typically one page or one screen, it would be injudicious to assume that it will be quick and simple to create and maintain.
Failing to match metrics to the goal	Instead of showcasing the activities of a single department, a dashboard should connect the department's efforts to the organization's actual goals and objectives
Using ineffective, poorly designed graphs and charts	While designing graphs and charts for dashboard, extreme care should be taken. Principles for designing good data visualizations should be followed to avoid dashboards populated with poorly designed graphs and charts.

3.2 Best Practices

Data visualization does not unleash a ready-made story on its own. There are no rules or protocols to guarantee a story. Instead, we need to look for *insights*, which can be artfully woven into stories in the hands of a good journalist (Jonathan Gray and Chambers 2012).

Here is a process for finding insights that tell a story. Each of these steps will be discussed further in this section.

3.2.1 Telling a Story with Insights.

Storytelling is an essential component of data visualization. The visualization must communicate complex ideas with precision and efficiency. The presenter must understand their audience's level of understanding and tailor their visualizations accordingly. An audience's level of analysis is key to creating and presenting a compelling story. Stikeleather's article outlined five key points to consider for telling a compelling story through a visualization(Jim Stikeleather 2013).

1. Find the Compelling Narrative
2. Understand your Audience
3. Be Objective and Offer Balance
4. Don't Censor
5. Edit

3.2.2 How to choose the best form of Visualization

Since just loading data into a table format could be a form of visualization, our focus should not be whether visualization is needed but on which form of data visualization is best for the situation.

Focus	Description
5 Second Rule	Research shows that the average modern attention span for viewing anything online is less than 5 seconds, so if you can't grab attention within 5 minutes, you've likely lost your viewer. Include clear titles and instructions, and tell people succinctly what the visualization shows and how to interact with it.
Design and layout matter	The design and layout should facilitate ease of understanding to convey your message to the viewer. Artists use design principles as the foundation of any visual work. If you want to take your data visualization from an everyday dashboard to a compelling data story, incorporate graphic designer Melissa Anderson's principles of design: balance, emphasis, movement, pattern, repetition, proportion, rhythm, variety, and unity, discussed in more detail in the design principles section (Anderson 2017).
Keep it simple	Keep charts simple and easy to interpret. Instead of overloading viewers' brains with lots of information, keep only necessary elements in the chart and help the audience understand quickly what is going on.
Pretty doesn't mean effective	There is a misconception that aesthetically pleasing visualization is more effective. To draw attention, sometimes we want them to be pretty and eye-catching. But if it fails to communicate the data properly, you'll lose your audience's interest as quickly as you gained it.
Use color purposely and effectively	Use of color may be prettier and attractive but can be distracting too. Thus, the color should be used only if it assists in conveying your message. Also another thing to keep in mind is to be consistent with the color scheme that the organization/consumer is used to and also try and follow the same color across dashboards while communicating a story.

3.2.2.1 Choosing suitable Visualization for various types of Data Analysis

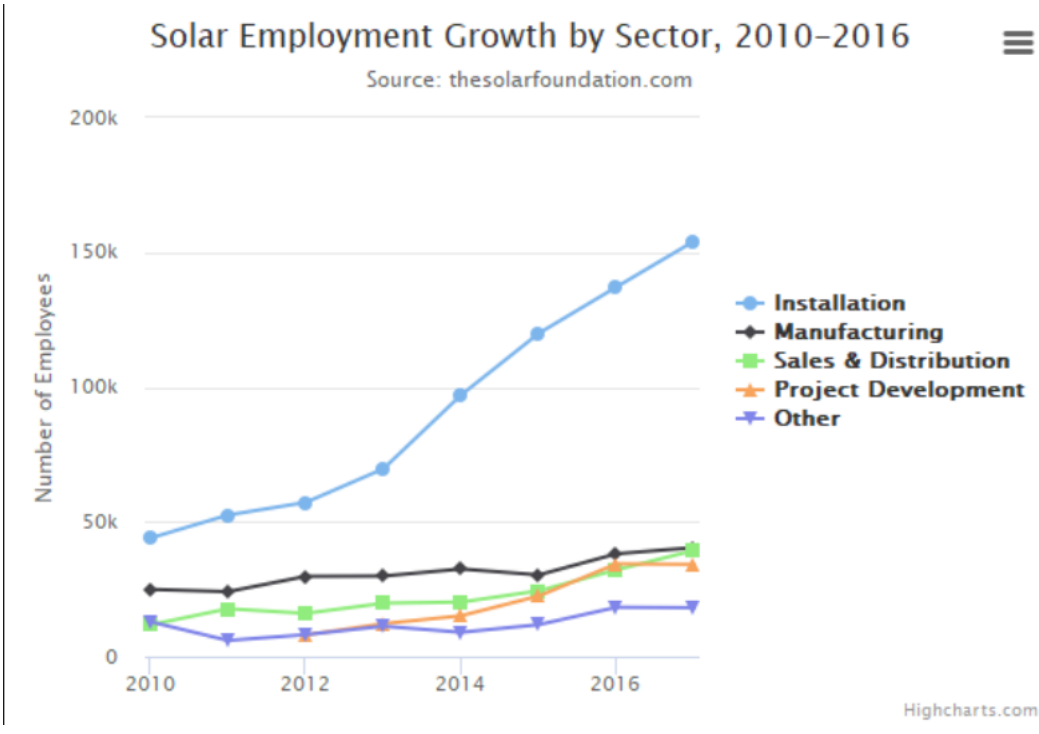
Choosing the right chart for data analysis helps in achieving the Visualization purpose. Here are some of the most commonly used analysis types and which chart types are most suitable for them.

3.2.2.1.1 Trend Analysis

It is an analysis of the rate of growth or decline (trend) between different periods of time. We can choose to compare any hour, day, month, or year with any other hour, day, month, or year. Also, we can visualize a trend in both real values and percentage changes. It allows to "see this" before "analyze this" and to take advantage of human eye ability to recognize trends quicker than any other methods.

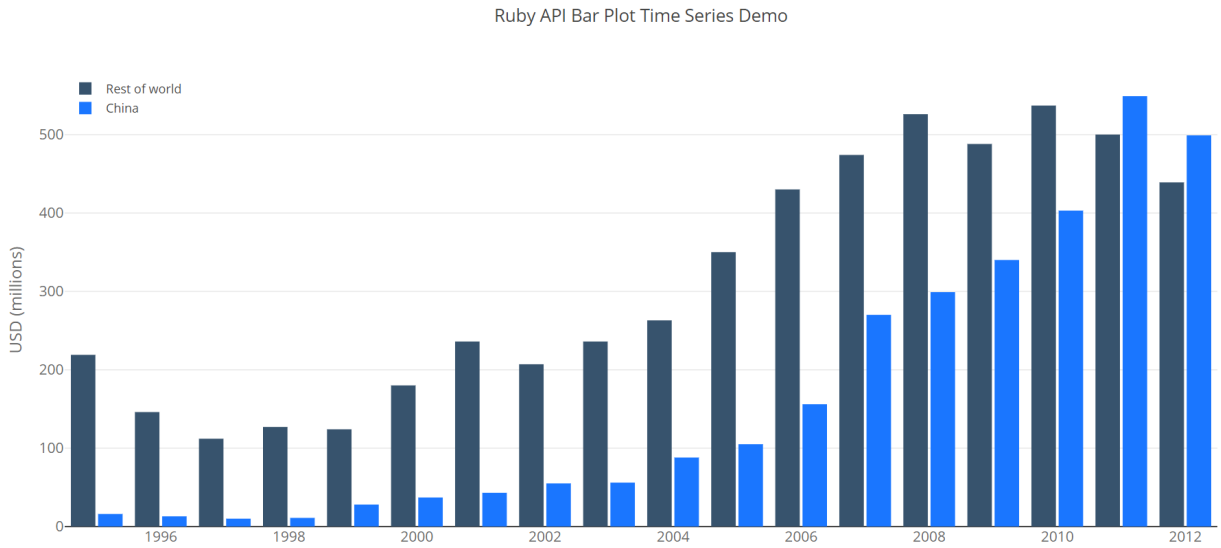
Type of Charts for Trend Analysis - Time Series Chart (using line or bar) - Motion Chart - Sparklines - Scatter

Examples of Trend Analysis - Time Series Chart (using line): A time series chart, also called a times series graph or time series plot, is tool that illustrates data points at successive intervals of time. Each point on the chart corresponds to both a time and a quantity that is being measured.



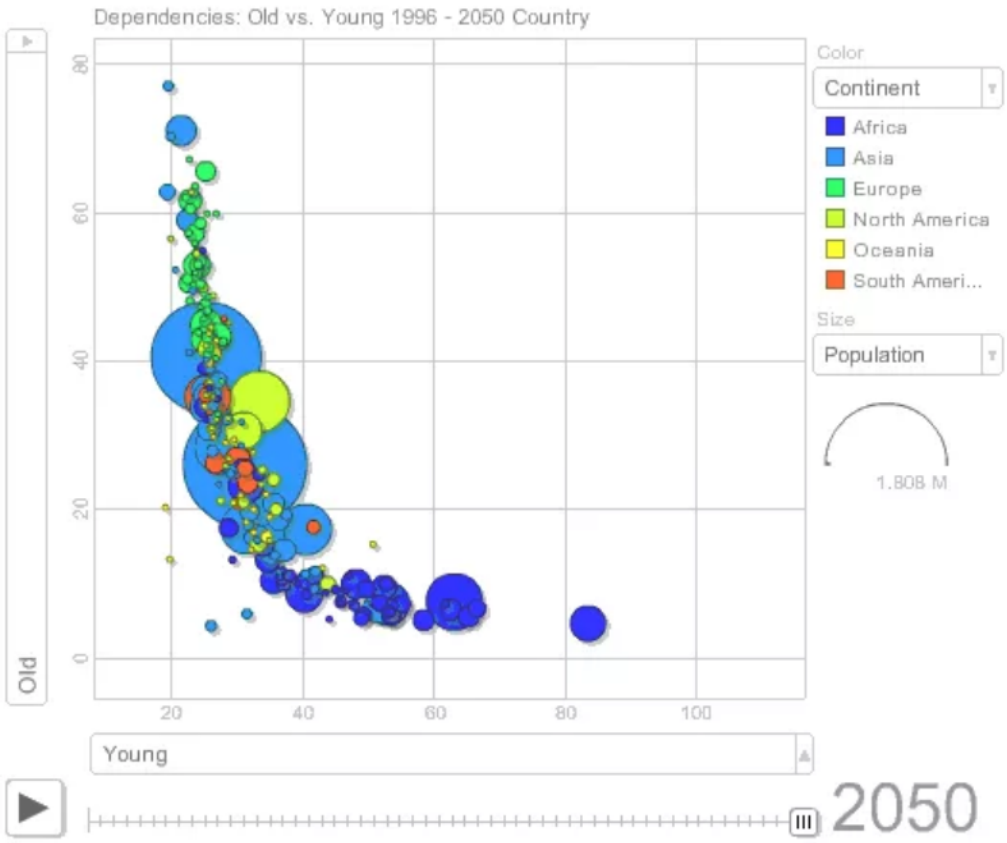
1.bb (Source: ("Time Series Chart Using Line," n.d.))

- **Time Series Chart (using Bar)**



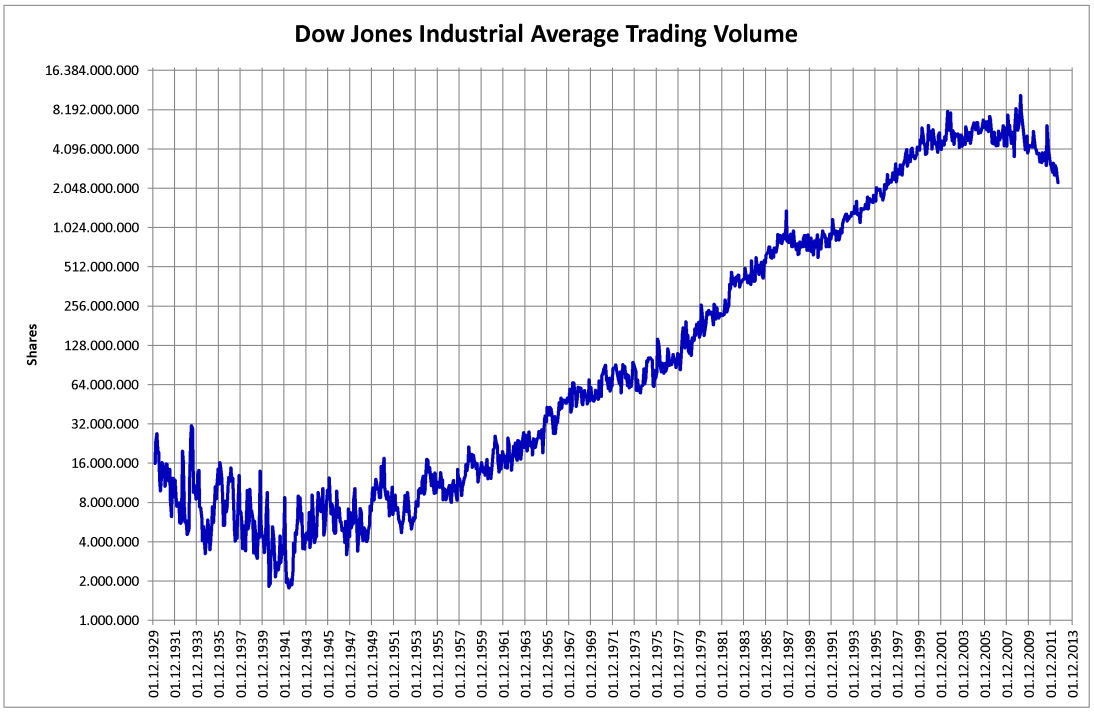
(Source: ("Time Series Chart Using Bar," n.d.))

- **Motion Chart**: A motion chart is a dynamic and interactive visualization tool for displaying complex quantitative data. Motion charts show multidimensional data on a single two dimensional plane. Dynamics refers to the animation of multiple data series over time. Interactive refers to the user-controlled features and actions when working with the visualization. Innovations in statistical and graphics computing made it possible for motion charts to become available to individuals. Motion charts gained popularity due to their use by visualization professionals, statisticians, web graphics developers, and media in presenting time-related statistical information in an interesting way. Motion charts help us to tell stories with statistics.



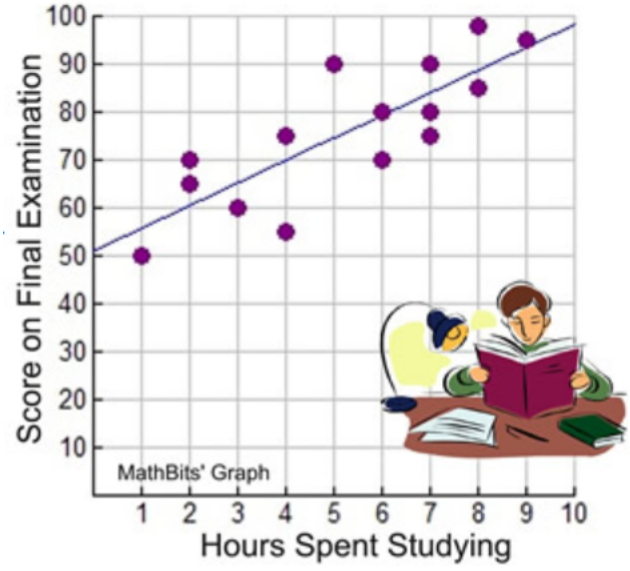
(Source: ("Motion Chart," n.d.))

- **Sparklines**: A sparkline is a small embedded line graph that illustrates a single trend. Sparklines are often used in reports, presentations, dashboards and scoreboards. They do not include axes or labels; context comes from the related content.



(Source: ("Sparkline Chart," n.d.))

- **Scatter**: A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. Scatter plots are very useful tools for conveying the relationship between two variables, but you need to know how to use them and interpret them properly

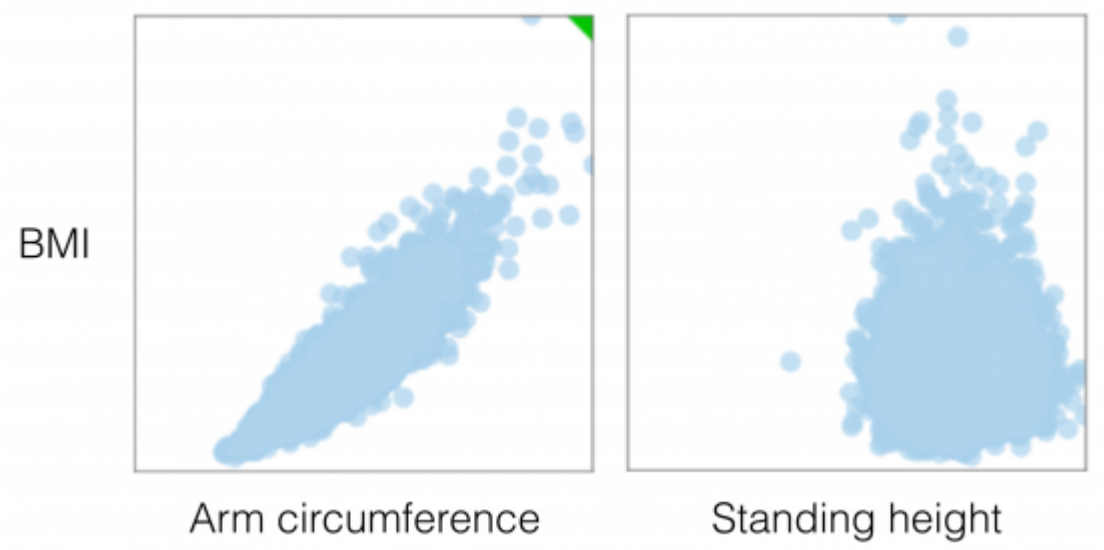


(Source: ("Scatter Chart," n.d.))

3.2.2.1.2 Correlation

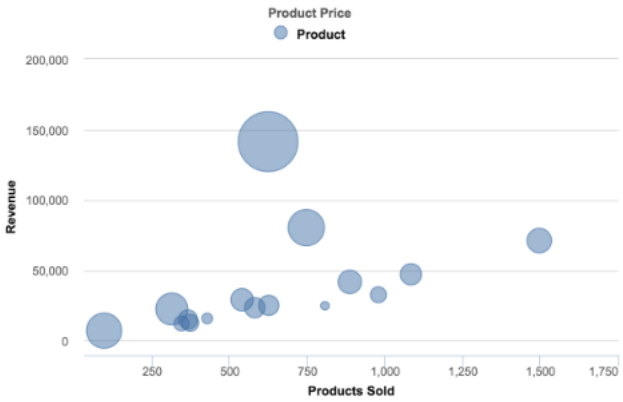
Discovering relationships between measures—it's something we do all the time in data analysis. Does smoking cause cancer? Does the price of a product impact the amount that gets sold? Running a simple correlation analysis is a crucial step in identifying relationships between measures. To confirm if the potential relationship truly exists, sophisticated methodologies are required to visually represent correlations between pairs of variables in a consistent way.

Scatter Plot : When two variables are numeric, a scatter plot is typically sufficient in representing their qualitative level of dependency. The two plots below, for example, were generated from data in the 2009-2010 National Health and Nutrition Survey, and compare two variables, arm circumference and standing height, with BMI (Body Mass Index) for adult individuals (18-65 years of age). In the first case there is a clear association between arm circumference and BMI, while in the second comparison, standing height is most likely independent from BMI. Visually, a functional relationship between two variables can be identified quite easily, even when there is a large amount of noise in the scatter plot.



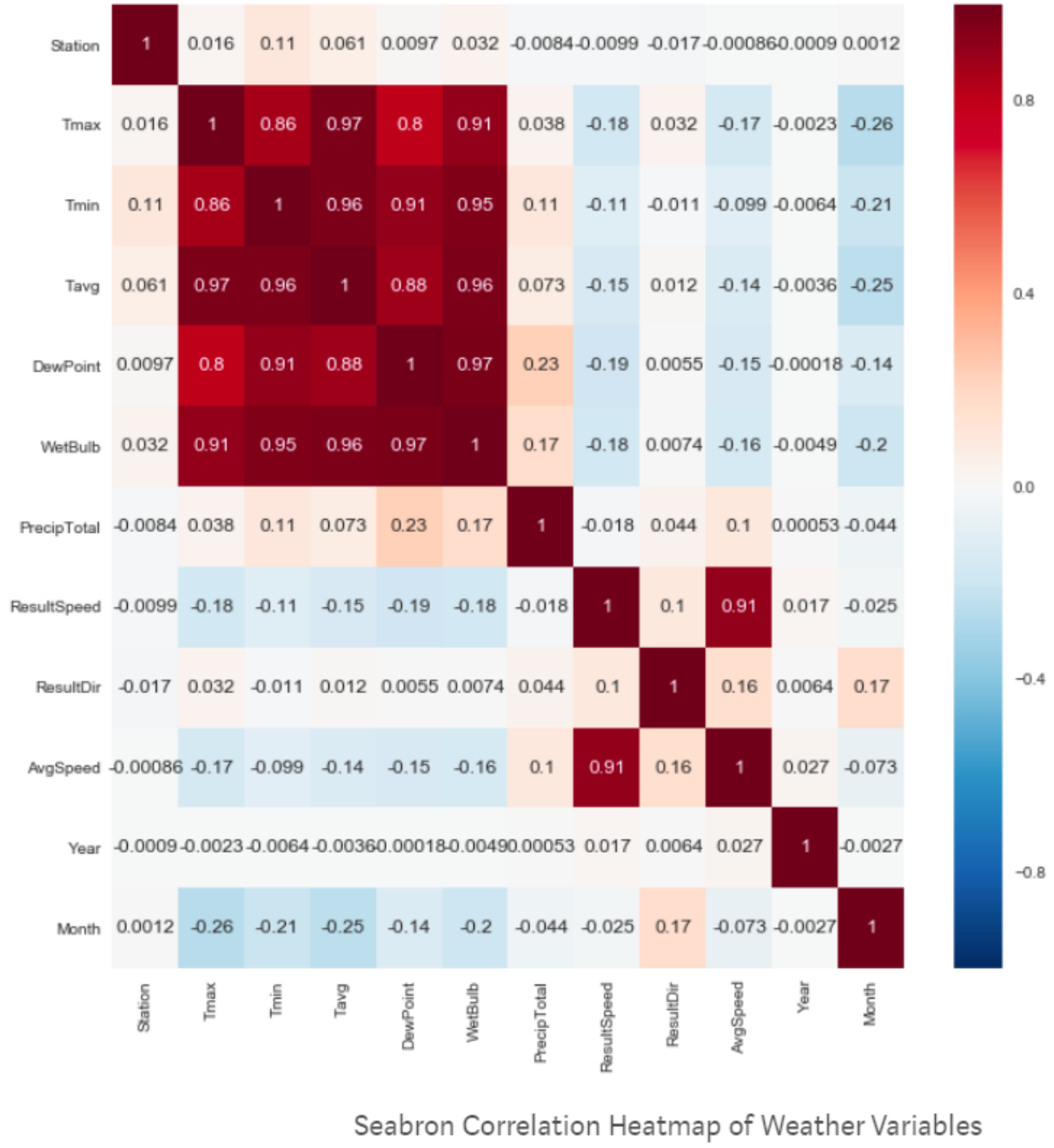
(Source: (Andres 2014))

Bubble Chart : It's a great option if there is need to add another dimension to a scatter plot chart. Scatter plots compare two values, but you can add bubble size as the third variable and thus enable comparison. A good example of a bubble chart would be a graph showing marketing expenditures vs. revenue vs. profit. A standard scatter plot might show a positive correlation for marketing costs and revenue (obviously), when a bubble chart could reveal that an increase in marketing costs is chewing on profits.



(Source: (Janis Gulbis 2016))

Heatmap : Heatmap depicts a pair-wise correlation matrix leveraged from different data attributes. This not only provides us with a numerical value of the correlation between each variable, but also provides us with an easy to understand visual representation of those numbers with high correlation to none or negative correlation. It is one of the simplest plots to create but is also one of the most informative and can guide our hand in generating other plots to investigate the numbers it has brought forward.

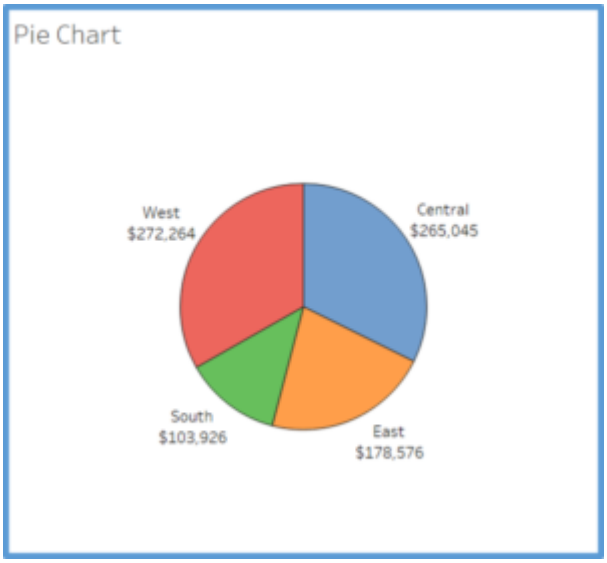


(Source: (Plapinger 2017))

3.2.2.1.3 Part to Whole

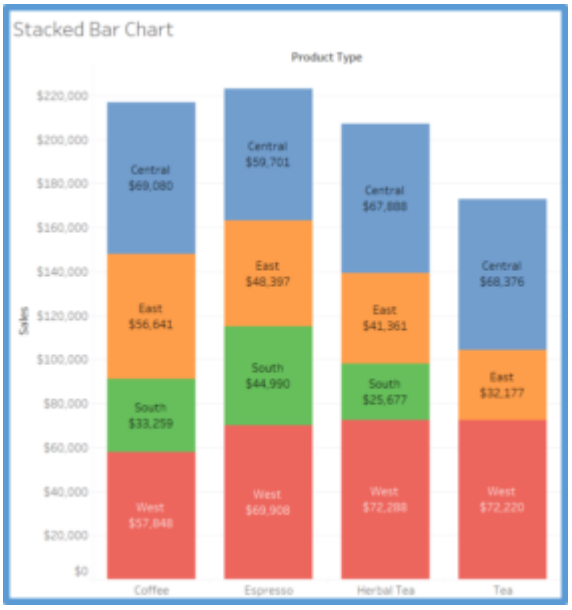
There are occasions when analysis requires visualisation methods that show part (or parts) of a variable to it's total. Often used to show how something is divided up.

Pie Chart : They are best suited to show proportional or percentage relationships. When used appropriately, pie charts can quickly show relative value to the other data points in the measure. They can be used if you really must but be aware that they are not always very accurate in depicting data. If there are more than six proportions to communicate, bar chart can be considered. It becomes too difficult to meaningfully interpret the pie pieces when the number of wedges gets too high. For example, if you didn't have the actual data points in the pie chart below we wouldn't be able to tell which region had more sales West or Central; as the slices of the pie are so similar in size.



(Source: (Strachnyi 2018))

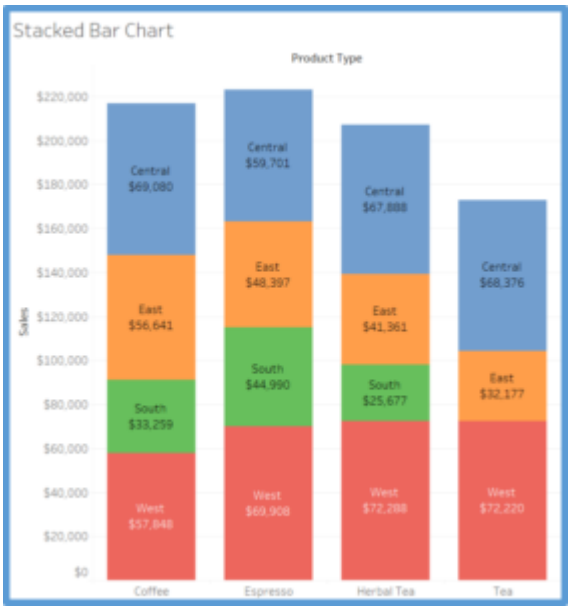
Stacked Bar Chart : A better option for visualizing the parts-to-whole relations of a data set is the bar chart. This is because it lets us compare the different objects by their length, which is one dimensional. Comparing objects along one dimension is a lot easier than along two, which makes comparing the length of bars a lot easier than the areas of pie slices. Stacked bar chart shows data in categories that are also stratified into sub-categories. In the example below we have sum of sales by product type and further divided into region. It allows us to see more details than the regular bar chart would provide.



(Source: (Strachnyi 2018))

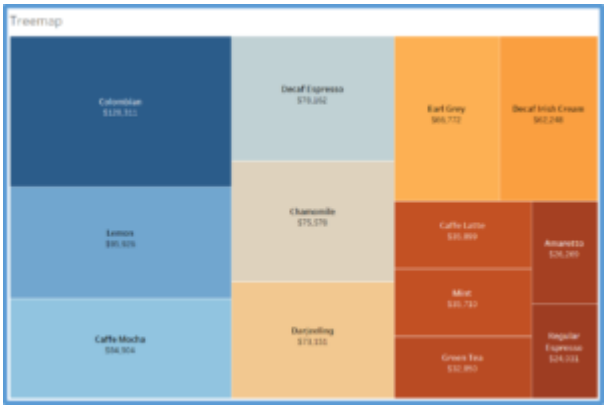
Side by Side Bar Chart : Similar to bar charts, this chart shows a side by side comparison of data. In the below example we are looking at regions and types of product (decaf vs. regular). The use of color makes it

easier to compare the sum of sales within each region for different product types. The side-by-side bar chart is similar to the stacked bar chart except the bars are un-stacked and put the bars side by side along the horizontal axis.



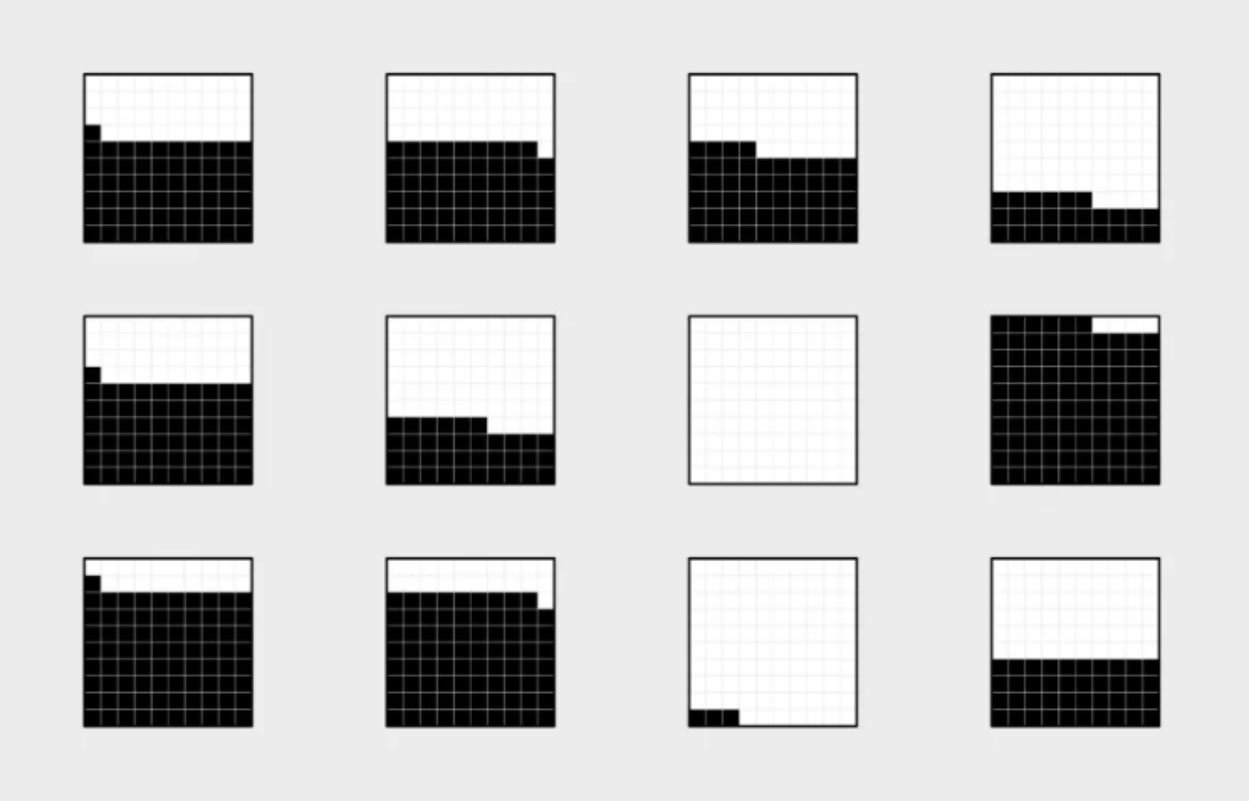
(Source: (Strachnyi 2018))

Treemap : Treemap is used to show hierarchical (tree-structured) data and part-to-whole relationships. Treemapping is ideal for showing large amounts of items in a single visualization simultaneously. This view is very similar to a heat map, but the boxes are grouped by items that are close in hierarchy.



(Source: (Strachnyi 2018))

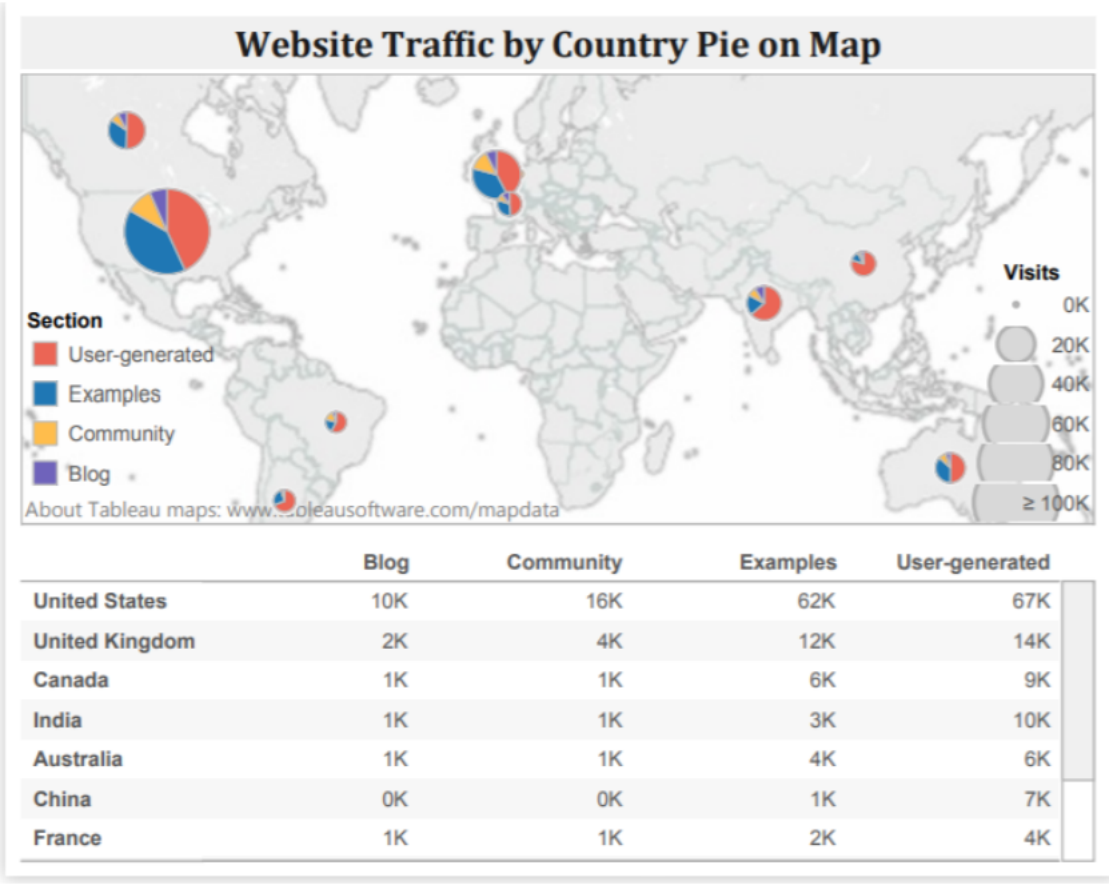
Square Pie Charts : Regular pie charts have their issues with the perception of angle and arc length, but their main advantage is that they represent parts of a whole. The metaphor is universally known. The square pie chart offers an alternative without sacrificing the metaphor, and are easier to read and more accurate at showing data. Designers like this one a lot when they want to focus on a single data point. It takes up a lot of space, but sometimes puts things in better perspective. Basically instead of showing each data point, you're showing every individual count within a data point.



(Source: (Yau 2018))

3.2.2.1.4 Geographical Data

Maps are ideal to show data on location. Maps are often best when paired with another chart that details what the map displays — such as a bar chart sorted from greatest to least, a line charts showing the trends, or even just a cross-tab to show the actual data. Although pie charts are not recommended for part-to-whole relationship, pie charts can be used on maps, such as in the website-traffic map below. By using pies on the map, we get a rough breakdown view of each country, which can be very useful when complemented by other chart types like the ones mentioned previously.



	Blog	Community	Examples	User-generated
United States	10K	16K	62K	67K
United Kingdom	2K	4K	12K	14K
Canada	1K	1K	6K	9K
India	1K	1K	3K	10K
Australia	1K	1K	4K	6K
China	0K	0K	1K	7K
France	1K	1K	2K	4K

(Source: (n.d.))

3.2.3 Analyze and Interpret

Once data is visualized, the next step is to learn something from visualization that was created. The single most important step to make a great visualization is to know what you're trying to say. It is vital that a visualization has a purpose and that one is selective about what to include in the visualization to fulfill that purpose. A few general questions that can be asked to determine the purpose of a visualization include: * Who is the audience of the visualizations? * What questions do they have? * What answers does the visualization provide them? * What other questions does it inspire? * What conversations will result? and so on…

The point is that the viewers should take something away from the time they spend with the visualization. (Software 2019)

3.2.3.1 Macro/Micro Views - Provide the user with both views (overview and detail) - (Tufte's Design Principle 2)

A carefully designed view can show a macro structure (overview) as well as micro structure (detail) in one space.

- What can be seen in this image? Is it what was expected?
- Are there any interesting patterns?
- What does this mean in context of the data?

Sometimes we might end up with visualization that, in spite of its beauty, might seem to tell that nothing of interest can be found from the data. But there is almost always something that we can learn from any visualization, however trivial.

3.2.3.2 Checking if the visualization answers the intended questions

Often it helps to ask what questions is the visualization trying to answer? Following which, it is easier to evaluate a visualization by asking some of the following questions: (Software 2019)

Overall message/ Warrant/ Claim * What is the overall message or claim or warrant that the visualization is trying to make? * Does the visualization jusity the claim?

Title * Does the visualization include a title? Is that title simple, informative, and eye-catching? * Is the purpose of the visualization clearly explained in its title or surrounding text?

Interpretation * Can one understand the visualization in 30 seconds or less, without additional information? * Does the visualization include subtitles to guide your viewers?

3.2.3.3 Using the right chart types

Another important aspect to consider is how effective are chart views in terms of measures & dimensions, colors, etc. It helps to ask the following questions:

- What types of analysis is being performed?
- Is the chart type(s) most suitable for the type of analysis?
- Are there alternative chart types that could work better than the chosen ones? More information on Chart Types can be referred here (Software 2019) and using Tableau online documentation.

3.2.3.4 Ensuring the dashboard has a holistic design

- Do all views fit together to tell a single story?
- Do all views flow well from one to the next? Are they in a good order?
- Does the most important view appear in the top or top-left corner?
- Are secondary elements in the dashboard placed well so they support the views without interrupting them?
- Are filters in the right locations?
- Do filters work correctly? Do views become blank or downright confusing if a filter is applied?
- Do filters apply to the right scope?
- Are filter titles informative? Can viewers easily understand how to be interactive with the filters?
- Are legends close to the views they apply to?
- Are there any filter, highlight or URL actions? If so, do they work?
- Are legends and filters grouped and placed intuitively?
- Are there scrollbars in the views? If so, are they acceptable?
- Are the views scrunched?
- Do the views fit consistently when filters are applied?

3.2.3.5 Evaluating effectiveness through measures, dimensions, colors etc.

- Are the most important data shown on the X and Y-axes and less important data encoded in color or shape attributes?
- Are views oriented intuitively? Do they cater to the way the intended audience would read and perceive data?
- Are number of measures or dimensions limited in a single view so that users can see the data?

- Is the usage of colors and shapes limited so that users can distinguish them and see patterns?

3.2.3.6 Final touches

- Do all colors on the dashboard go together without clashing?
- Are there less than 7-10 colors on all the dashboards?
- Are fonts used consistently in all of the dashboards/ views and there are no more than three different fonts on one dashboard?
- Are the labels clear and concise? Are they placed optimally to help guide the viewers? Make sure subtitles are formatted to be subordinate to the main title.
- Are tooltips informative? Do they have the right format so that they're easy for the audience to use?

3.2.4 Document Your Insights and Steps

If you think of this process as a journey through the dataset, the documentation is your travel diary. It will tell you where you have traveled to, what you have seen there and how you made your decisions for your next steps. You can even start your documentation before taking your first look at the data.

In most cases when we start to work with a previously unseen dataset, we are already full of expectations and assumptions about the data. Usually, there is a reason why we are interested in that dataset that we are looking at. It's a good idea to start the documentation by writing down these initial thoughts. This helps us to identify our bias and reduces the risk of misinterpretation of the data by just finding what we originally wanted to find.

Personally speaking, the documentation is the most important step of the process, and it is also the one people most likely to skip. As you will see in the example below, the described process involves a lot of plotting and data wrangling. Looking at a set of 15 charts you created might be very confusing, especially after some time has passed. In fact, those charts are only valuable (to you or any other person you want to communicate your findings) if presented in the context in which they have been created.

3.2.5 Transform Data

Naturally, with the insights that you have gathered from the last visualization, you might have an idea of what you want to see next. You might have found some interesting pattern in the dataset which you now want to inspect in more detail. Possible transformations are the following.

Focusing the attention: What can be removed? Realize that consistency can help eliminate unnecessary distractions. There may be a trade-off between losing information but conveying the ultimate meaning more clearly. Label important things rather than relying on a legend, which requires the viewer to hold on to too much information at once.

Transformation	Description
Zooming	This allows us to have look at a certain detail in the visualization Aggregation To combine many data points into a single group
Filtering	This helps us to (temporarily) remove data points that are not in our major focus
Outlier handling	This allows us to get rid of single points that are not representative of 99% of the dataset.

Let's consider the following example: You have visualized a graph and what came out of this was nothing but a mess of nodes connected through hundreds of edges (a very common result when visualizing so-called densely connected networks), one common transformation step would be to filter some of the edges. If, for instance, the edges represent money flows from donor countries to recipient countries, we could remove all

flows below a certain amount (n.d.).

3.2.6 Adapt your story to a different set of audiences

Jonathon Corum is a graphics designer for The New York Times and he provided a very informative talk to a strictly scientific audience on how to create and design visualizations that explain material originally created for a certain audience, i.e. the scientific community, but now is to be related to a different audience (in his case, the readership of the Times or maybe the public at large). The talk is filled with examples and breakdowns of how he has moved from his base content to the final product, all of which are illuminating examples by themselves. There is also great power in the broader themes that he is trying to convey.

Of course, it is easy to assume that we know the audience we are producing the work for, but even in this step, we should focus on the ultimate goal of conveying, understanding and explaining a concept. Some of the main highlights to help make this connection with the audience involved are mentioned below:

Principle	**Description**
Focusing the attention	What can be removed? Realize that consistency can help eliminate unnecessary distractions. There may be a trade-off between losing information but conveying the ultimate meaning more clearly. Label important things rather than relying on a legend, which requires the viewer to hold on to too much information at once.
Involving your audience	Give them opportunities to connect their own general knowledge on the topic. Use real world comparisons or examples to help build and relate context. Encourage comparisons and make this easy for the viewer to process and see.
Explaining why	Providing context, adding time sequence details, showing movement, change and mechanism will all guide your audience in connecting the dots and understanding the significance of what you are trying to communicate.

3.2.7 Developing Intuitive Dashboards

Often, data visuals end up too intricate and overly complicated. A dashboard should be appealing but also easy to understand. Following these rules will lead to the effective presentation of the data.

Best Practice	Description
The dashboard should read left to right	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 since dashboards are often used as a quick convenient way to look up data.
Group related information together	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.

Best Practice	Description
Find relationships between seemingly unrelated areas and display visuals together to show the relationship.	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 are 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.
Choose metrics based on why they matter	Chosen metrics should be important and relevant to the current task. That doesn't mean that each metric ought to be incorporated. You ought to be highly selective in determining which metrics earn a spot on your dashboard. Organization's core objectives, availability of data that can shed light on the objectives, effectiveness of metric to explain contribution to the objectives etc. are some of the aspects to consider while choosing metrics. In short, every metric on your dashboard should connect to the organization objectives.
Keep it visual	Dashboards are meant to be fast and easy to read. A well-designed, highly visual dashboard will be more widely adopted by audiences. Since metrics are also chosen in line with corporate objective, it will help in speeding peoples' understanding. This will also help see the translation of individual department objectives into broader organizations objective.
Make it interactive	Interactive, highly visual dashboards should enable audience to perform basic analytical tasks, such as filtering the views, drilling down, examining underlying data etc. Viewers should be able to get the big picture from the dashboard and then be able to drill down into a view that tells them the information they need to get their jobs done.
Keep it current or don't bother	Selected metrics should reflect current business challenges. You don't need up-to-the-minute data. Data can be current quarterly, weekly, hourly, etc. as relevant to the timeline of the organization. Ability to change and update the metrics represented in the dashboard is an important aspect.
Make it simple to access and use	Making dashboards easily accessible is critical. Web distribution is ideal for this - especially if dashboards can constantly pull current data and can adhere to IT protocols and security standards. Another alternative is posting files on websites, Wiki's or blogs.

3.2.8 More on Best Practices

Five Practices	**Explanation**
Find the compelling narrative	Along with giving an account of the facts and establishing the connections between them, don't be boring. You are competing for the viewer's time and attention, so make sure the narrative has a hook, momentum, or a captivating purpose. Finding the narrative structure will help you decide whether you actually have a story to tell. If you don't, then perhaps this visualization should support exploratory data analysis (EDA) rather than convey information.

Five Practices	Explanation
Think about the audience	If you think about data visualization as storytelling, then you realize you need to tailor your story to your audience; Novice: first exposure to the subject, but doesn't want oversimplification; Generalist: aware of the topic, but looking for an overview understanding and major themes; Managerial: in-depth, actionable understanding of intricacies and interrelationships with access to detail; Expert: more exploration and discovery and less storytelling with great detail; Executive: only has time to glean the significance and conclusions of weighted probabilities. When you tell the right story to the right audience you are able to identify data points that resonate with the audience.
Be objective and offer balance	A visualization should be devoid of bias. Even if it is arguing to influence, it should be based upon what the data says–not what you want it to say. There are simple ways to encourage objectivity: labeling to avoid ambiguity, have graphic dimensions match data dimensions, using standardized units, and keeping design elements from compromising the data. Balance can come from alternative representations (multiple clustering's; confidence intervals instead of lines; changing timelines; alternative color palettes and assignments; variable scaling) of the data in the same visualization.
Don't censor	Don't be selective about the data you include or exclude, unless you're confident you're giving your audience the best representation of what the data "says". This selectivity includes using discrete values when the data is continuous; how you deal with missing, outlier and out of range values; arbitrary temporal ranges; capped values, volumes, ranges, and intervals. Viewers will eventually figure that out and lose trust in the visualization (and any others you might produce).
Edit, Edit, Edit	Take care to really try to explain the data, not just decorate it. Don't fall into "it looks cool" trap, when it might not be the best way explain the data. As journalists and writers know, if you are spending more time editing and improving your visualization than creating it, you are probably doing something right.

3.3 Dashboards

(Taylor 2018),(tableau, n.d.)

(Few 2007) > "A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance." -Stephen Few

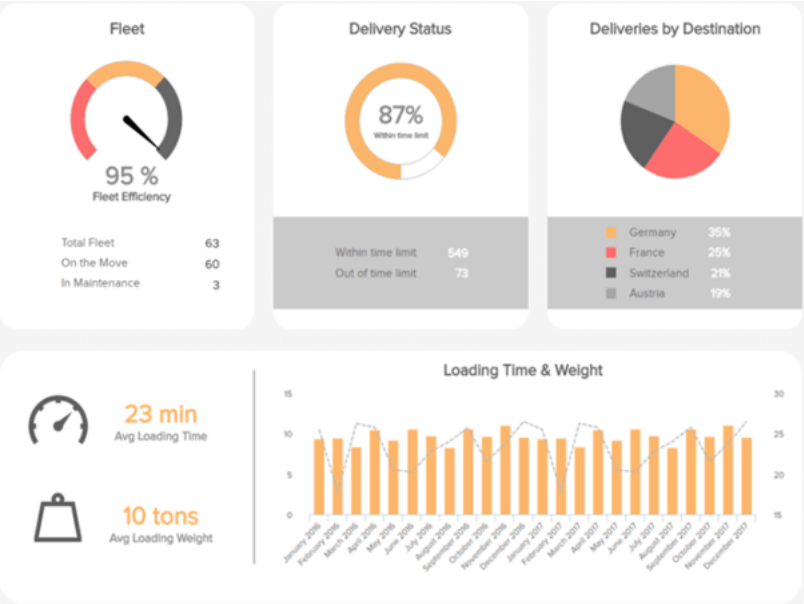
Dashboards display aggregated information visually and understandably. They give a comprehensive overview of a company, business department, process or project concerning achieving specific objectives, that are measured by Key Performance Indicators (KPIs). Also, they provide easy and immediate access to actionable analytics that can affect the bottom line of an entity.

Some of the advantages associated with dashboards are:

1. They are customizable
2. They are interactive
3. They allow for Real-Time monitoring
4. All of the data is in one place
5. They are intuitive
6. They get everyone on the same page
7. They are capable of gaining viewers focus

Below is an example of a dashboard for a logistics of a transportation industry. When it comes to logistics,

every moment matters, and maximum deliveries are expected to be on time. This transportation dashboard makes it easy with delivery status, fleet efficiency, average loading time and other logistics KPIs. (Lebied 2017)



3.4 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.

3.4.1 Brief Description of popular tools

Tool	Description
Tableau	The most popular in 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	The main competitor to Tableau and 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.
FusionCharts	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	It 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.

Tool	Description
Datawrapper	It 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	It 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	It can bring together multiple sources of data for easier access. It can even work with large data sets. Sisense makes it easy to share finished products across departments, ensuring everyone can get the information they need.
Altair	It is a statistical visualization library for Python, based on Vega and Vega-Lite. Its sources are widely available on GitHub. With Altair, we can understand the data and its meaning in a better way. Altair's API is very simple to use. This is simple, elegant and produces beautiful and effective visualizations with a minimal amount of code.
Shiny	Shiny is an open package from RStudio, which provides a web application framework to create interactive web visualization called Shiny apps. The ease of working with Shiny has what popularized it among R users.
Microsoft Office	Microsoft Office uses a variety of tools and combined with their data source, excel, it can create simple, well designed, and intuitive graphs. Excel is a very popular took for viewing raw data, and there are tools within the program to create graphs based on the data.
Google Suite	Google Suite includes their docs, sheets, and slides (among others). Sheets is a great tool for holding raw data, similar to Excel. Within sheets, theres are simple tools to create graphs that quickly update based on the changing data. One significant quality Google Suite provides is the interactivity and availability for collaboration on the same documents.

3.4.2 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 (Kerschberg 2014). Interactive Data Visualization Software has the following benefits:

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

There are multiple ways by which interactive data visualizations can be developed. D3.js is one of the ways to build an interactive data visualization.

3.4.3 Python for Data Visualization 10 Useful Python Data Visualization Libraries

(Bierly 2016)

It starts with the insights of learning d3.js by showing interviews with those top visualization practitioners. Then the author gives key concepts and useful features for learning visualization like d3-shape, d3 selection, d3-collection, ds-hierarchy, ds-zoom as well as d3-force. Sample charts for each

Library	Description
Matplotlib	Because matplotlib was the first Python data visualization library, many other libraries are built on top of it or designed to work in tandem with it during analysis. While matplotlib is good for getting a sense of the data, it's not very useful for creating publication-quality charts quickly and easily.
Seaborn	Seaborn harnesses the power of matplotlib to create beautiful charts in a few lines of code. The key difference is Seaborn's default styles and color palettes, which are designed to be more aesthetically pleasing and modern. Since Seaborn is built on top of matplotlib, you'll need to know matplotlib to tweak Seaborn's defaults.
Ggplot	ggplot is based on ggplot2, an R plotting system, and concepts from The Grammar of Graphics. ggplot operates differently than matplotlib: it lets you layer components to create a complete plot. For instance, you can start with axes, then add points, then a line, a trendline, etc. Although The Grammar of Graphics has been praised as an "intuitive" method for plotting, seasoned matplotlib users might need time to adjust to this new mindset.
Bokeh	Like ggplot, Bokeh is based on The Grammar of Graphics, but unlike ggplot, it's native to Python, not ported over from R. Its strength lies in the ability to create interactive, web-ready plots, which can easily give the output as JSON objects, HTML documents, or interactive web applications. Bokeh also supports streaming and real-time data.
Pygal	Like Bokeh and Plotly, pygal offers interactive plots that can be embedded in the web browser. Its prime differentiator is the ability to output charts as SVGs. As long as you're working with smaller datasets, SVGs will do you just fine. But if you're making charts with hundreds of thousands of data points, they'll have trouble rendering and SVG will become sluggish.
Plotly	You might know Plotly as an online platform for data visualization, but did you also know you can access its capabilities from a Python notebook? Like Bokeh, Plotly's forte is making interactive plots, but it offers some charts you won't find in most libraries, like contour plots, dendograms, and 3D charts.
Geoplotlib	geoplotlib is a toolbox for creating maps and plotting geographical data. You can use it to create a variety of map-types, like choropleths, heatmaps, and dot density maps. You must have Pyglet (an object-oriented programming interface) installed to use geoplotlib. Nonetheless, since most Python data visualization libraries don't offer maps, it's nice to have a library dedicated solely to them.
Gleam	Gleam is inspired by R's Shiny package. It allows you to turn analyses into interactive web apps using only Python scripts, so you don't have to know any other languages like HTML, CSS, or JavaScript. Gleam works with any Python data visualization library. Once you've created a plot, you can build fields on top of it so that users can filter and sort data.
Missingno	Dealing with missing data is a pain. Missingno allows you to quickly gauge the completeness of a dataset with a visual summary, instead of trudging through a table. You can filter and sort data based on completion or spot correlations with a heatmap or a dendrogram.

Library	Description
Leather	Leather's creator, Christopher Groskopf, puts it best: "Leather is the Python charting library for those who need charts now and don't care if they're perfect." It's designed to work with all data types and produces charts as SVGs, so you can scale them without losing image quality. Since this library is relatively new, some of the documentation is still in progress. The charts you can make are pretty basic but that's the intention.

3.4.4 R for Data Visualization: Grammar of Graphics

Chapter 3 of Grolemund and Wickham's "R for Data Science" (Grolemund and Wickham 2017)

3.4.4.1 Layered Grammar of Graphics:

The grammar of graphics is based on the implication that you can uniquely describe any plot as a combination of

a dataset

a geom

a set of mappings

a stat

a position adjustment

a coordinate system

a faceting scheme.

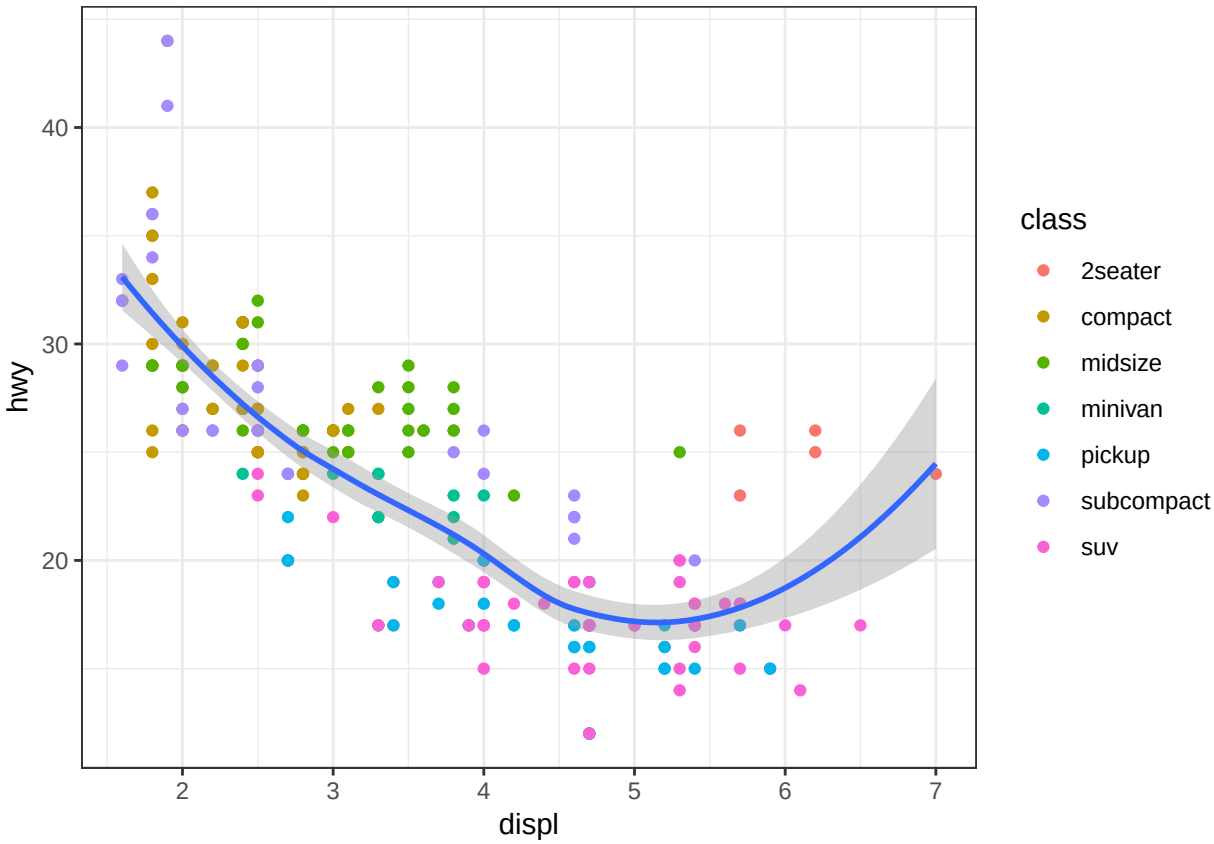
3.4.4.2 Aes/Mapping: Global Mapping and Local Mapping

Formula:

$$ggplot(data = DATA) + GEOMFUNCTION(mapping = aes(MAPPINGS), stat = STAT, position = POSITION) + CO$$

```
library("tidyverse")
library("gapminder")
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +  geom_point(mapping = aes(color = class)) +  ge
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Here, "mapping = aes(x = displ, y = hwy)" is a **global mapping**, where "mapping = aes(color = class)" is a **local mapping**.

3.4.4.3 Position Adjustment

a) "Identity" position will place each object exactly where it falls in the context of the graph. This is not very useful for bars.
b) "Fill" position works like stacking, but makes each set of stacked bars the same height. This makes it easier to compare proportions across groups.
c) "Dodge" position places overlapping objects directly beside one another, which makes it easier to compare individual values.
d) "Jitter" position adds a small amount of random noise to each point. This spreads the points out because no two points are likely to receive the same amount of random noise.

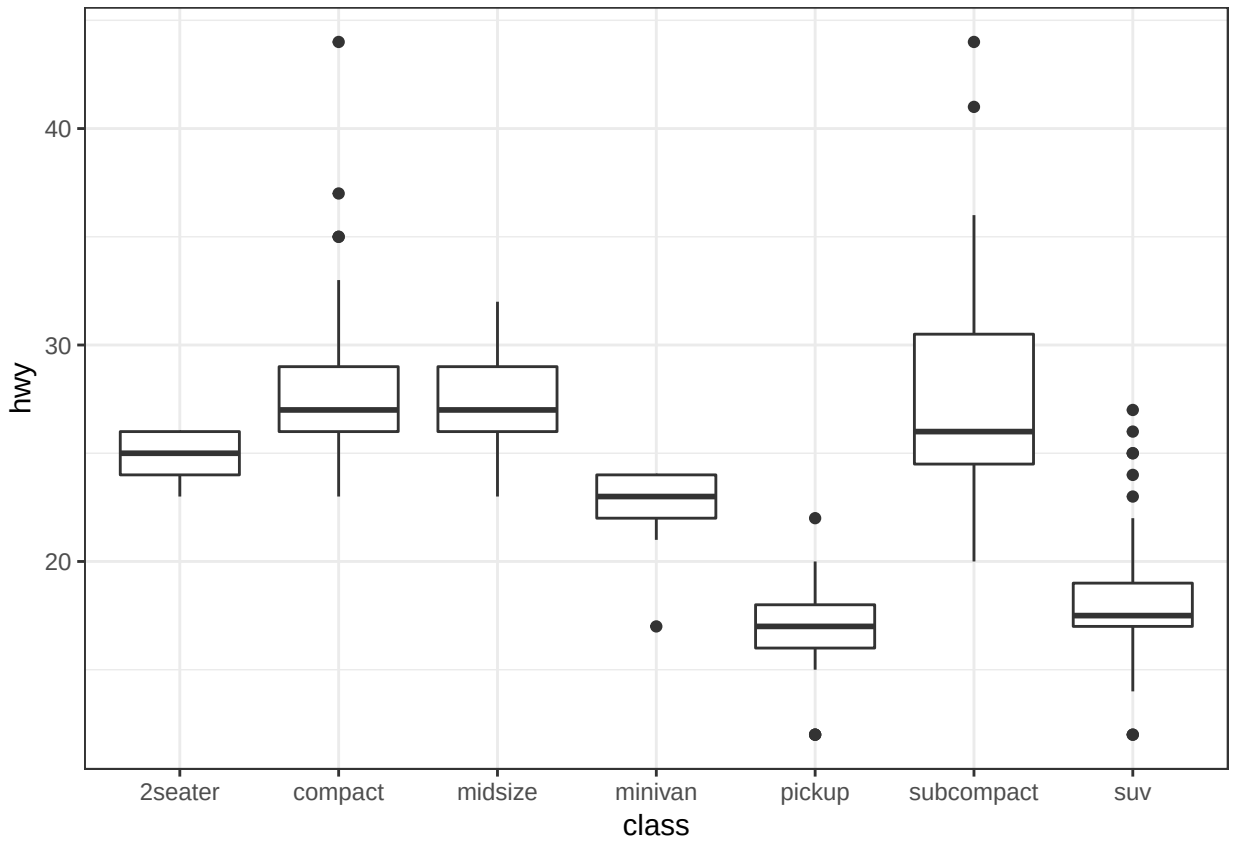
3.4.4.4 Coordinate Systems

The default coordinate system is Cartesian.

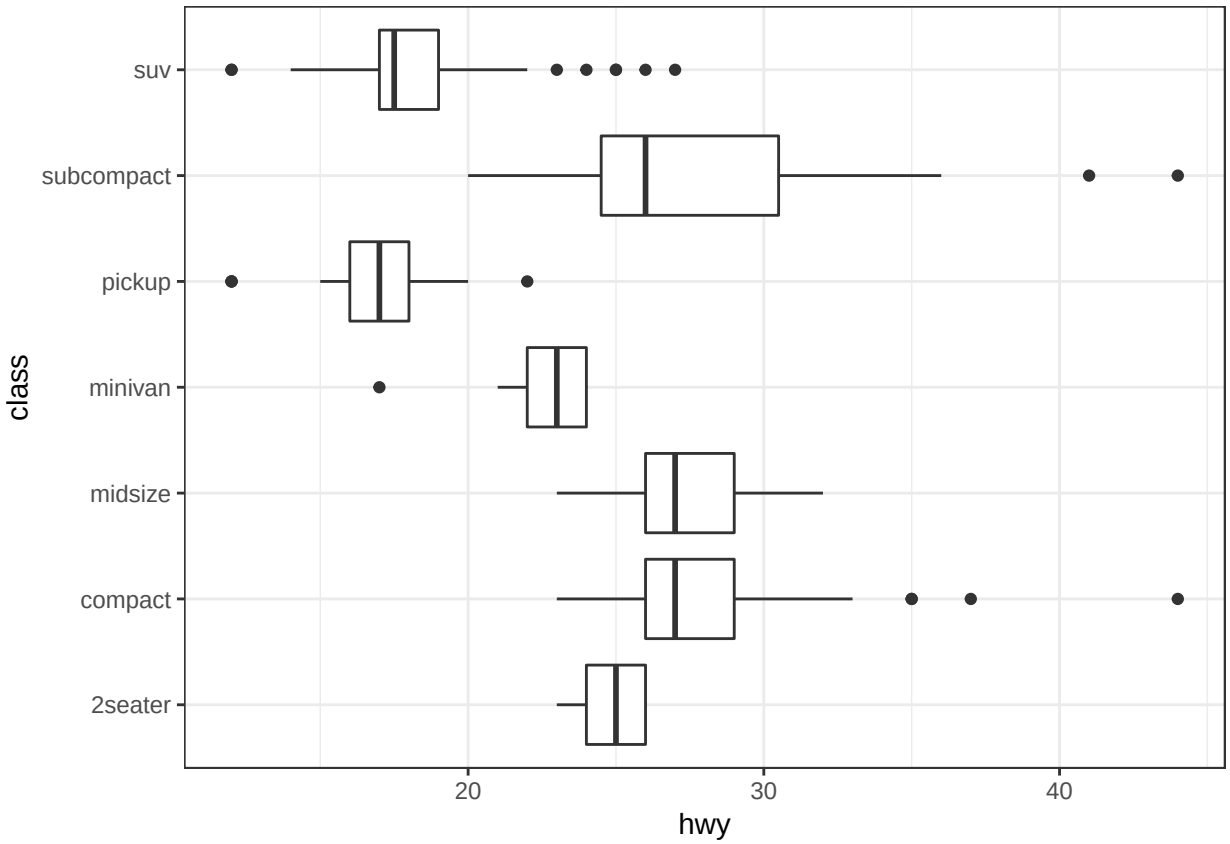
Command	Description
coord_flip()	It switches the x- and y-axes. Very useful if you want horizontal boxplots.
coord_quickmap()	It sets the aspect ratio correctly for maps. This is very important if you draw a map.
coord_polar()	It uses polar coordinates. Polar coordinates reveal interesting connections between a bar chart and a Coxcomb chart.
coord_quickmap()	sets the aspect ratio correctly for maps. This is very important if you draw a map.

Command	Description
coord_polar()	uses polar coordinates. Polar coordinates reveal interesting connections between a bar chart and a Coxcomb chart.

```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) + geom_boxplot()
```




```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +  geom_boxplot() +
  coord_flip()
```



3.4.5 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 (Castañón 2016). 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 advantage of using R Shiny is : 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 interactive data experience without (or with much less) web development and without the limitations or cost of proprietary BI tools.

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

3.4.6 D3.js

D3.js stands for Data-Driven Document, a JS library for interactive Big Data visualization in literally ANY way required real-time(Cabot Technology Solution 2017). This is not a tool, mind you, so a user should have a solid understanding of JavaScript to work with the data and present it in a humanly-understandable form. To say more, this library renders the data into SVG and HTML5 formats, so older browsers like IE7 and 8 cannot leverage D3.js capabilities.

The data gathered from disparate sources like huge-scale datasets is bind in real-time with DOM to produce interactive animations (2D and 3D alike) in an extremely rapid way. The D3 architecture allows the users to intensively reuse the codes across a variety of add-ons and plug-ins. Some of the key advantages are: It is a dynamic, free and open source and very flexible with all web technologies, the ability to handle big data and the functional style allows to reuse the codes.

The Hitchhiker' Guide to d3.js (Ian 2017) is a wonderful guide for self-teaching d3.js. This guide is meant to prepare readers mentally as well as give readers some fruitful directions to pursue. There is a lot to learn besides the d3.js API, both technical knowledge around web standards like HTML, SVG, CSS and JavaScript as well as communication concepts and data visualization principles. Chances are you know something about some of those things, so this guide will attempt to give you good starting points for the things you want to learn more about.

It starts from the insights of learning d3.js by showing interviews with those top visualization practitioners. Then the author gives key concepts and useful features for learning visualization like d3-shape, d3 selection, d3-collection, ds-hierarchy, ds-zoom as well as d3-force.

The guide is helpful as it lists a lot of useful resources links for learning d3.js. For example, it recommends d3 API Reference, 2000+ d3 case studies and tutorials for d3. It contributes tremendously in doing exploratory analysis version of group project of this class on d3. Further, the guide provides information such as some meetup groups in the bay area, which can be helpful in connecting with data professionals and building up networks.

3.4.7 Tableau

Tableau is amid the market leaders for the Big Data visualization, especially efficient for delivering interactive data visualization for the results derived from Big Data operations, deep learning algorithms and multiple types of AI-driven apps (AbsentData 2018).

Tableau can be integrated with Amazon AWS, MySQL, Hadoop, Teradata ,and SAP, making this solution a versatile tool for creating detailed graphs and intuitive data representation. This way the C-suite and middle-chain managers are able to make grounded decisions based on informative and easily-readable Tableau graphs.

Tableau is a 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.

Key advantages:

It provides a 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 visual 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 ("Data Visualization Best Practices" 2017)("The Extreme Presentation Method," n.d.).

It is possible to create new calculated fields within Tableau by using functions on existing fields

A Tableau file can be saved with the data attached so that it does not need to remain on the same hard drive/cloud as the data.

[Tableau Public]https://public.tableau.com/s/ is available to showcase work on a user profile online

3.4.7.1 Using Multiple Data Sources

(Tableau 2019c)

In Tableau data source there are two ways to add in data from multiple sources. The first is joining the data, which will add two datasets together at the row level related by specific columns. When joining tables, the fields being joined must be of the same data type. There are four types of joins: inner, left, right and full outer. Note that when possible perform joins outside of Tableau and import one dataset in order to maximize performance.

The second is blending multiple data sources, which keeps two or more data sources separate from each other but displays the information together. Blending allows for data to have different levels of detail such as aggregate number of transactions per month vs. individual transactions. It also is used when joins will not work, such as having transactional data in one source and quota data in another. Blending requires at least one common field between both data sources. If the field names are different but the two columns contain the same values, the relationship can be defined by changing the column names in one data source to match the other, or defining the relationship manually (Technology mart 2017).

Steps For Joining Data

Connect the first data source (dragging the file wanted to the canvas if there are multiple options)

Add another connection (there should be two overlapping circles on the canvas where the two datasets overlap)

Click on the join relationship (the circles) to add a join type and data-match !(joins)[https://onlinehelp.tableau.com/current/pro/desktop/en-us/Img/joins_joindialogbox.png]

Choose from the different types of joins and then which column in each dataset that will match (ie. the orderID column in both a sales dataset and a shipping dataset)

Steps For Blending Data

Add your first data source

Go to Data > New Data Source

Choose your second dataset

Your primary data source (the first dataset used when dragging dimentions or measures in the sheet) should have a blue tick mark beside it

Secondary data source will have an orange tick mark next to it.

Tableau will try to automatically define the relationship between the two datasets using columns with the same names.

Define the relationship manually by going to Data > Edit Relationships > Custom > Add a Relationship

Joins and Data Blending work when it is required to append columns from one table to another. During a situation where we need to add rows from one table to another, Union functionality can be used. For instance, when data is available in separate tables for different months; each table contains the same information but only for the relevant month. Now if it's required to club these data, then in this case union functionality will be useful.

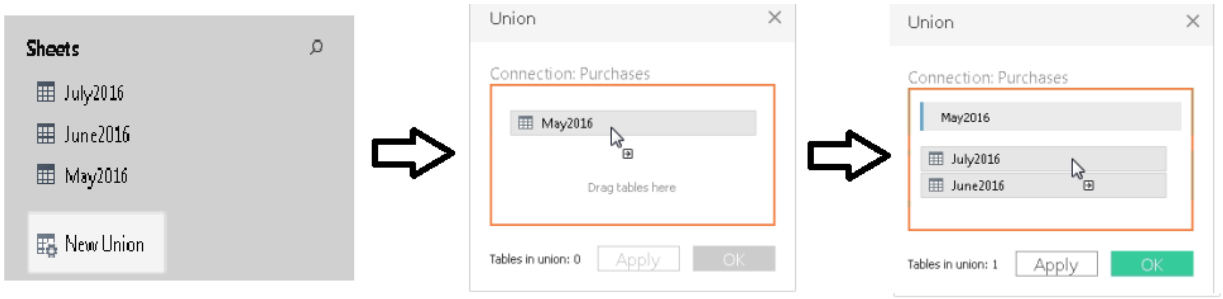
Steps for using Union

On the data source page, double-click New Union to set up the union.

Drag a table from the left pane to the Union dialog box.

Select another table from the left pane and drag it directly below the first table.

Click Apply or OK to union

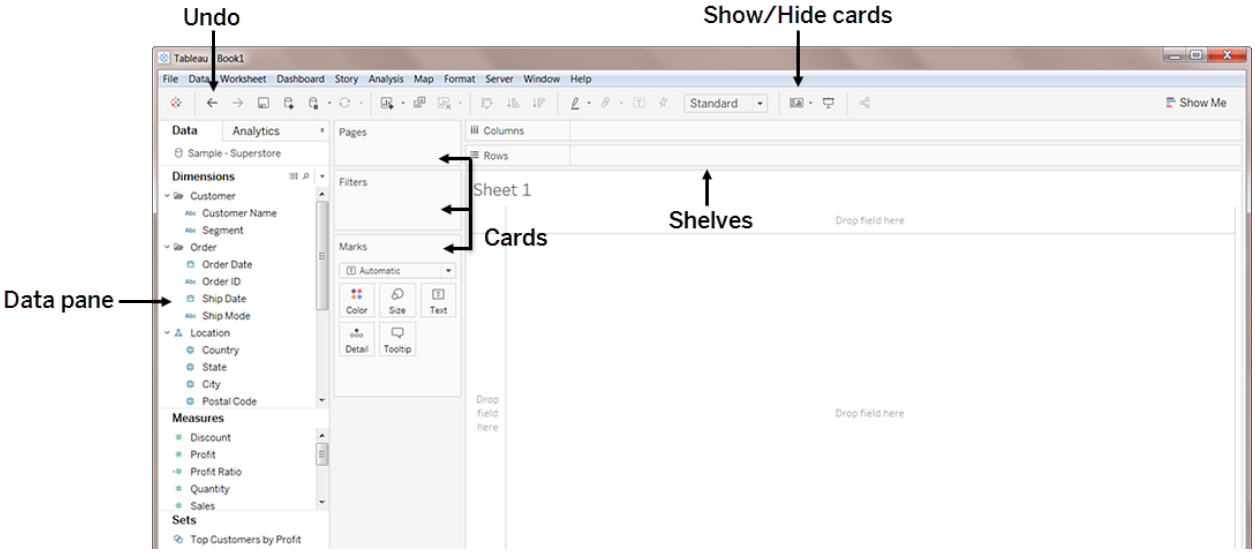


("Union Your Data" 2019)

3.4.7.2 Getting Started by Creating a First View

(Tableau.com 2019)

The types of visualization will change for each dataset, but Tableau's layout will be the same regardless, so it is important to know more about the main features for creating visual of your data.



Element	Description
Data Pane	Displays the data set that is in the view and the fields (columns) in the data set that are automatically sorted into dimensions or measures (explained below).
Cards	Cards are containers for different controls available in Tableau.
Marks	This is a main way of manipulating visual aspects of graphs and charts. You can encode your data fields with elements like type, color, size, shape etc. You can have multiple *measures* and *dimensions* added to the marks cards to add different properties such as different regions denoted by shapes and quantity denoted by size.
Summary	The summary card is available on the *show/hide cards* toolbar menu to add summary statistics either for the data selection, or the entire data source.
Shelves	Shelves are a type of card.

Element	Description
Rows and Columns	Rows and columns are the shelves where you drop data fields in order to create the structure of your visualization. When creating a table visualization, the columns will create the columns of the table, while the rows would create the rows of the table. Any number of fields can be dragged onto the shelves to create a more granular picture of the data. When a *dimension* is placed on the shelf, headers for the member of the dimension are created. When a *measure* is placed on a shelf, the quantitative axes for the measure are created.
Filters	The filters shelf is where you specify the data to include and exclude by dragging *measures*, and/or *dimensions*. For example, you can put your date into the filter if you only want to see revenue from the last quarter, or you can put a categorical dimension into the filter if you only want to see data for specific products.
Pages	The pages shelf allows for a series of separate page views with a common axis to better analyze how a certain *measure* affects the rest of the data. When a field is added to the pages shelf a control panel is added to the right side of the view to move through the pages (either manually or automatically).
Measure Values	At the bottom of the *measures* data pane is a special field that contains all measures of the data collected into one field.

Calculation Type	Description
Dimension	Qualitative or categorical data that would normally appear as column headers for rows of data that normally defines the granularity that shows in the view, i.e. customer name or order date.
Measure	Quantitative or numerical data that can be aggregated based on a given dimension, i.e. total sales (measure) per region (dimension)

3.4.7.3 Using Shapes as Filters in Tableau When Your Fields Are Measures

This article (Brett 2018) introduces the methodologies on how to use shapes as filters in Tableau when your fields are measures. It teaches you how to load custom shapes as action filters and use them for showing different graphs with those filters, which can make your visualization more interesting and interactive. You can also download the Tableau file for practice.

3.4.7.4 Creating Calculated Fields

(Tableau 2019a) On the Data Source tab of a Tableau file there will be the complete set of data imported that is available to use for your visualizations. This list also contains the opportunity to create new fields, based on those existing. "You can use calculated fields for many, many reasons. Some examples might include:

To segment data

To convert the data type of a field, such as converting a string to a date.

To aggregate data

To filter results

To calculate ratios"

(Tableau 2019a) These calculated fields can also be created in the visualization pages. This allows the user to combine fields for more robust vizualizations, without the need to create another field.

Calculation Type	Description
Basic calculations	Basic calculations allow you to transform values or members at the data source level of detail (a row-level calculation) or at the visualization level of detail (an aggregate calculation).
Level of Detail (LOD) expressions	Just like basic calculations, LOD calculations allow you to compute values at the data source level and the visualization level. However, LOD calculations give you even more control on the level of granularity you want to compute. They can be performed at a more granular level (INCLUDE), a less granular level (EXCLUDE), or an entirely independent level (FIXED) with respect to the granularity of the visualization.
Table calculations	Table calculations allow you to transform values at the level of detail of the visualization only. (Tableau 2019a)

Basic Calculations are the most common, and many of the functions found in Excel can be used such as:

If/Then Statements and other logical functions

Date Functions

AVG, COUNT, MAX, MIN and other Aggregation functions

Functions involving simple math using two or more of the existing fields

3.4.7.5 Creating Groups

(Tableau 2019b) Creating a group in Tableau will combine items in a field that are related to each other in some way and allow you to visualize the new grouped data, while the old non-grouped data still remains. The paperclip icon in Tableau is used for grouping.

There are many ways to create groupings of data, and different reasons for grouping data in a certain way:

Selecting multiple data points in the view and group them together using the group icon.

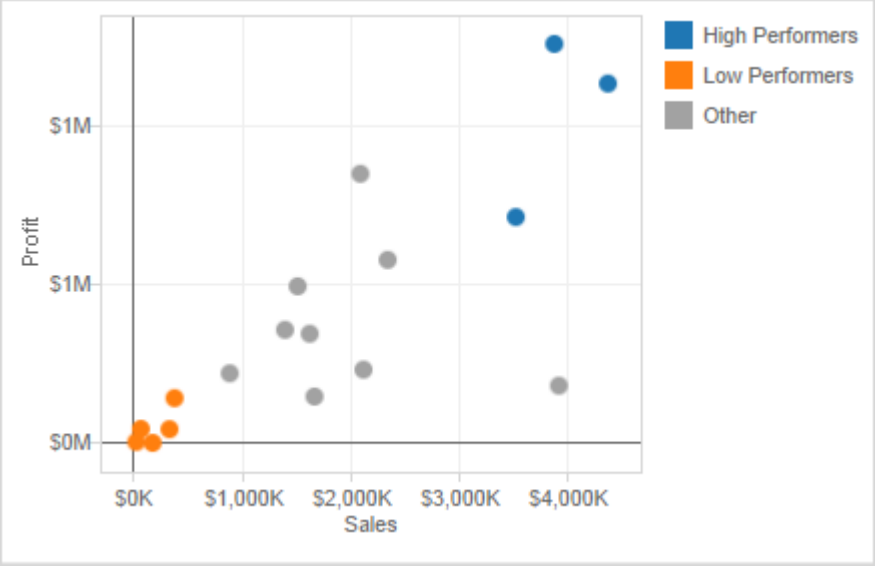
For example, you might want to do this with mislabeled data from input error. i.e. columns that are titled "CA" and "California" would be depicting the same data but would be graphed separately. We can combine these into a group and rename the group according the rest of the dataset's naming convention for states.

Groups can be created from a field in the data pane by clicking on the field containing the data needing to be grouped and selecting **create** > **group**. Select the members of field you want to group and click **group**.

For example, you might want to do this to view smaller items in a larger aggregate grouping. i.e. If expenditure on office supplies were being graphed and we wanted to put all the small items like paperclips, pencils, pens, and sticky notes in an aggregate category called "small supplies".

Creating an "other" group when making multiple grouped categories is useful for grouping all the remaining un-grouped members. This can be done by clicking on the grouped field, and selecting **Edit Group**, and selecting **Include Other**.

This can help highlight certain groups against everything else. For example, if the visualization is intended to show high performing products and low performing products, then creating an "other" group will help draw



attention to the focus of the graph.

Coloring a view using groups helps to visually identify a group of related marks. You can select the marks directly by pressing and holding the **Shift** key to select multiple marks and click the *paperclip* (Group) button on the toolbar and select **Group**.

It is possible that when creating groups this way, the marks will represent more than one dimension. You can choose to group on all dimensions or just a particular dimension.

This is useful for showing things like outliers, or highest performers.

3.4.8 TabPy

(Beran 2017)
Imagine a scenario where we can just enter some x values in a dashboard form, and the visualization would predict the y-variable! TabPy allows us to integrate and visualize data from Python in Tableau.

The author here has given an example in which he tries to identify criminal hotspots in the area using data from Seattle's police department's 911 calls. The author uses machine learning (spatial clustering) and creates a great interactive visualization which allows viewers to click on the type of criminal activity to show various clusters.

There are other examples and use cases that may be downloaded, and the scripts are also given by the author to anyone who is interested in replicating the visualizations.

3.4.9 Google chart

A free and powerful integration of all Google power. The tool is rendering the resulting charts to HTML5/SVG, so they are compatible with any browser. Support for VML ensures compatibility with older IE versions, and the charts can be ported to the latest releases of Android and iOS. What's even more important, Google chart combines the data from multiple Google services like Google Maps. This results in producing interactive charts that absorb data real-time and can be controlled using an interactive dashboard ("Top 4 Big Data Visualization Tools" 2018).

3.5 Data Visualization in Business

Companies tend to rely on dashboards (a compilation of several related data visualizations) to give them high-level insights on company-wide, market-level, or employee-level performance. The following are some common applications of dashboards in business.

Application	Description
Sales & Marketing	This is one of the most popular uses for dashboards. Companies like to regularly track their revenue, conversions, lead sources, etc. and rely on data visualization to synthesize these large and constantly updated data into visual summaries.Funnel reporting in terms of sales velocity and efficiency, Camparing ROI, distribution of opportunities and leads across region, time, etc are some of the matrices which requires dashboards and visualization on latest as well as historic data. Sales and Marketing teams are one of the major consumers of BI tools driven reporting and monitoring dashboards.
Customer Success	These dashboards can be created by the team, but are also often built into customer service platforms such as Zendesk. They include various KPIs of the customer success team, such as the ratio of tickets open to tickets closed and time to resolution.
Product Management	These dashboards tend to synthesize sales, marketing, and customer research data together and are typically used for executive reporting. The visuals display metrics such as dollars and hours devoted to various projects and most requested features by customers.
Clinical Performance Management	Data visualizations are also helpful in the healthcare industry to monitor healthcare systems operations, clinical performance monitoring and patient profiling. Healthcare provider organizations (hospitals and health systems) can better examine their clinical, administrative and financial data to support clinical costing and resource coordination, better-planned care for patients and provide competitive advantage alongside maintaining quality standards.
Finance	Finance is another popular domain where dashboards help cover a variety of aspects such as- profit & loss, cash flow management, revenue, profit margin, cost heads etc. Finance dashboards can often be helpful in identifying trends on revenue, profitability, cash flows, accounts payable, day sales outstanding and so on. A key area to leverage finance dashboards is identifying performance of key metrics over a period of time and creating and comparing performance against internal (and/or external) benchmarks. There is no dearth of data that can flow into a financial dashboard!
Human Resources	Human resource is another critical functional domain where dashboards and reporting play a key role. HR and People Analysts are actively hired across organizations for measuring employee productivity, attrition or turnover rates, understanding training costs per employee, recruiting conversion rate, average employee retention period, cost per hire and so on. Human resource can constitute one of the major cost heads for most service companies and hence drives the need for HR management and reporting.

3.5.1 Data Visualizations in Industry

3.5.1.1 Healthcare

Data visualization is also used across many different industries. One popular area right now is healthcare, especially involving big data. The benefits and uses of interactive data viz are detailed in a paper from the University of Maryland (2013). (Schneiderman 2013)

The paper highlights three types of data that can and should be visualized to help in decision-making: personal, clinical, and public health information. Examples include: exploration of prescription patterns of different drugs and tracking personal health and fitness statistics. (Even the nice, clean Fitbit app home screen is a comprehensive dashboard!) Importantly, making sense of all this data collected from individuals will help healthcare organizations and companies provide more personalized and effective health treatment.

With data volumes increasing exponentially, health care can no longer rely on antiquated data presentation tools like spreadsheets and tables any more than modern computers can still use transistors. Spreadsheets and tables are outdated means of data-sharing which are time-consuming to produce and ineffective to consume, particularly with large amounts of data. ("A Healthcare Data Revolution – the Case for Data Visualization" 2019)

According to AHIMA (American Health Information Management Association), the healthcare industry is in the midst of a data revolution, storing more information than ever before. Healthcare data volumes are increasing at a 49% clip annually, according to a recent report in CIO Magazine.

Many organizations produce data visualizations in the areas of healthcare delivery, patient-facing applications, population health, public health, or global health. Some examples include: * The Institute for Health Metrics and Evaluation, a population health research center at UW Medicine, regularly features data visualization on its site regarding topics such as the social determinants of health and obesity. * Visualizing Health is a project of the Robert Wood Johnson Foundation and the University of Michigan Center for Health Communications Research that provides visualizations that communicate healthcare risk information. * The Center for Disease Control's National Center for Health Statistics offers a data visualization gallery based on the data the organization collects. * The Agency for Healthcare Research and Quality (AHRQ) offers a data visualization site that highlights findings from the Agency's Medical Expenditure Panel Survey, the Healthcare Cost and Utilization Project, and other AHRQ data sources. ("The Rise of Healthcare Data Visualization" 2017)

3.5.1.2 Media and Entertainment

The Data Visualization & Analytics trend has impacted all industries, including the media industry. As new technologies are being developed to automate and simplify the process of data analysis, and as throngs of data analysts are being trained and hired to meet the demand for the analysis of these data. For newspapers, television, magazines and Internet-only publishers, Data Visualization & Analytics strategies can include audience analytics to enable a better understanding and targeting of customers; tools to understand public and private databases for journalistic storytelling; tools to manage and search the exploding amount of video, social media and other content; tools to target advertising and ad campaigns; tools to automate the production of text and video stories, tools to identify waste and enable efficiency; and much more. ("Data Visualization and Analytics Transforming Media Industry" 2017)

There are several ways in which media and entertainment companies can benefit from visualization and analytics applications such as: * Audience Interest Analysis: Viewing history, searches, reviews, ratings, location and device data, click streams, log files and social media sentiment are just a few data sources that help take the guess work out of identifying audience interest. * Enhanced Program Scheduling: With the help of insights gained through data visualization and analytics, the entertainment companies are able to understand when customers are most likely to view content and what device they'll be using when they view it. Benefiting from the scalability of visualization and analytics solutions, this information can be analyzed at a granular ZIP code level for localized distribution. * Increasing Acquisition and Retention: Smart data visualization and analytics tools may help media and entertainment companies in understanding why consumers subscribe and unsubscribe, which will allow them develop the best promotional and product strategies to attract and retain customers. Unstructured data visualization and analytics sources best handled by data visualization and analytics applications such as call detail records, email and social media sentiment reveal often overlooked factors driving customer interest and churn. ("Big Data in Media and Entertainment," n.d.)

3.5.2 How visualization impacts Industry/business

(Lazarevich 2018b) 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:

3.5.2.1 Cleaning

The simplest way to explain the importance of visualization is to look at visualization as a means of 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. Data cleansing is typically done by using instance reduction techniques.

Instance reduction: It helps to reduce the size of the data set without compromising the quality of insights that can be extracted from the data. It removes instances and generates new ones to make the data set compact. There are two major instance reduction algorithms:

Instance selection: It is used to identify the best examples from a very large dataset with many instances in order to curate them as the input for the analytics system. It aims to select a subset of the data that can act as a replacement for the original dataset while completely fulfilling the goal. It will also remove redundant instances and noise.

Instance generation: It involves replacing the original data with artificially generated data in order to fill regions in the domain of an issue with no representative examples in the master data. A common approach is to relabel examples that appear to belong to wrong class labels. Instance generation thus makes the data clean and ready for the analysis algorithm.

Tools you can use: Drake, DataWrangler, OpenRefine

3.5.2.2 Extraction

Known versatile tools for data visualization and analytics like 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. The new generation of data visualization based on AR and VR technology, however, provides formerly unfeasible 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.

Data has shown phenomenal growth over the past decade and its widespread application by businesses as a growth catalyst continues to deliver positive results. The scale of data is massive and the volume, velocity, and variety of data call for more efficient processing to make it machine-ready. Although there is a multitude of ways to extract data such as public APIs, custom web scraping services, internal data sources, etc., there would always remain the need to do some pre-processing to make the data perfectly suitable for business applications.

Data pre-processing techniques that play a key role in the process are :

- Data cleansing
- Data Manipulation
- Data normalization
- Data Transformation
- Missing values imputation
- Noise identification
- Minimizing the pre-processing tasks

3.5.2.3 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 businesses. Then how data visualization become an important tool in your strategic kit?

The use of dashboards to present business statistics in a graphical manner charts, tables, and graphs helps the stakeholders keep track of the key indicators of the business and to focus on the areas that need to be improved. Building the dashboard application to impact a better decision-making process is the important aspect of business intelligence. It is observed that the higher adoption of latest technologies in business is resulting in higher return on investment and the low adoption causes the loss in the business.

According to Aberdeen Group, managers who utilize data visualizations are 28 percent more likely to find relevant information compared with those who use managed dashboards and reporting tools.They also discovered that 48 percent of those who use data visualizations are able to find the information they want without the help of tech support. Data Informed provides an excellent example of this on their blog: Business leaders for a supermarket chain can use data visualization to see that not only are customers spending more in its stores as macro-economic conditions improve, but they are increasingly interested in purchasing ready-made foods.(import.io, n.d.)

3.5.3 Corporate Scorecards and Data Visualization

Corporate transparency, flat organizations, open book policies, etc. are terms executives and entrepreneurs learn about all the time (Boost Labs 2015). As the corporate world shifts towards a more open culture, the demand for open data and insights have increased dramatically. This shift has helped the overall corporate strategic planning and management process easing the alignment of business activities towards a series of goals. Being transparent top down aligns the culture to sail towards the same North Star.

The growth of corporate transparency is not only important internally, but externally as well. Corporate certifications like B Corporations certifications (B Corp), require companies to provide a transparent view of their social conscious efforts to the general public. Achieving the certification is one step of the process; the true goal is to show the world how and why the certification is truly deserved.

Here's the process on how to get it done.

Step	Name	Description
1	Perform Data Discovery and Determine the Story	Before this step it is easy to underestimate the effort level it takes to pull the best insights from the data. Data manipulation products like Tableau, Domo, Pentaho, IBM's Many Eyes, and R, among others, make insight extraction that much easier to gain understanding of data using a visual medium. The key is to start with a simple portion of your data and to start pulling basic insights to visualize and correlate with each other. This process leads towards a compound series of questions, which helps provide an overall vision to the end product. We see the effect during our discovery process, which leads to unforeseen avenues for data intelligence.

Step	Name	Description
2	Data Infrastructure Setup	Data infrastructures can be simple or complex depending what the end goal is. Many clients prefer to go the route of complete data integration in order to centralize their data repositories. Technologies such as Hadoop have helped by unifying disparate data sources, but other options such as data cloud environments can help produce API's for future product deployments. Why is this important? Accessibility of data is an important foundation not only within the context of dashboards, but also the possibility of branching out to other products.
3	Product Design & Development	Wireframing, prototyping, and application development are the main engines to transform an idea into a final product. Products can range from static presentations/reports to full interactive applications. Mobile, tablet, TV, and workstation platforms can all be mediums to help deliver the final product. The secret to a great end product is how well the data story is conceptualized. If the story is weak then the end product will also suffer.
4	QA & Product Release	The best part of any project is to get it finalized and released for all to see. All data gets verified for accuracy, functionality testing (if applicable), application flow (if applicable), design testing, and remaining items are all completed. The end result is an engaging visual product for all intended audiences to see and use.

3.5.4 Demand for Data Literacy

The demand for data literacy is at an all-time high. Originally, data science was focused on the finance and tech industries but the demand for data science skills is increasing for every industrial section. ([Lifeline)(https://lifelinedatacenters.com/data-center/business-intelligence/))

Businesses generate more data everyday with what knowing the robust use cases are. Forbes states that a big hinderance which slows business progress is poor data literacy. Employers are now training their employess on data skills with the advantage of already knowing the business. But the demand still persists and this article states that automation is a key factor that can affect all of the sectors in section 3.5.1 but not enough employees have that toolset. (Forbes)

The end message is clean; no matter what industry or title you may have, adding skills related to data as mentioned in previous sections can truly increase business efficiencies.

3.6 Special Topics

3.6.1 Data Mining and Data Visualization

According to a paper in 2018(EDUCBA 2018), there are some key differences between data mining and data visualizations as suggested below:

Data Mining involves different processes such as data extraction, data management, data transformations, data pre-processing, etc.

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.

The author has also listed top 7 comparisons between data mining and data visualization, and 12 key differences between them. The article provides a very clear understanding of each of these techniques.

BASIS FOR COMPARISON	Data Mining	Data Visualization
Definition	Searches and produces a suitable result from large data chunks	Gives a simple overview of complex data
Preference	This has different applications and is preferred for web search engines	Preferred for data forecasting and predictions
Area	Comes under data science	Comes under the area of data science
Platform	Operated with web software systems or applications	Supports and works better in complex data analyses and applications
Generality	New technology but underdeveloped	More useful in real time data forecasting
Algorithm	Many algorithms exist in using data mining	No need of using any algorithms
Integration	Runs on any web-enabled platform or with any applications	Irrespective of hardware or software, it provides visual information

3.7 Implications of Good Data Visualization

Raw data is often meaningless or at the very least is difficult to derive immediate meaning from. When people face a broad set of measurements and/or in large quantities, they are unable or unwilling to spend the time required to process it. Technological advances of the Digital Age contribute to an ever-growing pool of "big data" and have dramatically improved our ability to collect such large amounts of information. Thus, filtering, visualization, and interpretation of data becomes increasingly important.

We should understand how to best derive meaning from data, but first we should understand why its presentation in graphical format is so powerful. Furthermore, while the ideal purpose of data visualization is to facilitate understanding of data, visualization can also be used to mislead. Some of the main methods of doing so are omitting baselines, axis manipulation, omitting data, and ignoring graphing convention. Examples of these methods will be explored later in this chapter.

SNo.	Principle	Description
1.	**Easy Recall**	People can process and remember images quicker than words. When data is transformed into images, the readability and cognition of the content greatly improves.
2.	**Providing Window for Perspective**	With infographics, you can pack a lot of information into a small space. Colors, shape, movement, the contrast in scale and weight, and even sound can be used to denote different aspects of the data allowing for multi-layered understanding (Mullis 2015).
3.	**Enable 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 correlation and analytical insights can be easily derived.
4.	**Increase in User Participation**	Interactive infographics can substantially increase the amount of time someone will spend with the content and the degree to which they participate in the information, both in its collection and its dissemination.

3.7.1 Typography and Data Visualization

Typography is the art and technique of arranging type to make written language legible, readable and appealing when displayed. (WIKI)

The arrangement of type involves selecting typefaces, point sizes, line lengths, line-spacing (leading), and letter-spacing (tracking), and adjusting the space between pairs of letters.

3.7.1.1 Preattentive visual attributes and typography

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, other typographic elements in data visualization are less prevalent.

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 disparate situation 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.

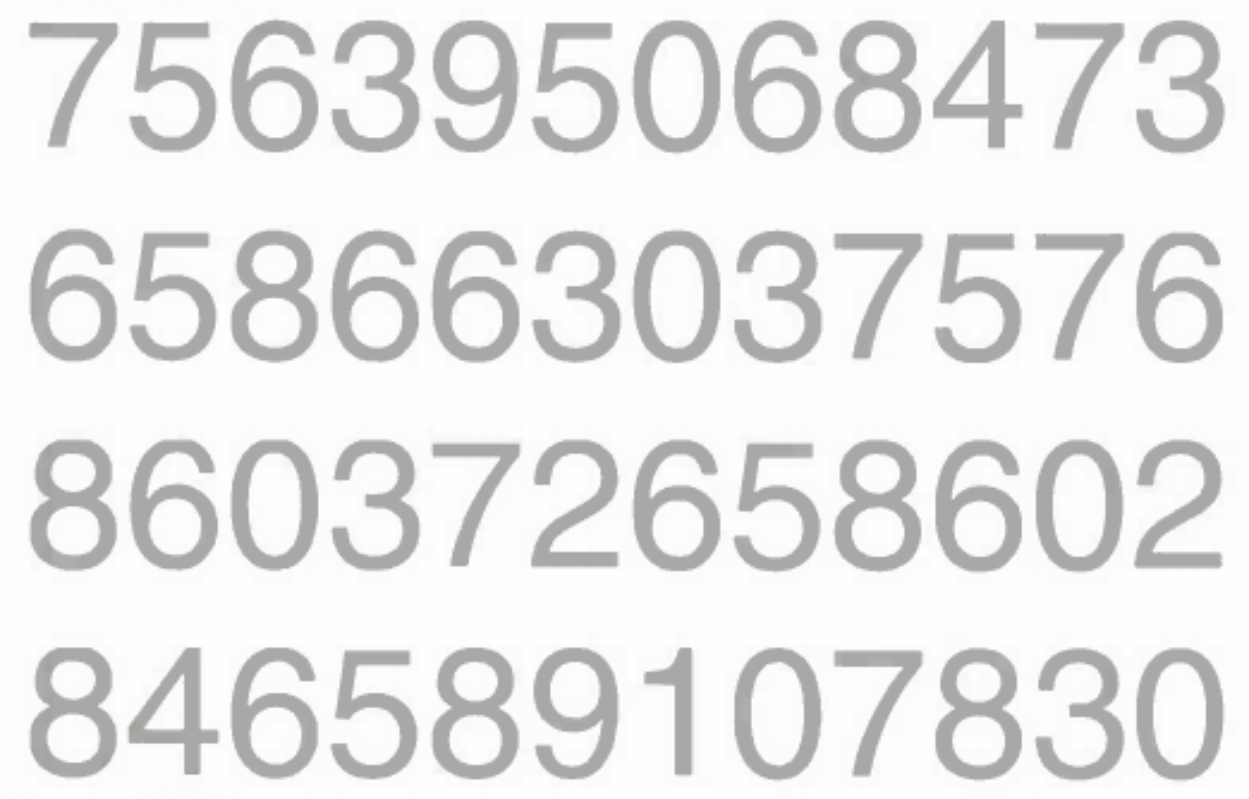
Furthermore, these visual attributes are inherently more viscerally powerful, and they are easy to code in a variety of programming languages. Technology has also perhaps previously limited the use of typographic attributes, for only recently have fine details such as serifs, italics, etc. been made readily visible to the audiences of data visualizations by technological advances.

Benefits of Using Preattentive Attributes
(Hepworth 2015) There are many benefits to using preattentive attributes in your visualization, mainly that it helps direct your audience's attention to where you want them to focus it. It can also be used to create a visual hierarchy of elements to lead your audience through the information you want to communicate in the way you want them to process it.

By understanding how our audience sees and processes information, we put ourselves in a better position to be able to communicate effectively.

Seeing it in Action
Here are some examples of preattentive attributes in action.

Taking note of how you process the information and how long it takes, quickly count the number of 3s that appear in the sequence below:



Count the 3s example without preattentive attributes.

Source: https://kathep.com/tools/readings/focus-your-audiences-attention/

The correct answer is six. In the above figure, there were no visual cues to help you in concluding your answer. Thus making it a challenging exercise, where you scan through the four lines of text, looking for the number 3 to count their occurances.

Now let's see what happens when we make a slight change to the block of numbers, by adding colours to it.

Repeat the above exercise of counting the number of 3s using the below figure.

Count the 3s example with preattentive attributes.

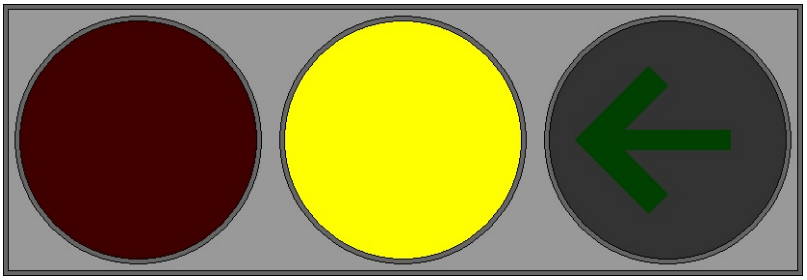
Source: https://kathep.com/tools/readings/focus-your-audiences-attention/

Note how much easier and faster the same exercise when we added the colour to the number 3. Even without blinking or thinking, all six 3s become very obvious and cannot be missed. This is because in this second iteration, your iconic memory is being leveraged. The preattentive attribute of intensity of color, in this case, makes the 3s the one thing that stands out as distinct from the rest. Our brain picks up on this without us having to dedicate any conscious thought to it.

This is remarkable. And profoundly powerful. What this means for us is that, if we use preattentive attributes strategically, they can help us enable our audience to see what we want them to see before they even know they're seeing it! (Hepworth 2015)

Aside from colour, we can also use movement to call the audience's attention to the point of focus in a visualization. Movement has two sub-attributes - flicker and motion. While both can be used very effectively to call someone's attention, we should use it with caution in information visualizations, as the audience may find the motion annoying or distracting from the rest of the information that is being presented. Movement is most commonly used in banner adverts and other forms of web advertising and may be a primary reason for the rise of ad-blocking software today.

Shading, or colour tints, can also be used to direct the reader's attention by increasing darkness on elements we do not want the reader to be focusing on. On the contrary, we increase the brightness/contrast on the element we want the reader to focus on.



Source: https://www.interaction-design.org/literature/article/preattentive-visual-properties-and-how-to-use-them-in-inform
Author/Copyright holder: P.Ctnt. Copyright terms and licence: Public Domain.

3.7.1.2 Why typography is not currently popular in data visualizations

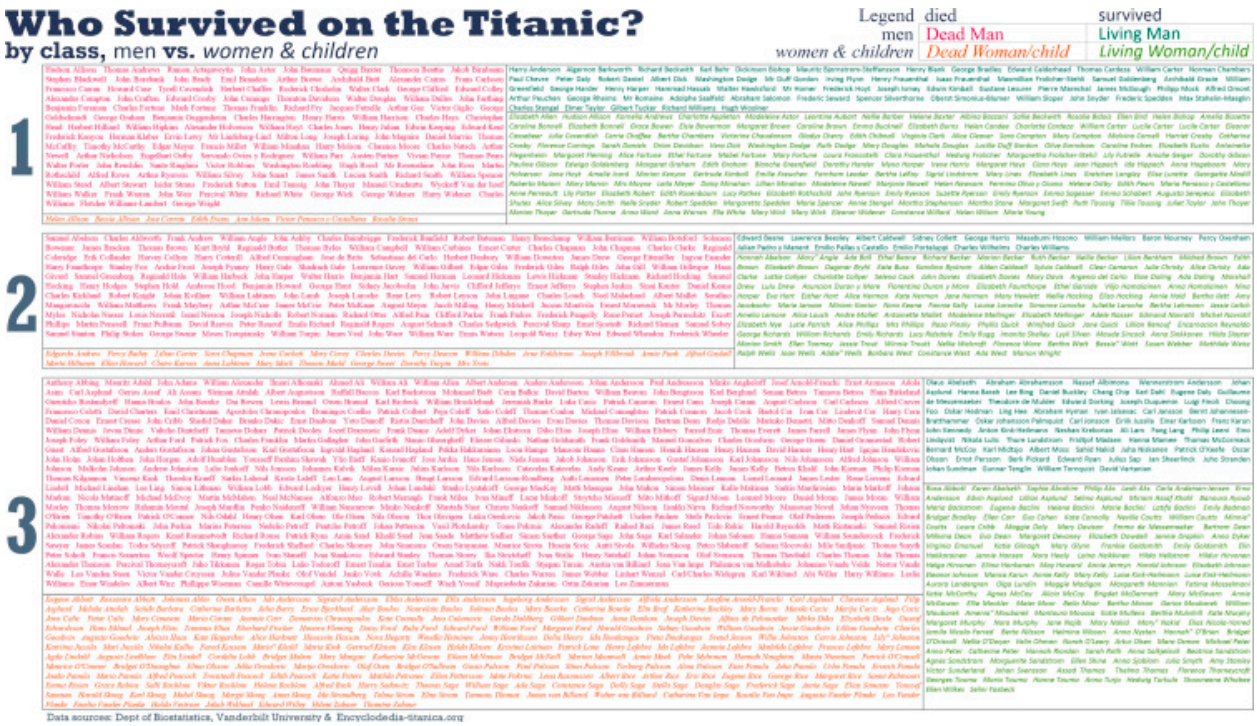
The authors remark that it is possible the lack of variety of typographic elements used in data visualizations is due to the limited knowledge of computer scientists and other individuals pursuing data visualization in how to apply these elements effectively. While the first few proposed explanations 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. 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 cartography, mathematics, chemistry, and programming have a rich history with type and font attributes that informs the scope of the parameter space?

3.7.1.3 Tips for using typographic attributes in visualizations

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 and 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 we 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.



3.7.1.4 Criticisms of typography

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.

3.7.2 Infographics vs. Data Visualizations

Data visualization and infographics both present visual information to users. While their purposes may seem similar, they have different use cases. This article explains the differences between an infographic and a data visualization. (Pritchard 2016).

3.7.2.1 Data Visualization

Data visualization usually involves the presentation of summary statistics using visual forms such as graphs, plots or charts; its goal is to provide clear and succinct information about your research. Data visualization also typically focuses on the two critical aspects of data and design. However, a design should depend on the

data itself; for example, the type of chart used in a data visualization should be selected based on which one best displays the particular data set. Since visualizations are essential in telling stories (such as trends), it should avoid adding extraneous and distracting details. Data visualizations should be self-explanatory, and users should be able to conclude on their own.

3.7.2.2 Infographics

An infographic, on the other hand, is typically a combination of illustrations, facts, and text. Infographics might include some components characteristic of data visualization but in general feature less data-driven storytelling. While infographics are not grounded in data, like data visualizations, infographics convey several ideas simultaneously; and like data visualization, the design should be both visually appealing and should base in the function of conveying the visual story.

3.7.2.3 Comparison

Data Visualization	Infographics
Illustrates raw values	Visualize stories
Delivers information	Provide stances
Offers objectivity	Offer subjectivity

3.7.2.4 When Should You Use Infographics or Data Visualization?

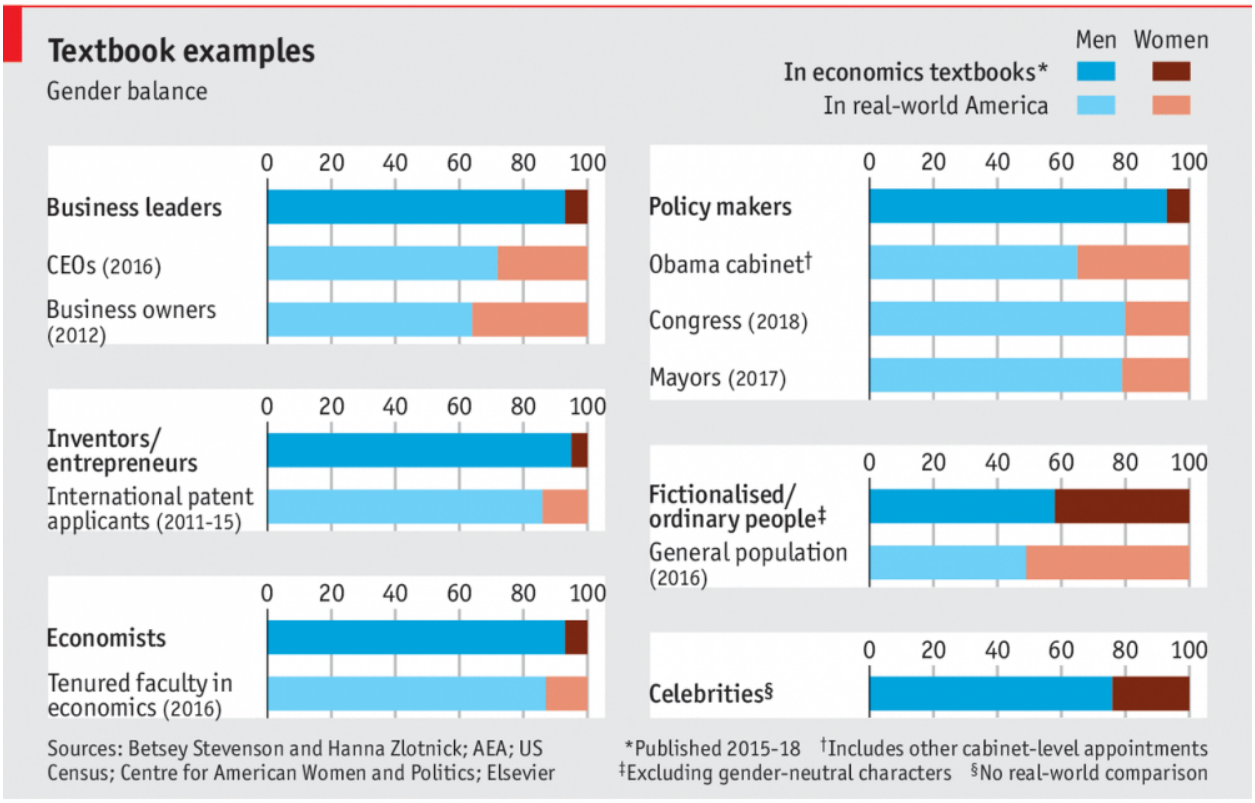
While both infographics and data visualizations have their distinct use cases, more often than not they can be used together. Some of the effective ways to choose between them are described below.

Data Visualization:

Use Case	Best Visual Representation	Rationale
Newsletters	Data Visualization	Newsletters have to catch the interest of viewers. Putting good data visualizations in newsletters makes them more interesting, and includes informative details such as a company's unique findings, statistics, or status.
White papers & eBooks	Data Visualization	Including data visualizations can help support the argument you make in the document.
Annual Reports	Data Visualization	Things like an overview of past year, success stories, and company performance can be done well using data visualization.
Blog Posts	Infographic	Blog posts are generally written for a specific purpose. Including infographics can reinforce the point you are trying to make.
Case Studies	Infographic	Paired with a case study, an infographic can provide engaging visuals and succinctly summarize a lengthy report, offering valuable insights to readers.
Marketing Content	Infographic	Marketing content generally tells a story. The best way to tell a story is using proper infographics. These can be great for social media campaigns since infographics can display all the main points.

3.7.3 Developing a Data Visualization Style

Type in 'economics chart' on your favorite web search engine and you might come across this chart:



If you're familiar with business news and publications, you might quickly recognize that this chart likely came from the famed economics publication, *The Economist*. Perhaps it was the light grey background, the short and attention-grabbing title, sparing use of colors, or even use of the tiny red bar at the top-left corner of the chart that told you this was from the *The Economist*. Search again for 'The Economist charts' and you'll notice that all of their published charts look consistent and carry an identical brand. That is because *The Economist* charts conform to a style unique to the magazine.

[This guide](Lab 2017) by the Trafford Design Lab makes a compelling argument for organizations to develop style guides when creating visualizations to ensure consistent and high quality charts are produced. Style guides might make users conform to certain typefaces (fonts) or color palettes, but can also include best practices for data visualizations.

3.7.4 Handbooks to improve your visualization design

How do we turn findings from a dense spreadsheet into something that really makes our point? Good information design is the key. There are many free handy ebooks that offer guidance. The ones listed below might not relate to data viz directly, but can guide us in designing better visualizations.

- **Design's Iron Fist** by Jarrod Drysdale (Drysdale 2016)

The free ebook, Design's Iron Fist, is a collection of Drysdale's previous work all wrapped up in one neat little package. Aside from practical tutorials and processes, this book also offers help on how to get into the mindset of being a truly great designer.

- **The Creative Aid Handbook** by Koo Roo (Roo 2013)

Creativity doesn't just happen overnight. It's something that each and every designer has to work at on a day-to-day basis. If you find that your innovative juices are running dry, The Creative Aid Handbook could be the answer. The helpful guide looks at how you can boost your intellect, foster your well-being, and, most importantly, become more creative.

- **Designbetter.co** by InVision (Invision 2018)

InVision released three fantastic design books that are available for free. Each book discusses various aspects of design like design process, management, and business. Moreover, some of the materials are available in audio format.

- **Type Classification** ("Type Classification Handbook" 2008)

Type Classification is a helpful beginner's guide to typography. It provides the foundations of typography and covers a history of each of the type forms.

3.8 Contemporary Research Results & What's Next

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 (UW Interactive Data Lab 2015):

New Approach	Description
Adapting 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.
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.
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 (Klemmer and Levis 2011). 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.

(Tufte 1986) ("Principles of Data Visualization - What We See in a Visual," n.d.)

Chapter 4

Case Studies

This chapter explores some interesting case studies of data visualizations. Critiquing these case studies is a valuable exercise that helps both expand our knowledge of possible visual representations of data as well as develop the type of critical thinking that improves our own visualizations. Furthermore, the examination and evaluation of case studies help show that new designs are just as usable as existing techniques, demonstrating that the field is suitable for future development.

4.1 Introduction

Visualization is like art; it speaks where words fail. The usefulness of data visualizations is not just limited to business and analytics; visualizations can explain almost anything in the world. Wars, rescue operations, social issues, etc. can be visualized to synthesize the details important details relevant to the issues. In particular, phenomena like the Syrian war, the number flights during Thanksgiving in the USA, the controversy of '#OscarsSoWhite,' etc. present such complexity that we can write endless paragraphs and still fail to convince readers. Below are visualizations of some of these important and complex topics - visualizations that are much more persuasive than an essay, and with a tiny fraction of the text.

Many of the case studies mentioned below come from the following articles:

Source	Description
(Nathan Yau 2015a)	This source picks the top 10 best data visualizations of 2015. For each pick, the author displays the project plot and also describes his reasoning for choosing that chart as an exemplary visualization. This article is useful for getting a basic understanding of what characteristics a good visualization should include.
(Kayla Darling 2017)	The author has chosen fifteen of the best infographics and data visualizations from 2016 and explained the reasoning behind these choices.
(Crooks 2017)	This author has chosen 16 examples of data visualization that demonstrate how to represent data in a way that's both compelling and easy to digest.
(Stadd 2015)	These 15 data visualizations show the vast range that data analysis is applicable to, from pop culture to public good. Take a look at them to get inspiration/understanding for your own work.
(Chibana 2016)	This source includes 15 data visualizations that cover current events, including politics, Oscar nominations, and immigration.

Source	Description
(Andy 2009)	Vizwiz is a blog about Tableau-based data visualization. It has case studies about how to improve visualizations, written by Andy Kriebel, a famous Tableau Zen Master. This blog is recommended because it is not only practical but also full of insights. One of the best parts of this blog is the "Makeover Monday," which develops a new visualization based on an original one. This blog also includes excellent tips for and examples of Tableau.
Viz of the Day	Tableau has a gallery that displays great data visualization examples created by Tableau. It is useful to see how people are using all kinds of data to create informative yet fun data visuals. Data being used is also attached to the example so we can try to mimic what other people did as well.

4.2 Geographic Visualizations

Geovisualization or geovisualisation (short for geographic visualization), refers to a set of tools and techniques supporting the analysis of geospatial data through the use of interactive visualization. Like the related fields of scientific visualization and information visualization geovisualization emphasizes knowledge construction over knowledge storage or information transmission.To do this, geovisualization communicates geospatial information in ways that, when combined with human understanding, allow for data exploration and decision-making processes. Source:(contributors 2019a) More specifically, Geovisualization is a process that alters geographic information so that we can consume it with our eyes. Its purpose is to capitalize on our affinity for visual things and convert the seemingly random collection of information available to us into a form that can be quickly understood. Many tools can be used for Geographic Visualization, such as Mapbox,Carto,ArcGIS Online and HERE Data Lens. Source:(Gloag, n.d.: Tools & Techniques)

Often, people use maps to visualize data that should not be mapped. Here are some examples of when a map visualization is a good choice.

4.2.1 Spies in the Skies

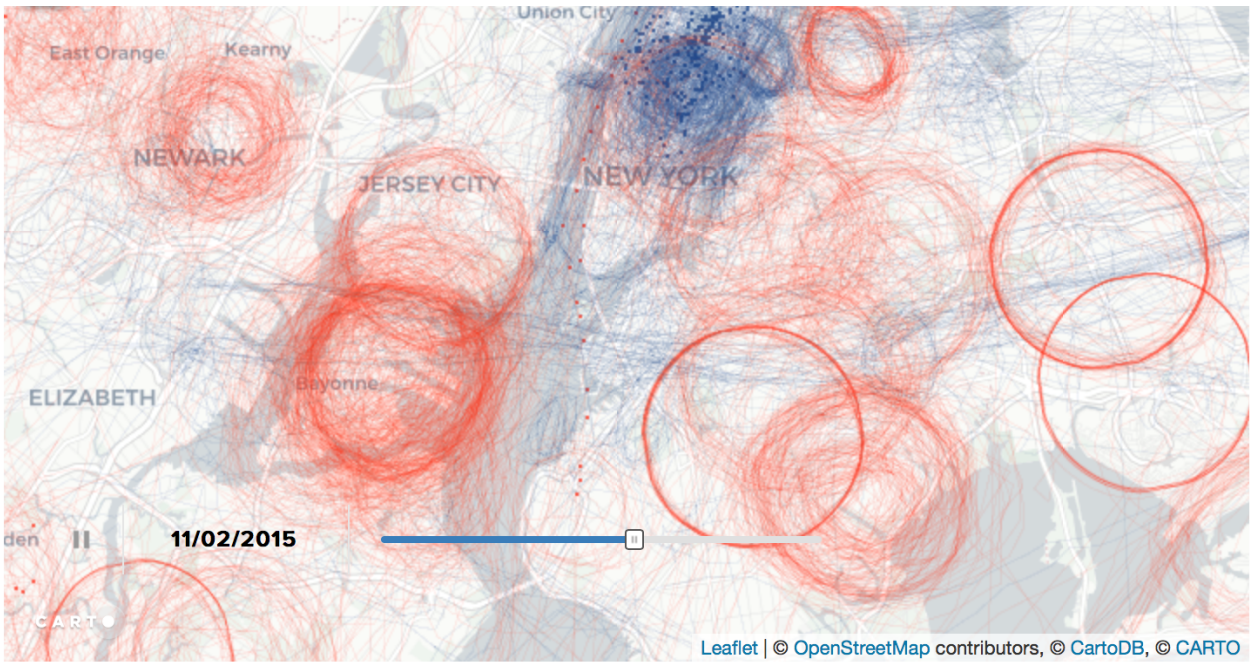
The map below is from a Buzzfeed article (Aldhous and Seife 2016) that shows how common it is for the government to observe people. It was filled with red and blue lines (representing FBI and DHS aircraft, respectively) which illustrate the flight paths of the planes. When planes circle an area more than once, the circles become darker. The circles change by day and time, and individual cities can be typed into a search bar to see the flight patterns over them. The visualization rather creatively looks almost like a hand-drawn map. While presenting an ordinarily uncomfortable topic, this allows individuals to check things for themselves, hopefully providing some peace of mind.

Source: (Kayla Darling 2017)

4.2.2 Two Centuries of U.S. Immigration

This interactive map from (Galka 2016) shows the rate of immigration into the U.S. from other countries over the last 200 years in 10-year segments. Each colored dot represents 10,000 people coming from the specified country. Countries then light up when they have one of the highest rates of migration. A tracker on the left indicates what countries sent the most people to the U.S. at what times.

This is a good visualization because it is engaging and easy to read and interpret. The movement of the dots draws the reader's attention while the brightly lit countries make it easy to pick out the highest total migrations. The bright colors and dark background help the information stand out. This map is a bit simple, but effective.



FBI **DHS** *PETER ALDHOUS / BUZZFEED NEWS*

Faint lines show flight tracks, which become dense circles when planes repeatedly monitor the same location. The animated dots show the flights of individual aircraft, over time. *Type a city into the box at top right to look at specific areas.*

Figure 4.1: New York Flight Patterns

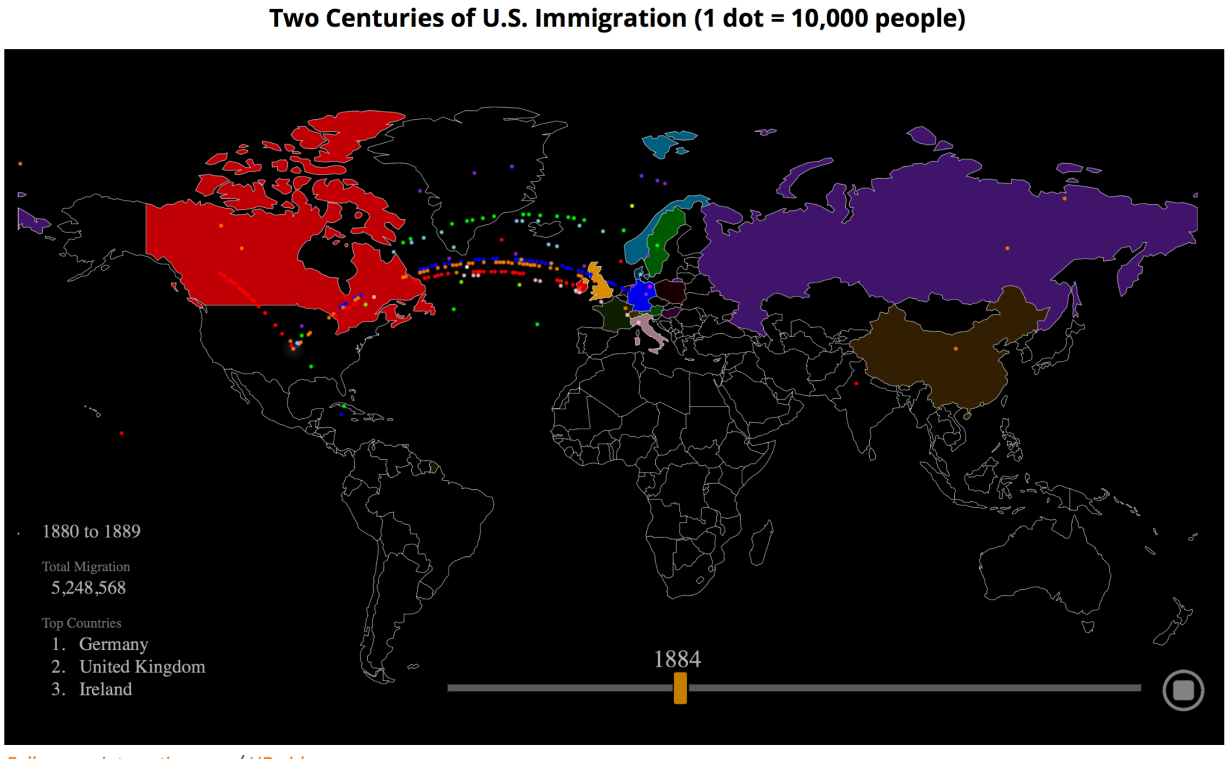


Figure 4.2: US Immigration

Source: (Kayla Darling 2017).

4.2.3 Uber: Crafting Data-Driven Maps

Map visualization is essential for companies like Uber that need to track metrics using geo-space points. In this article, the designer from Uber talks about the challenges of designing such visualizations and the possible solutions (Klimczak 2016).

The challenges that Uber faced when crafting geospatial visualizations:

There are great individual maps but as a whole lack of consistency across the company.

Common graphing tools like Sketch does not support GIS file, which is essential to Uber's insights.

The scale of the framework includes more than 400 cities in the world with a variety of different geographic features and data types.

To tackle these problems, Uber started by defining base map themes by optimizing detail, color, and typography. Based on that, data layers are added using scatter plots and hex bins, with careful color selection to help their team make decisions. To make it even better, Uber took a further step by adding trip lines (see images below), which became a signature visualization of Uber. Choropleths are also used to help visualize how metrics and values differ across geographic areas. Uber uses US postal codes as geographic boundaries and infuses various datasets to create the color variation.

The visualization in this article is a classic problem of visualizing geographic data. The detailed explanation of the problems and how they are solved can be beneficial for people or startups trying to conceptualize and make appropriate visualizations that support the decision-making process.

Source:(Klimczak 2016)

Figure 4.3: Uber Route Maps

4.3 Demographic Comparisons

One common use of visualization is to compare different groups against each other, such as political parties or generations.

4.3.1 Young Voters, Class and Turnout: How Britain Voted in 2017

This article's goal is to convey the change in party votes in the 2017 UK general election compared to votes in 2015 (Holder, Barr, and Kommenda 2017). 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 the party's standard color. The change in the 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.

This a good visualization because it is straightforward to read and interpret. The color-coding of the arrows and party names makes it easy to pick out the different parties. The index is smartly spread across the visualization to reduce cross-referencing, and color in the graph represents the actual party colors in the campaign. The arrow lengths highlight just how significant of a change happened. For example, in the Age section, it is easy to see the pattern between the Labour party gaining many voters aged 18 to 44 and the Conservative party gaining voters aged 45 and up.

Source: (Holder, Barr, and Kommenda 2017)

4.3.2 U.S. Migration Patterns

The New York Times data team mapped out Americans' moving patterns from 1900 to present, and the results are fascinating to interact with (Aisch, Gebeloff, and Quealy 2014). We can see where people living in each state were born, and where people are moving to and from. The groupings of the destinations vary based on that state's trends, preventing unnecessary clutter while still showing detail when vital, as can

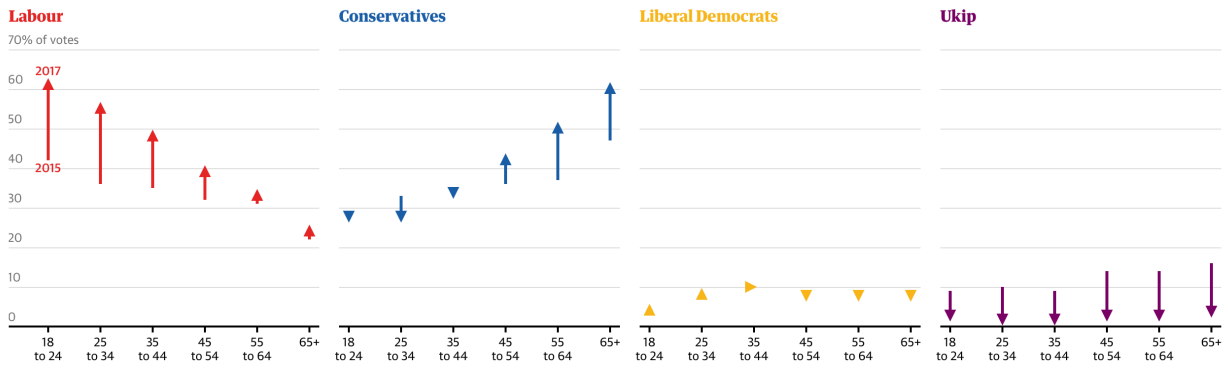


Figure 4.4: UK Party Votes by Age

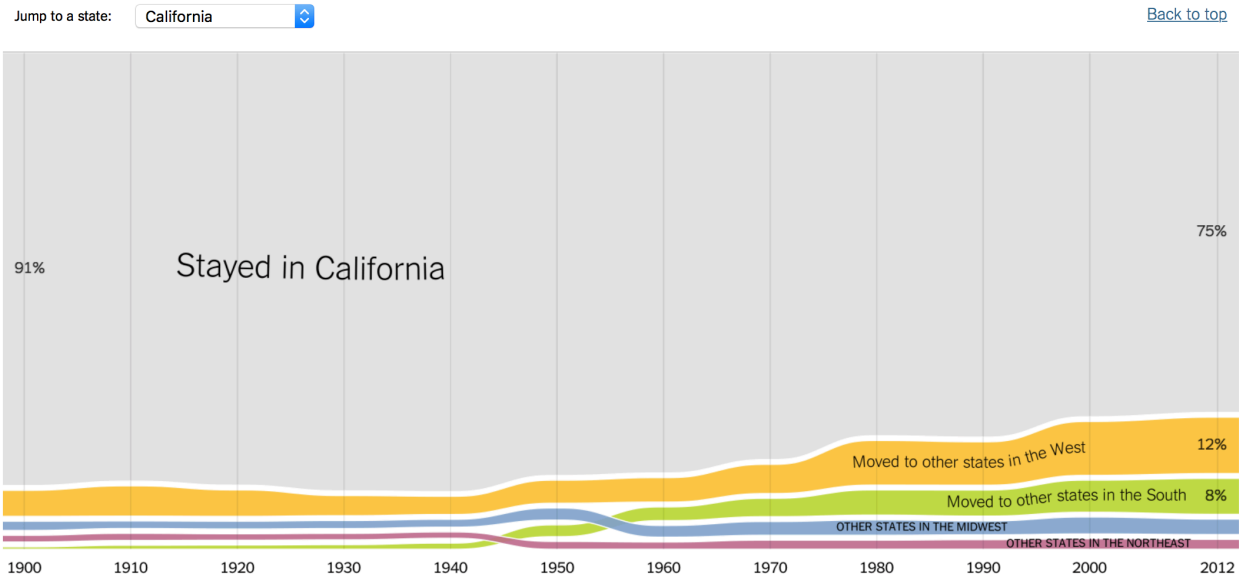


Figure 4.5: Migration from California

be seen by the difference between the charts for California and Pennsylvania. When generating interactive charts, one must always assume that the audience will not interact with it. The message of a chart has to be clear enough that anyone just viewing the generic chart can understand.

Overall, this type of chart can work well to visualize movement in data over time, such as with migration. However, it must be done carefully to maintain clarity. Too many categories with colors and crossing lines can make it difficult for a reader to keep track of what the data is saying and it can quickly go from a very graphic visualization to a chaotic mess of lines. The designer does a pretty good job with these visualizations by limiting the number of categories in grouping states by region (West, South, Midwest, etc.). But when introducing many dimenional variables such as Migration from Pennsylvania, the chart can quickly turn convoluted and hard to read which costs the audience. Finally, it is not completely clear why so many crossing lines are necessary for the Pennsylvania chart. The crossing lines, along with the use of the same color for different lines within the same regional categories, can introduce unnecessary complexity.

Source:(Aisch, Gebeloff, and Quealy 2014)

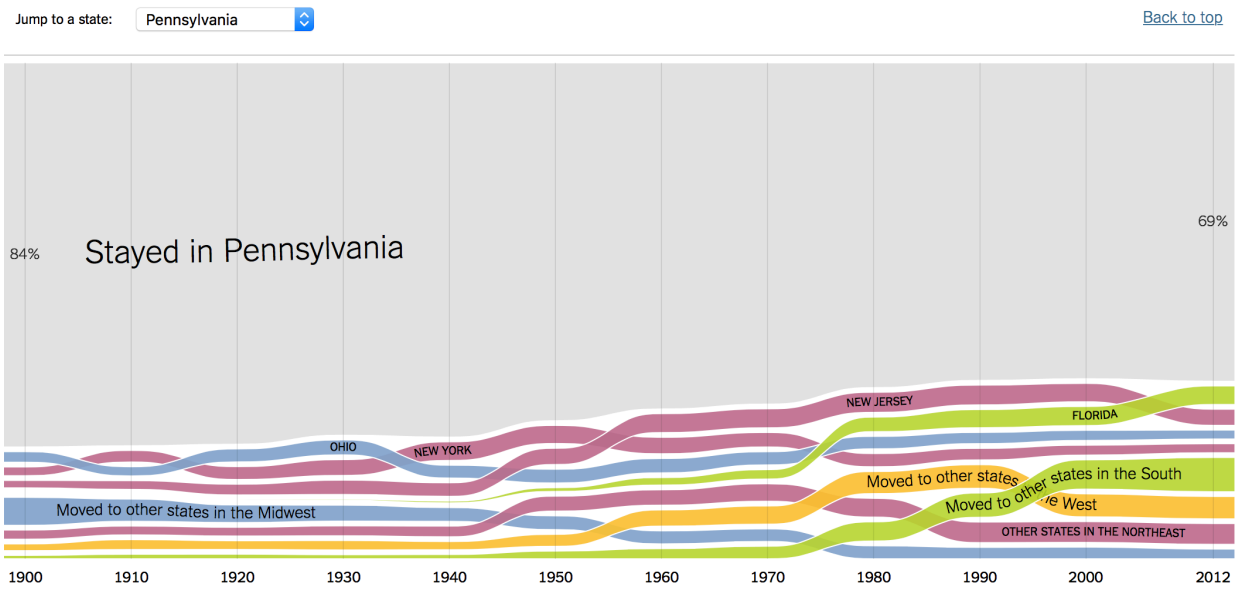
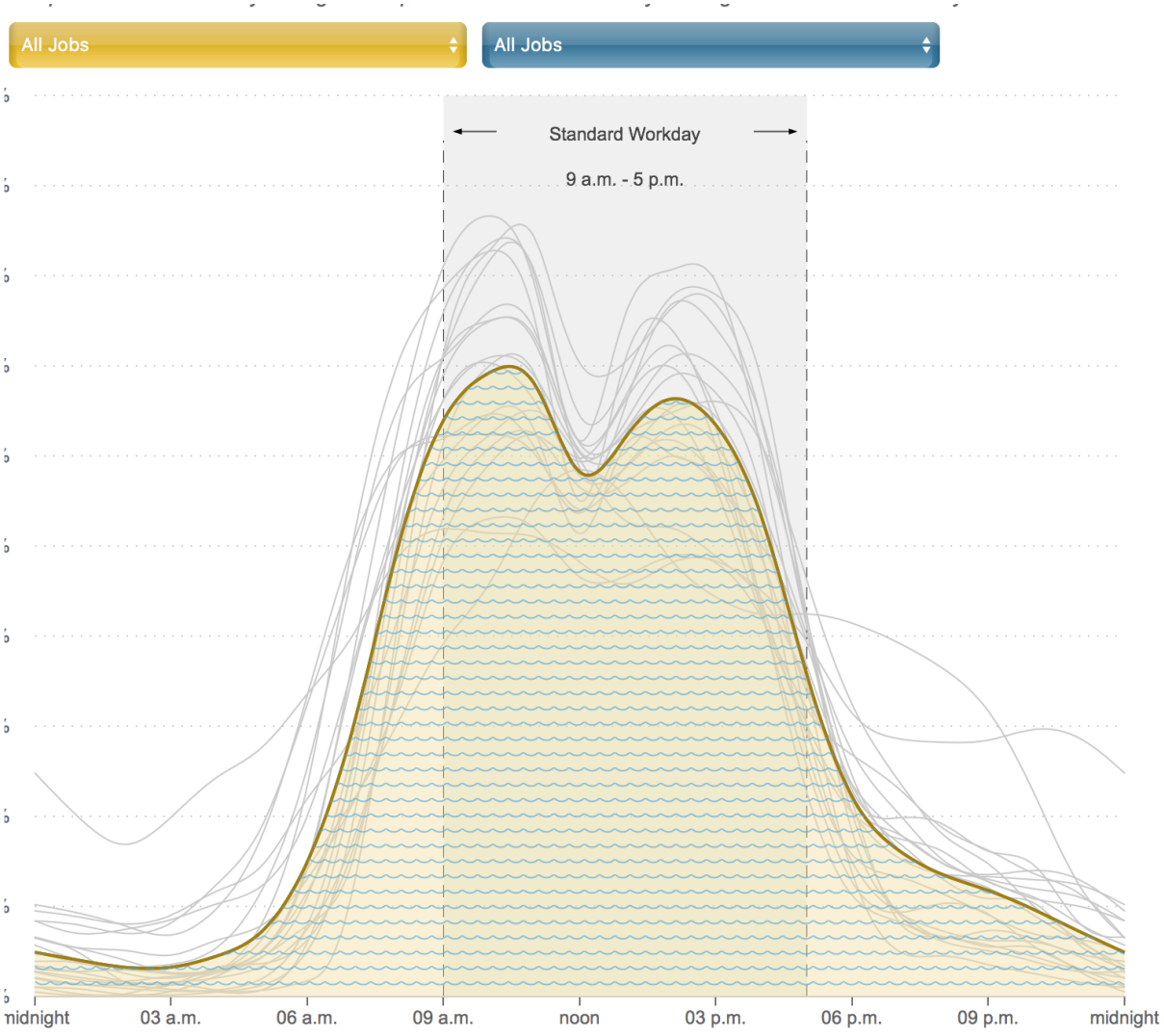


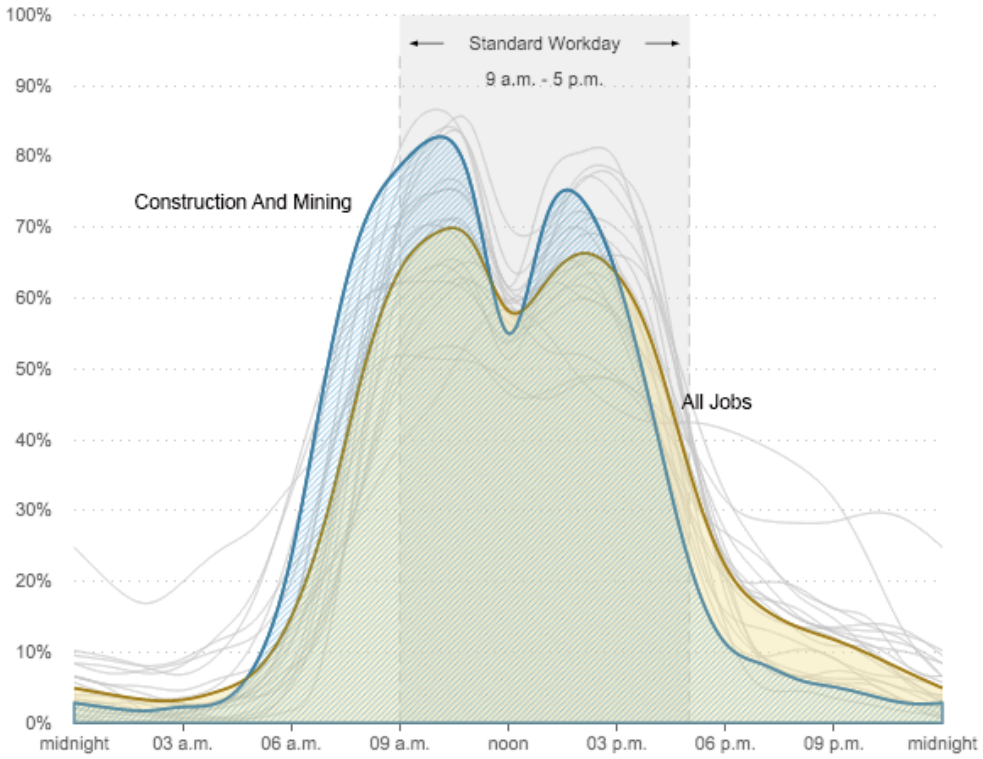
Figure 4.6: Migration from Pennsylvania

4.3.3 The American Workday

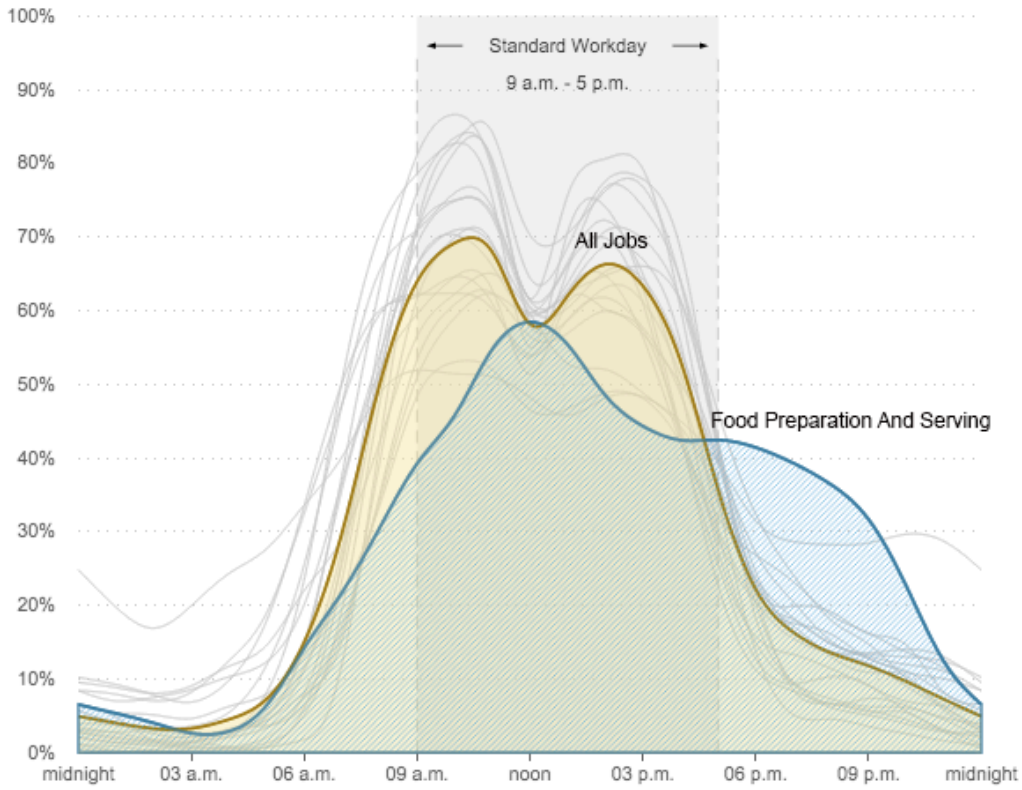
NPR tapped into American Time Use Survey data to ascertain the share of workers in a wide range of industries who are at work at any given time (Quoctrung Bui 2014). The original question of when Americans work, rather than the number of hours worked, is answered in the graph. The chart overlays the traditional 9 AM-5 PM standard workday as a reference point, helping the audience draw exciting conclusions. Below is a screenshot of the data product; the original graph is more interactive and allows the audience to explore when people are working for different occupations.



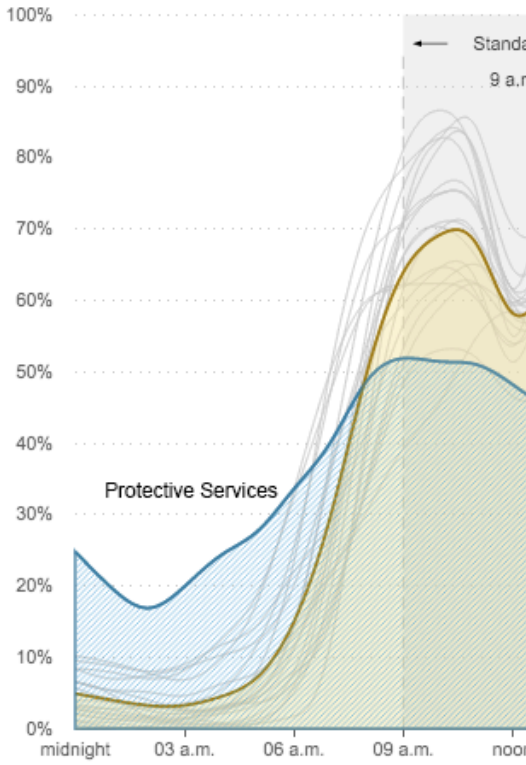
Some interesting findings include: 1. Construction workers both start and finish their workday earlier and generally do not work at lunch hours as there is a massive drop at noon.



2. Servers and cooks' schedule are the opposite of all other occupations with the peak from lunch through the evening.



3. Protective services, e.g., police officers, firefighters, and detectives, have many workers working through-



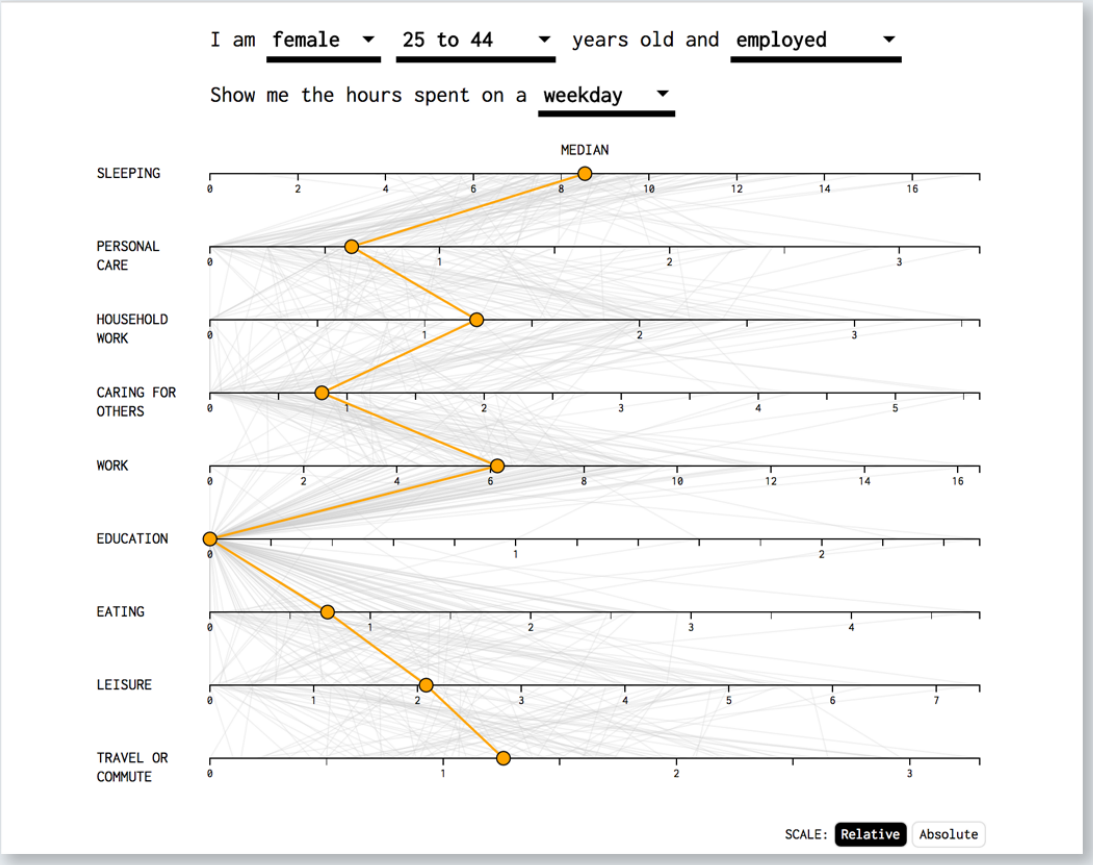
out the night, which is entirely different from all other occupations.

This data product is an excellent example because the analytic design has been applied to contrast specific occupations to the traditional 9-5 working hours. This is easy to understand and make particular occupations stand out more manageable. The use of color for highlighting the selected occupation in the graph helps to categorize different occupations as well.

4.3.4 How People Like You Spend Their Time

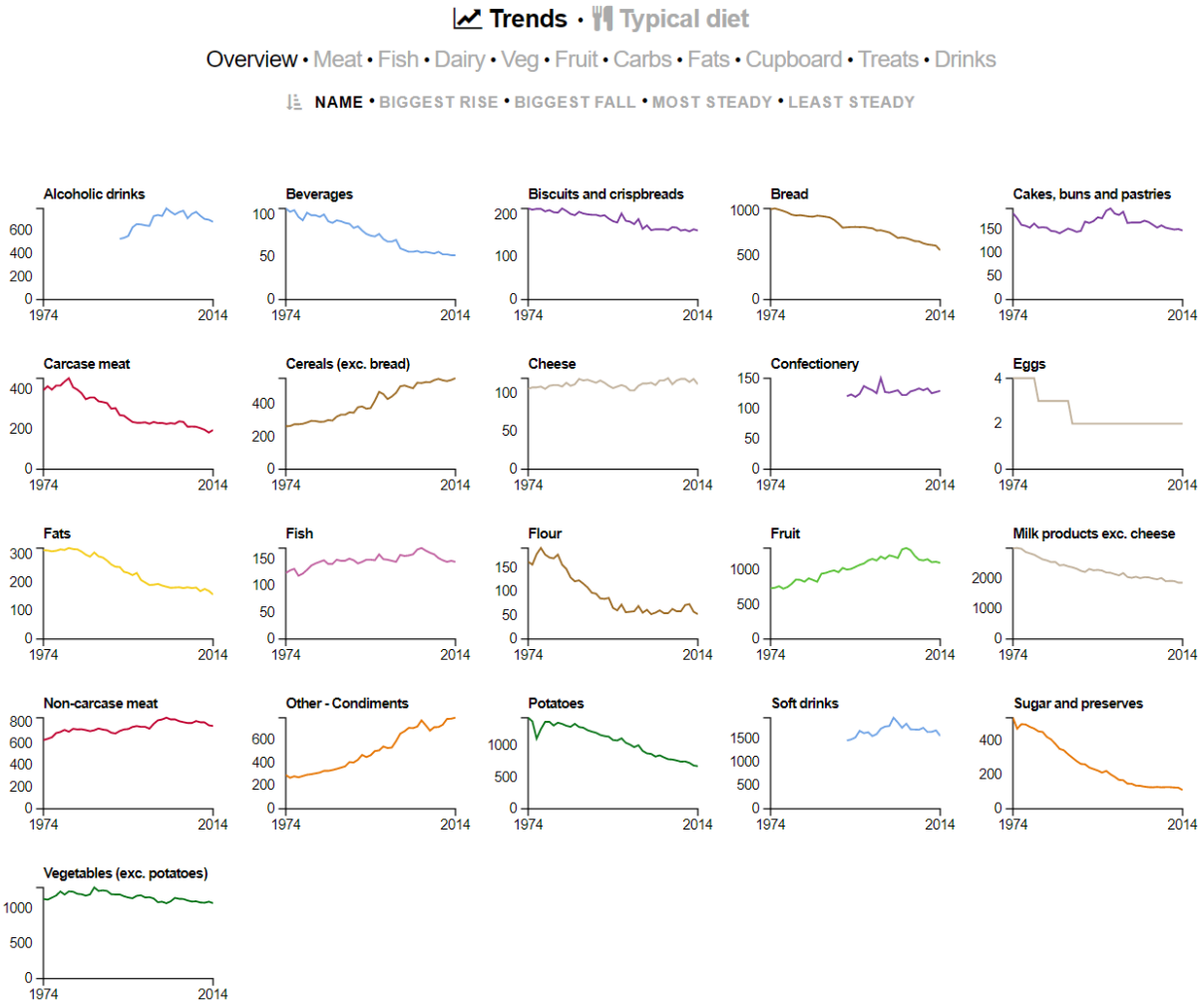
This visualization from (Yau 2016) lists several categories such as “personal care” and “work” along one side of a graph with a line illustrating the amount of time the average person in a particular demographic spends on each subject. Entering different parameters at the top, such as changing gender or age, causes the lines to shift to feature that demographic. The simplicity of this visualization helps the information get across and avoids bogging down the statistics. Sometimes, less is more.

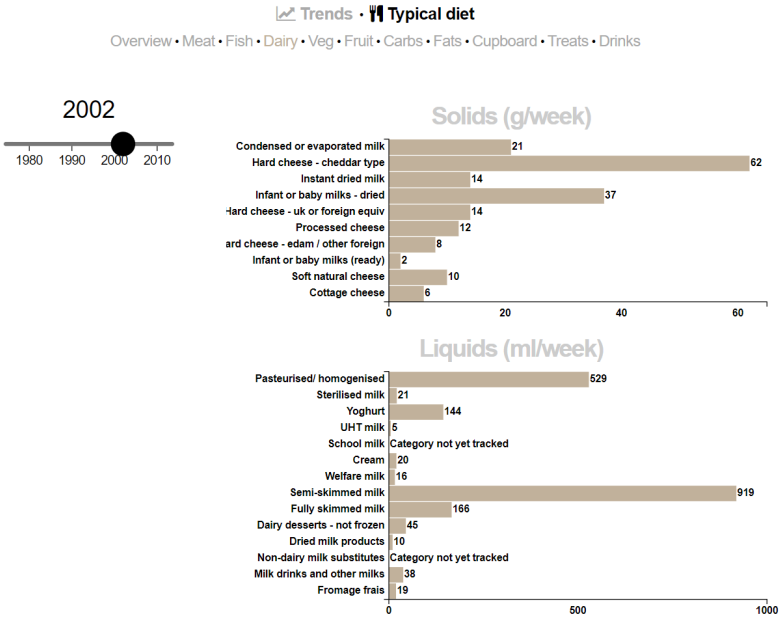
Source: (Kayla Darling 2017)



4.3.5 Britain's Diet In Data

This is an excellent example about how to present a significant amount of comprehensive data - distributed across different categories and measured in different metrics - in a simple yet effective manner, while still maintaining interest and aesthetics. The data product attempts to show how the average Briton's diet has changed over the last four decades for the better (Institute 2016). It does this by displaying simple trend lines that show that more harmful and fatty foods are being consumed less while consumed more healthier and leaner foods. It further breaks down every major food category into tens of its constituent products, and in both the overview and deep-dive versions, provides further levers to massage more meaning out of the data. It also shows how the contribution of different foods to the typical diet has changed over the years. Here, we can toggle the year to see exactly how much of each food was consumed, again with another deep-dive into the constituents of every primary food group.





Source: (Institute 2016)

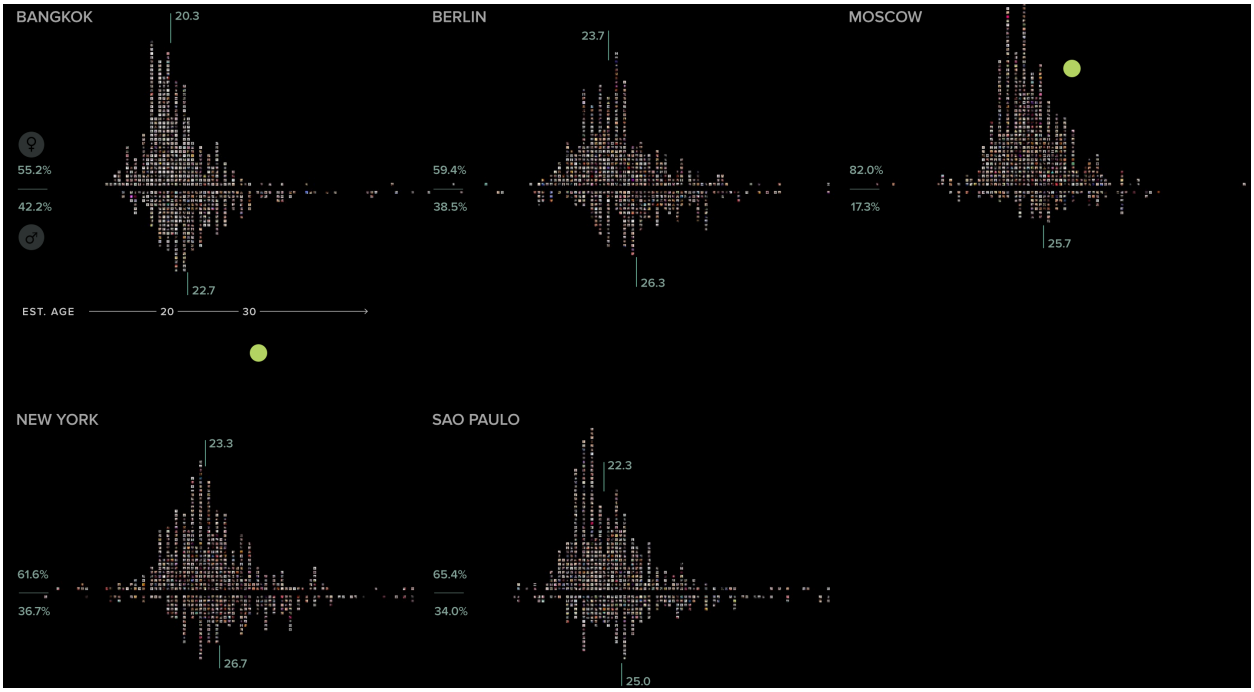


Figure 4.7: Estimated Age and Gender Distribution

Such a visualization is ideal for a layman who would want to walk away with an immediate and accurate understanding of the overall dietary changes. It also provides plenty detail on demand for the more discerning viewer who might have more time and inclination to dissect and parse through the graphs. It is difficult to use the same data product to cater to both types of viewers in such an adequate capacity, which is what makes this particular data product so impressive and useful. It satisfies the principles of graphical excellence as stated by Edward Tufte : >"Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

Source: (Tufte 1986)

4.3.6 Selfie City

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 (Manovich et al. 2014). This project is based on a unique dataset compiled by analyzing tens of thousands of images from each city, both through automatic image analysis and human judgment. The team behind the project collected and filtered the data using Instagram and Mechanical Turk. Rich media visualizations (imageplots) assemble thousands of photos to reveal interesting patterns. It provides a demographic and regional comparison of selfies.

Source: (Manovich et al. 2014)

4.3.7 Evolving Demographics

Another frequent use is to look at how something changes over time. Time-series data can be shown many ways, and these are some examples.

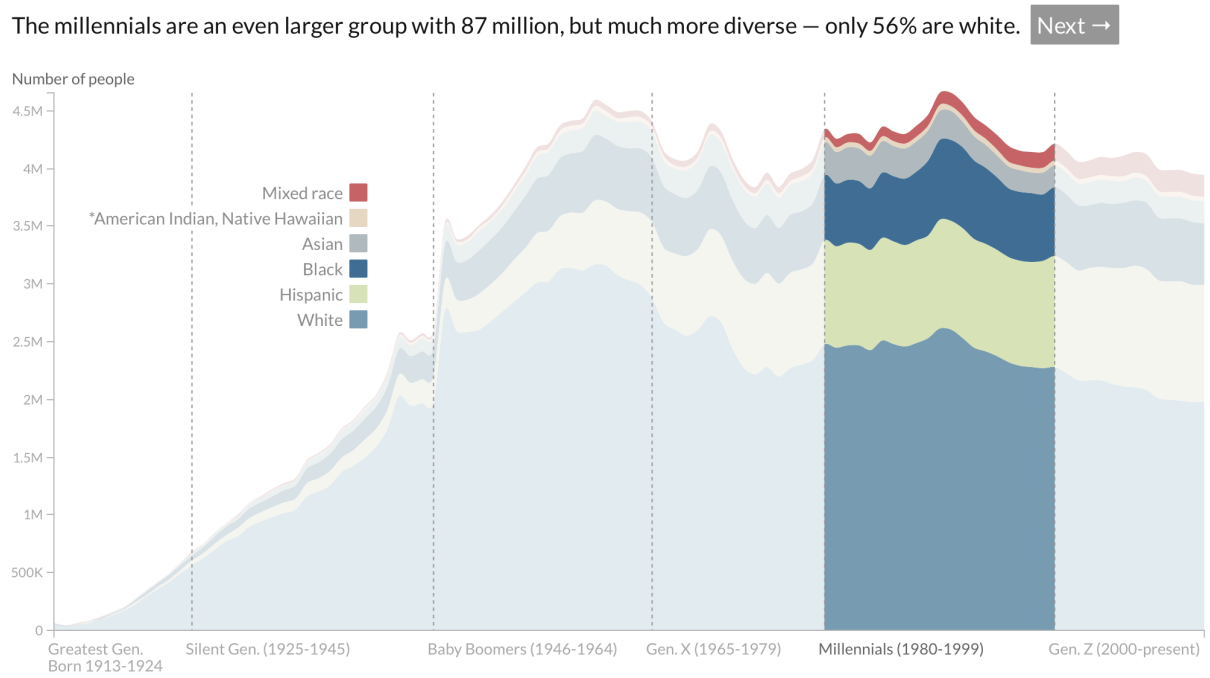


Figure 4.8: Racial Diversity of US Generations

4.3.7.1 Millennial Generation Diversity

CNNMoney created an interactive chart using U.S. Census Data to show the size and diversity of the millennial generation compared to baby boomers (Kurtz and Yellin 2018). While the article's main point is that the millennial generation is bigger and more diverse than the baby boomer generation, it also contains information about all of the other living generations. It turns hard numbers into an intriguing story, illustrating the racial makeup of different age groups from 1913 to present.

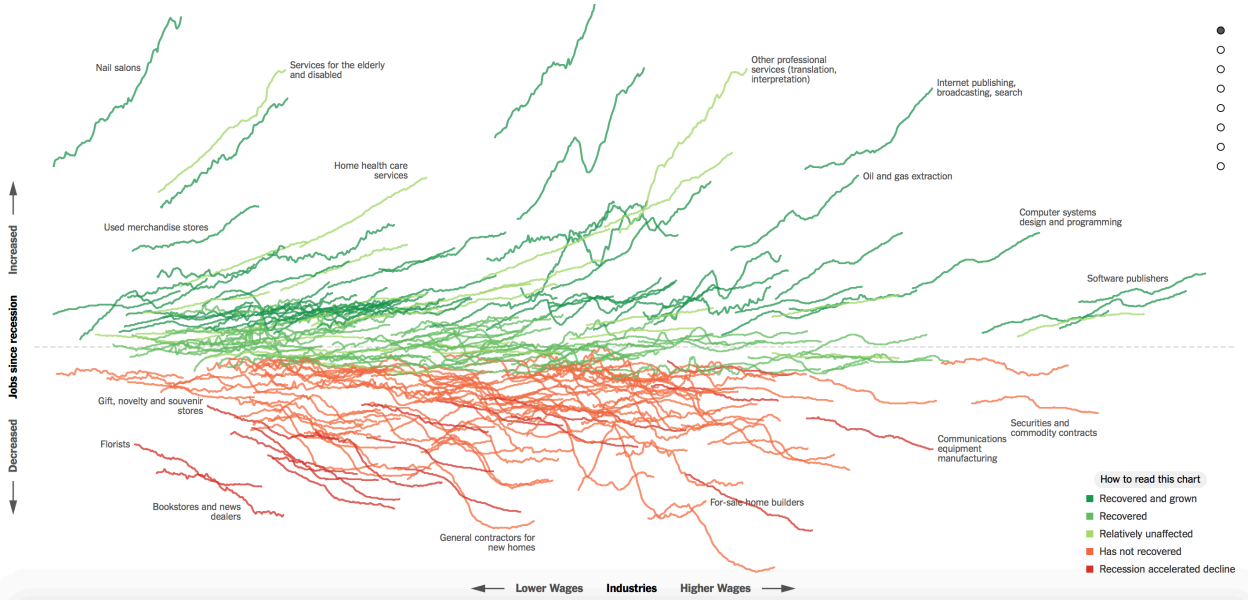
The author also summarized three key findings from the graph: 1| The most common age in the US is 22 years old. 2| The median age in the US is 37.6 years old. * 3| Among the youngest generation, only 50% of the population is white with the potential of dropping from the biggest race in the US.

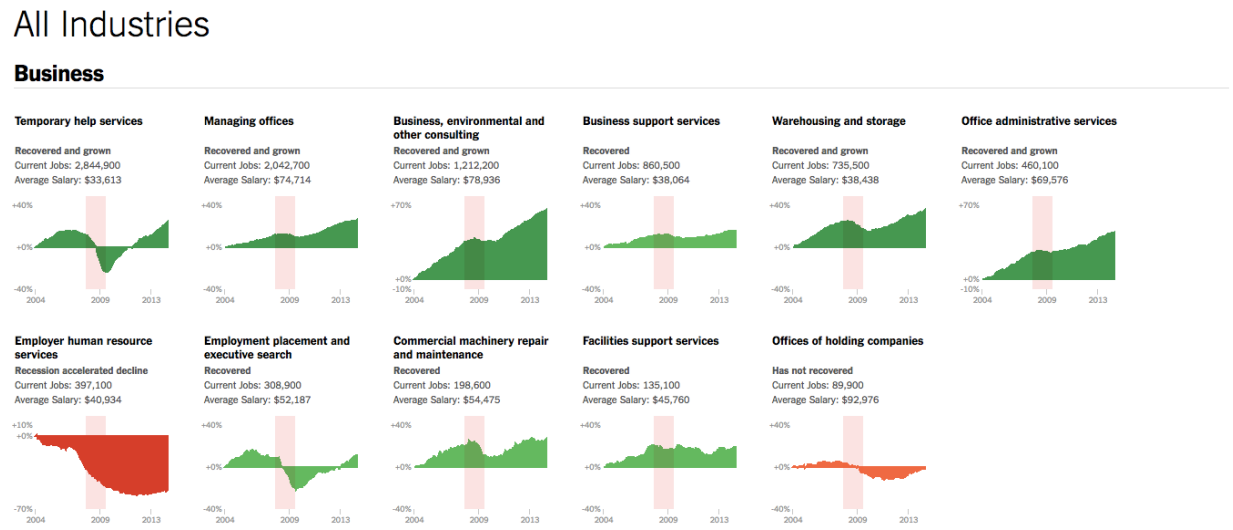
Source:(Kurtz and Yellin 2018)

This is an effective graph because while it contains many data points, it makes the overall trends very clear without sacrificing much detail. You can see the drop in some white people and the increasing growth of the other racial categories.

4.3.7.2 How the Recession Reshaped the Economy, in 255 Charts

The first large graph contains 255 lines to show how the number of jobs has changed for every industry in America, using color to highlight the lines and let viewers see the specifics for each industry (Ashkenas and Parlapiano 2014). By hovering over a line, viewers can get the detailed information of that industry's job trend. Keeping this extra data hidden until needd will make it easier for readers to absorb the bigger picture from this vast data visualization.
Following charts are subsets categorized by job sector and sub-industries. Readers can choose the industry or sector they are interested in and, similar to the first graph, view the more detailed information by hovering over a line.

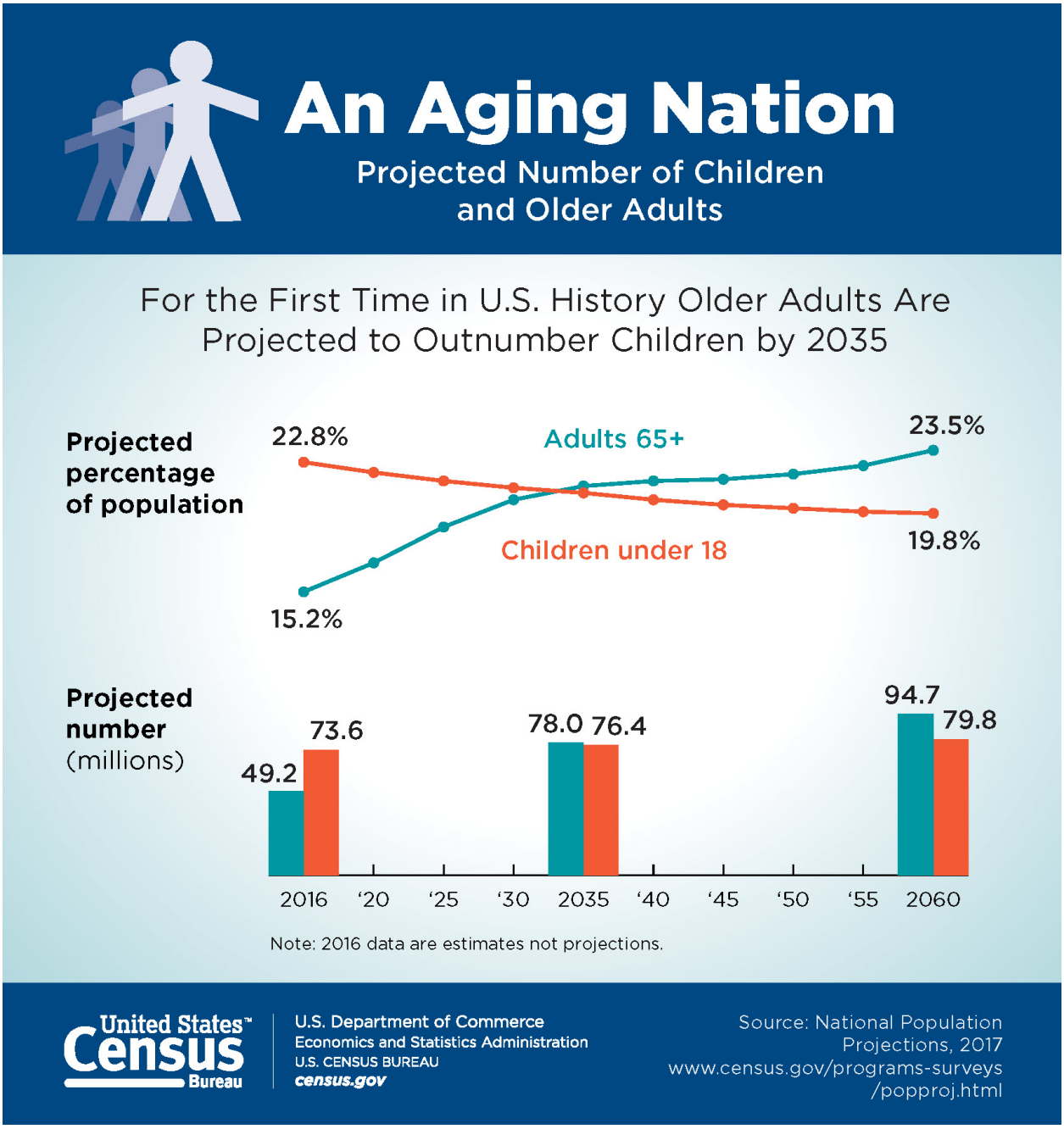




Source:(Ashkenas and Parlapiano 2014)

4.3.7.3 An Aging Population: Projected Number of Children and Older Adults

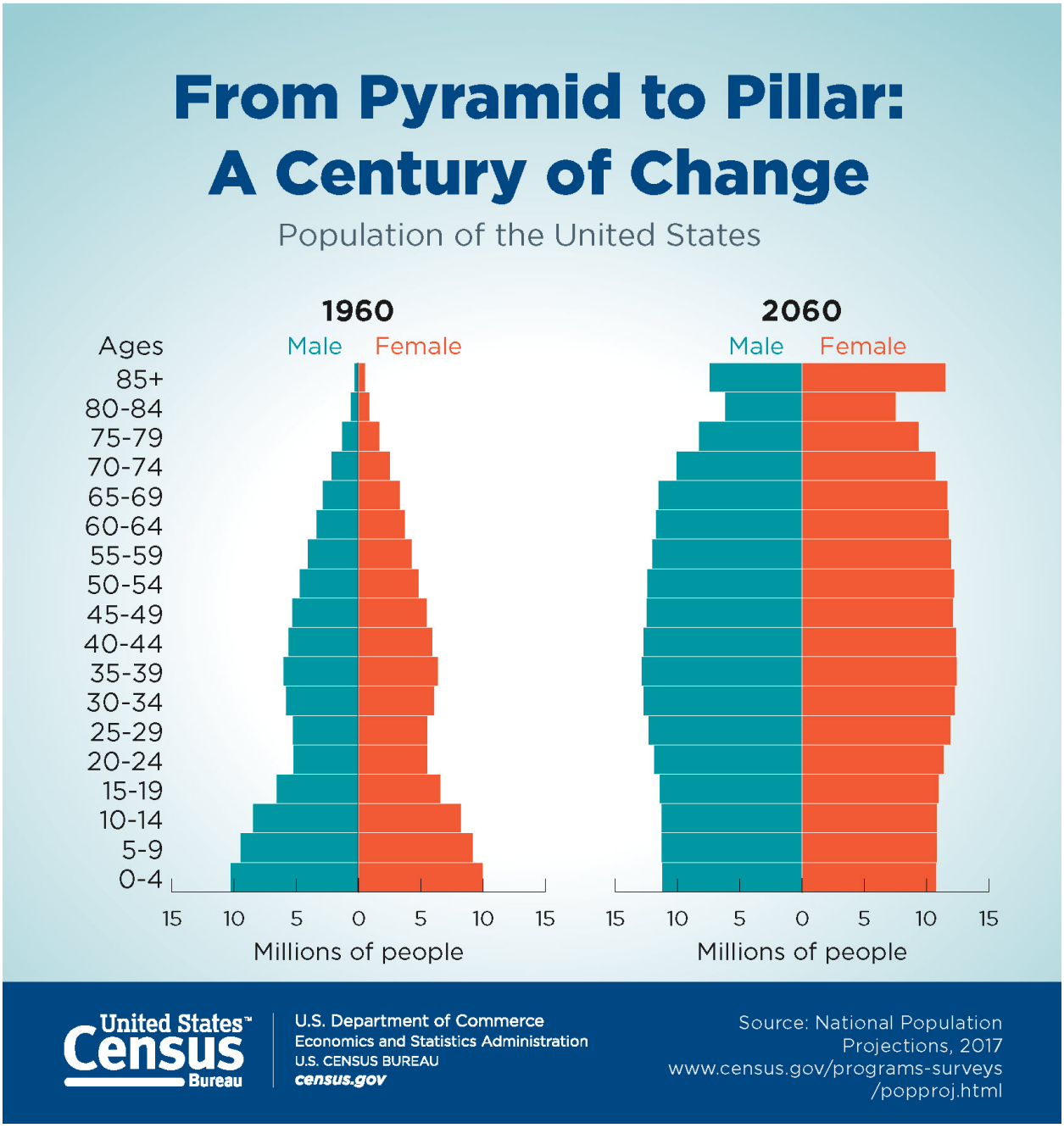
An aging population is always a hot topic in social economics and politics (United States Census Bureau 2018). Here we explore a collection of data visualizations showing the aging population in the U.S. and the world.



Source:(United States Census Bureau 2018)

This example includes a bar chart and a line graph to demonstrate the aging population compared with the population of children. This visualization allows easy comparison, employs color to differentiate the categories, and highlights the intersection point.

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



This is a **population pyramid**. "A **population pyramid** is a pair of back-to-back histograms for each sex that displays the distribution of a population in all age groups and in gender" (Bureau 2018b). It is good to visualize changes in population distributions (sex, age, year). The shape of a pyramid is also used to represent other characteristics of 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 to make sense of census data. ("An Aging Population," n.d.) offers an animated pyramid.

Source:("An Aging Population," n.d.)

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

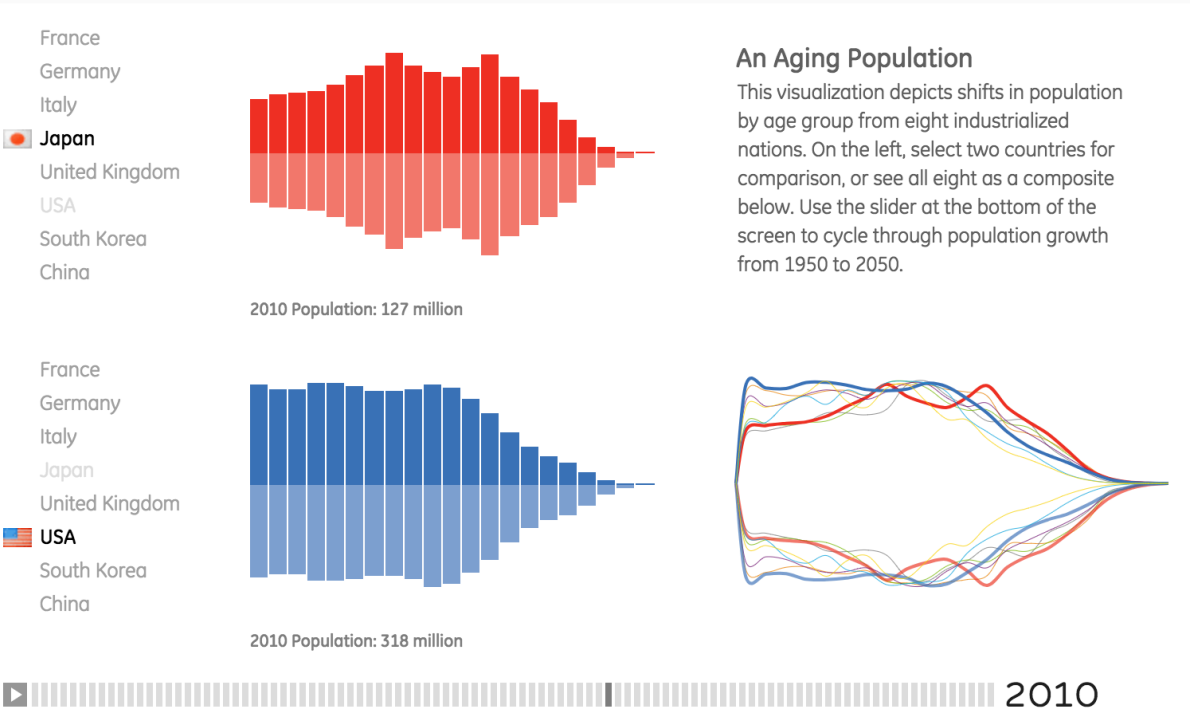
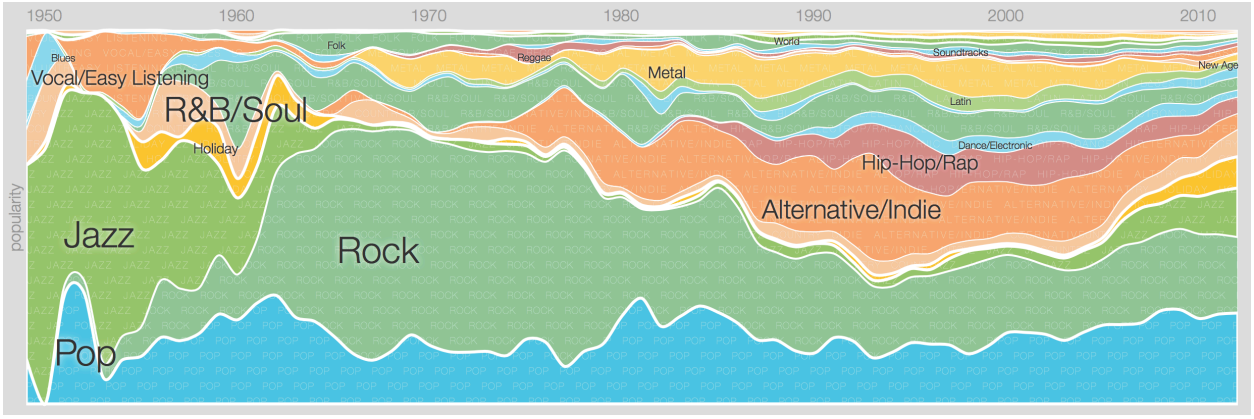


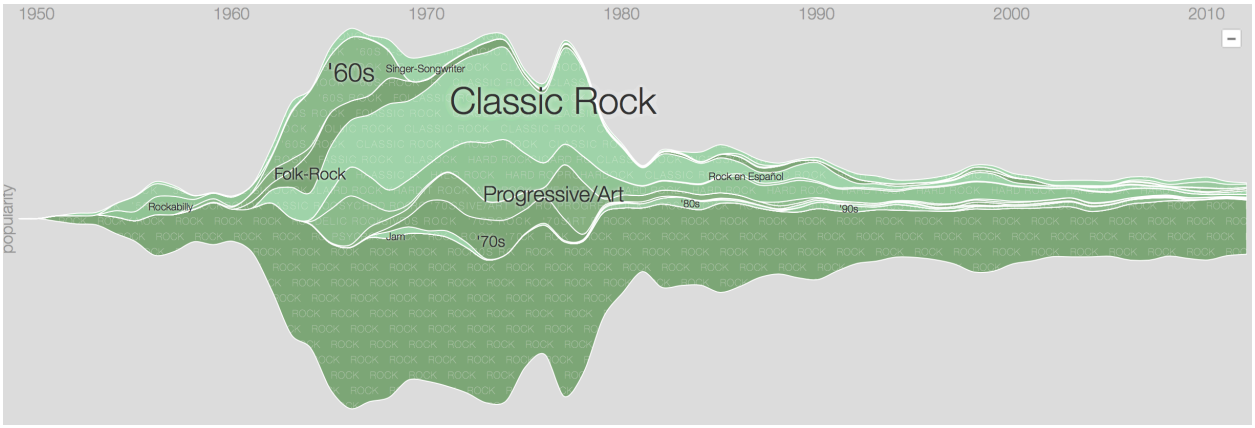
Figure 4.9: Comparison of aging population in US and Japan

which is useful for time-series comparison. A similar project with R code is here.

4.3.7.5 Music Timeline

Google's Music Timeline illustrates a variety of music genres waxing and waning in popularity from 2010 to the 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 (Google 2014). One useful feature of this graph is the reader's ability to explore one specific genre and its subgenres at a more detailed level, as well as view the general timeline of all music. The drill-down interaction allows for more details without cluttering the overview of the visualization. Embedding the graph with names (e.g., Rock/Pop) makes similar color lines easy to distinguish.





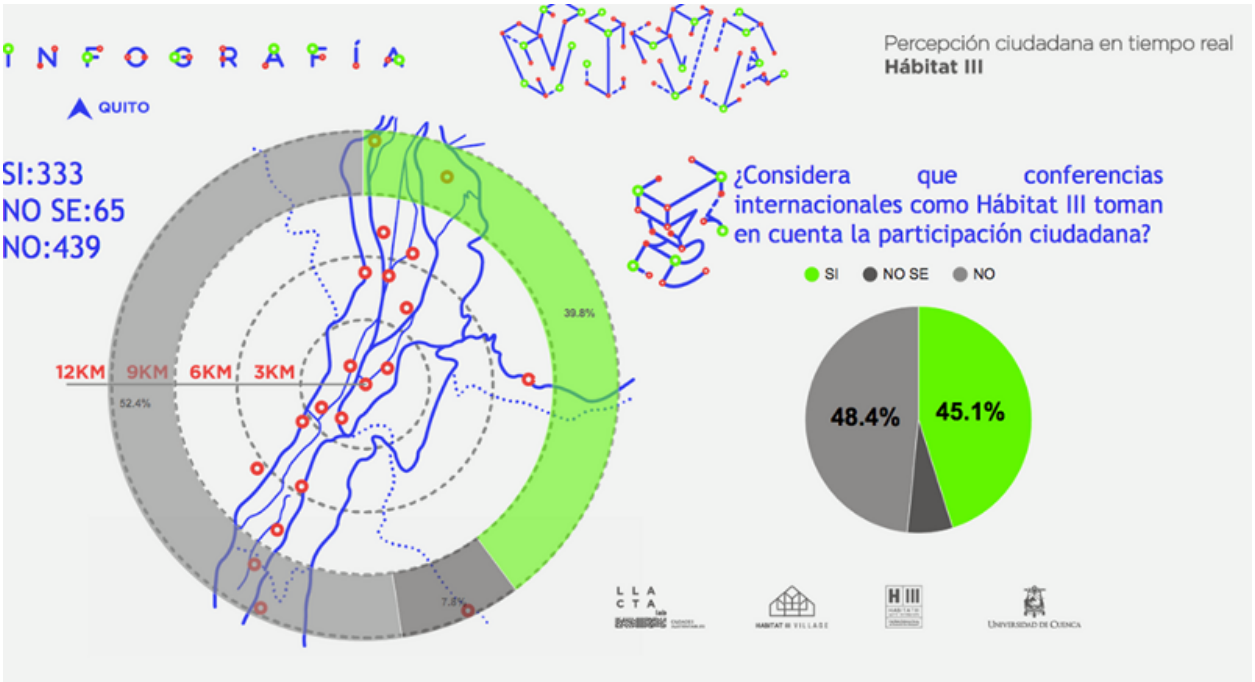
Source:(Google 2014)

4.4 Visualizing Urban Data for Social Change

(Neira 2016)

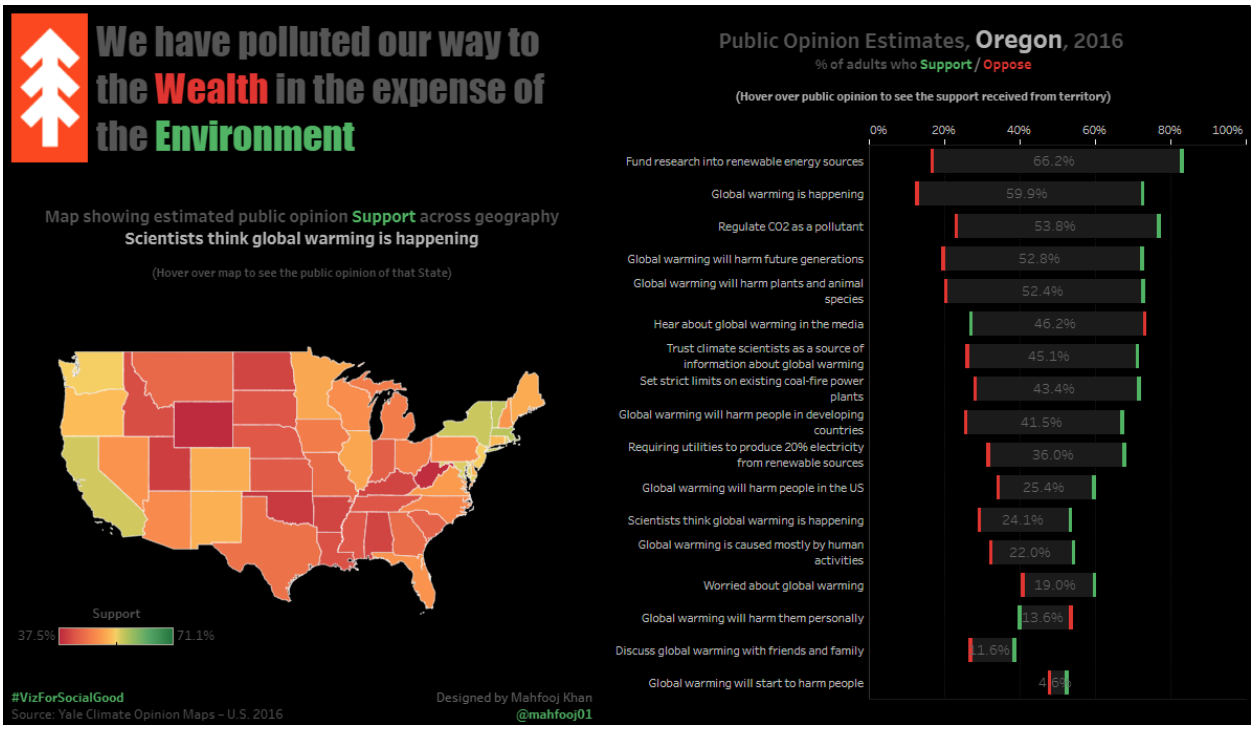
One field in which visualization can have a meaningful social impact is promoting understanding of and generating discussions around cities. With the development of a city, demographic changes, economic, environmental and social problems become important issues. Visualization plays an important role in promoting understanding of how the cities and the societies within them work, debating the problems that cities face, and engaging citizens to work toward their dream cities.

Recently, as part of Habitat III side event , LlactaLAB - Sustainable Cities Research Group, presented a project called Live Infographics. It was an interactive methodology that put citizens and experts opinions about the New Urban Agenda on one platform to help generate a 'horizontal governance'. The different opinions were materialized with a dynamic map to visualize the generated data. The primary objective of the project is to generate citizen-led data collection and to enable governments to build a better understanding of public sentiment, and then engaging people in the process.



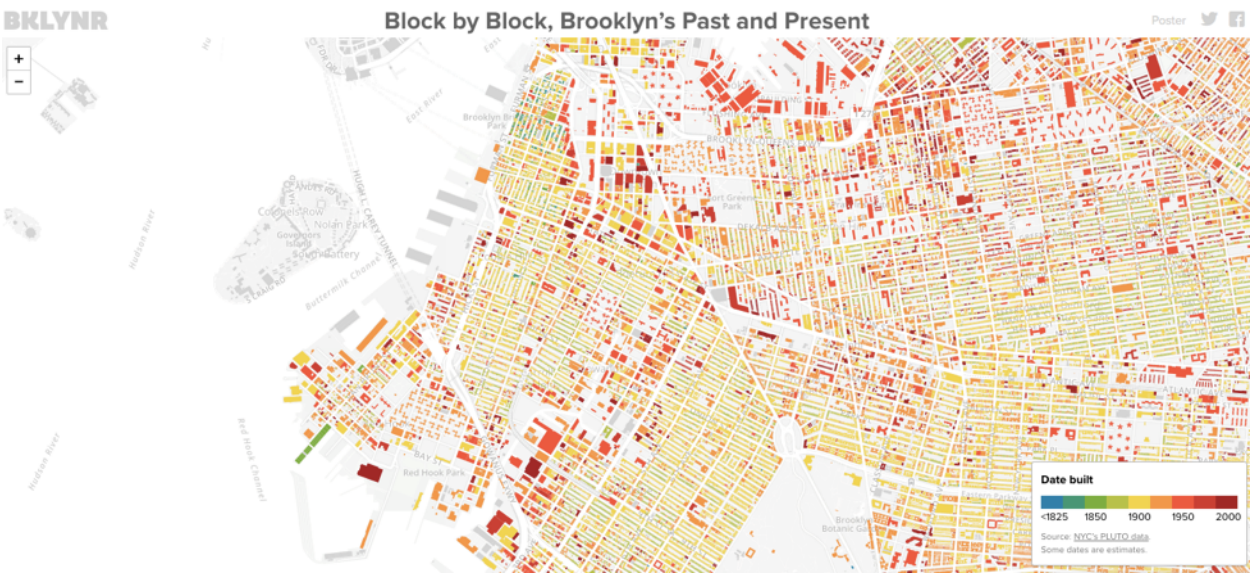
A great Urban Data Visualization ought to have the capacity to start "Sociological Imagination". It should provoke individuals to consider how their individual choices, issues, struggles, and in general their daily lives, are a extension of society, and how their choices collectively influence public opinion. Another key aspect of these kinds of data visualizations is their ability to make the audience understand how their activities impacts the cities they live in and help them work towards the betterment of the cities.

The following is an example of a visualization that is trying to effect social change. It shows how different states are populated on our way to wealth at the cost of the Environment and the percentage of adults who support the cause by estimating public opinions. Source : ("We Have Poluted Our Way to the Wealth in the Expense of the Environment," n.d.)



Urbanization and the spread of information technologies transform Cities into huge data pools, that data will play a major role in understanding how city areas have changed and are likely to change in the future. Urban Data Visualization gives us a quick view of the architectural contrast of Urban changes in Cities. (MORPHOCODE 2019)

This Urban Data Visualization based on the NYC Department of City Planning Data set, the result is a snapshot of Brooklyn's evolution, revealing how development has rippled across certain neighborhoods while leaving some pockets unchanged for decades, even centuries. The visualization is interactive, the reader can check every block's name and built year.(MORPHOCODE 2019)



As urban areas continue to develop, diverse and complex issues evolve along with them. Disparity, isolation, loss of biodiversity and environmental quality, etc. are all important but thorny issues, and finding successful solutions will require uniting strategy producers, academics, designers, and citizens. Visualization, if done right, can help jumpstart important discussions between these diverse groups of people and help solve the issues that emerge as the world becomes more urbanized.

4.5 Animated Data Visualization

Like evolving demographics, these visualizations are demographics that change over time. These, however, are self-animated instead of interactive.

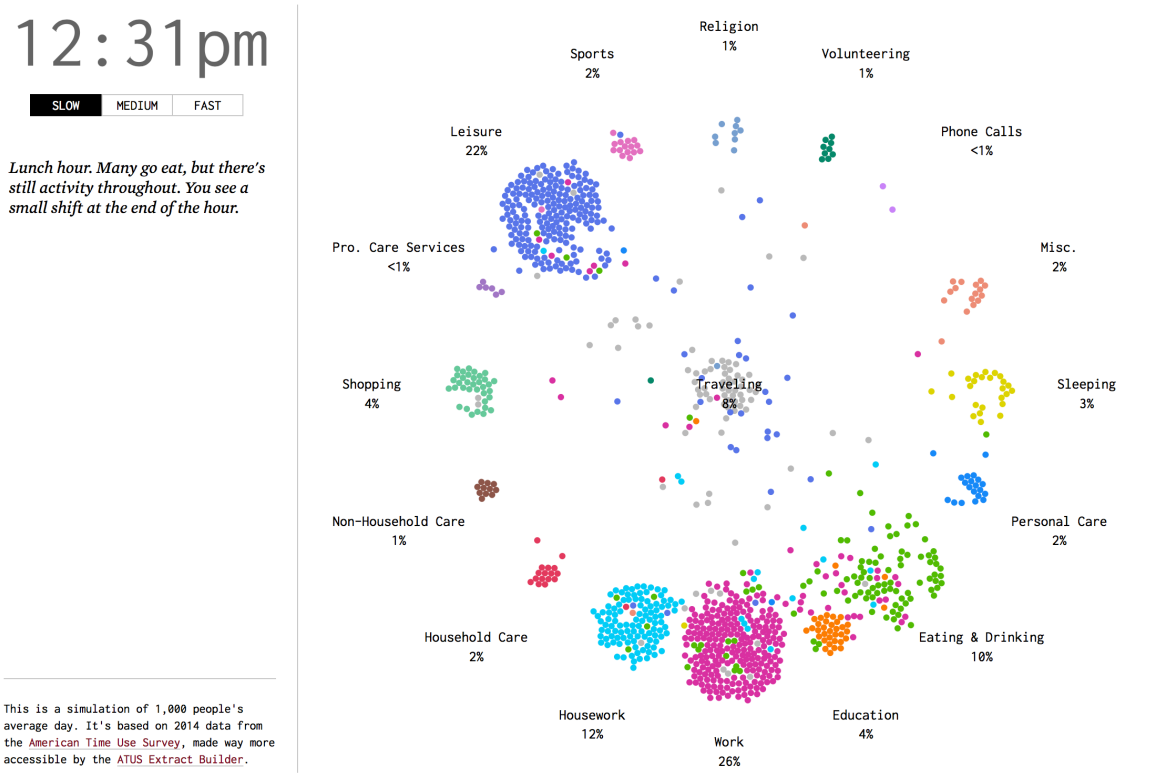
4.5.1 A Day in the Life of Americans

This animated data visualization shows the time people spend on daily activities throughout the day (Nathan Yau 2015b). 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 drill down to the level of an individual and follow him or her throughout the day. The accumulation of dots for each particular activity also gives us an aggregate-level view of the same data, so that we get both 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 crucial times of the day. This represents people's movements from one activity to another without overwhelming the reader.

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



Source:(Nathan Yau 2015b)

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

Global health data expert Hans Rosling's famous statistical documentary "The Joy of Stats" aired on BBC in 2010, but it is still turning heads. In the remarkable segment "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 (Rosling, Hans 2010).

Source:(Rosling, Hans 2010)

What makes this visualization so well-known is its use of animation and narration to highlight different stories within the overall data. While the visualization could have been made as an interactive chart where the audience can select the year, instead it is a video. Rosling's narration of how various regions have fluctuated over the last two hundred years is necessary for his argument since there is no other description or explanation.

4.6 Dust in the Wind: Visualization and Environmental Problems

Environmental issues can quickly become extremely complex. When dealing with assessments of site, environmental remediation design, monitoring, environmental litigation, the quantity of data involved can quickly become overwhelming. Maintaining and organizing that data and keep a balance is insufficient. Visualization is the only means for condensing and communicating vast quantities of data. Visualization provides an invaluable tool to communicate complex data in a form that makes it intelligible to all parties. There are many case studies on visualization of environment-related issues. Some of them are mentioned below:

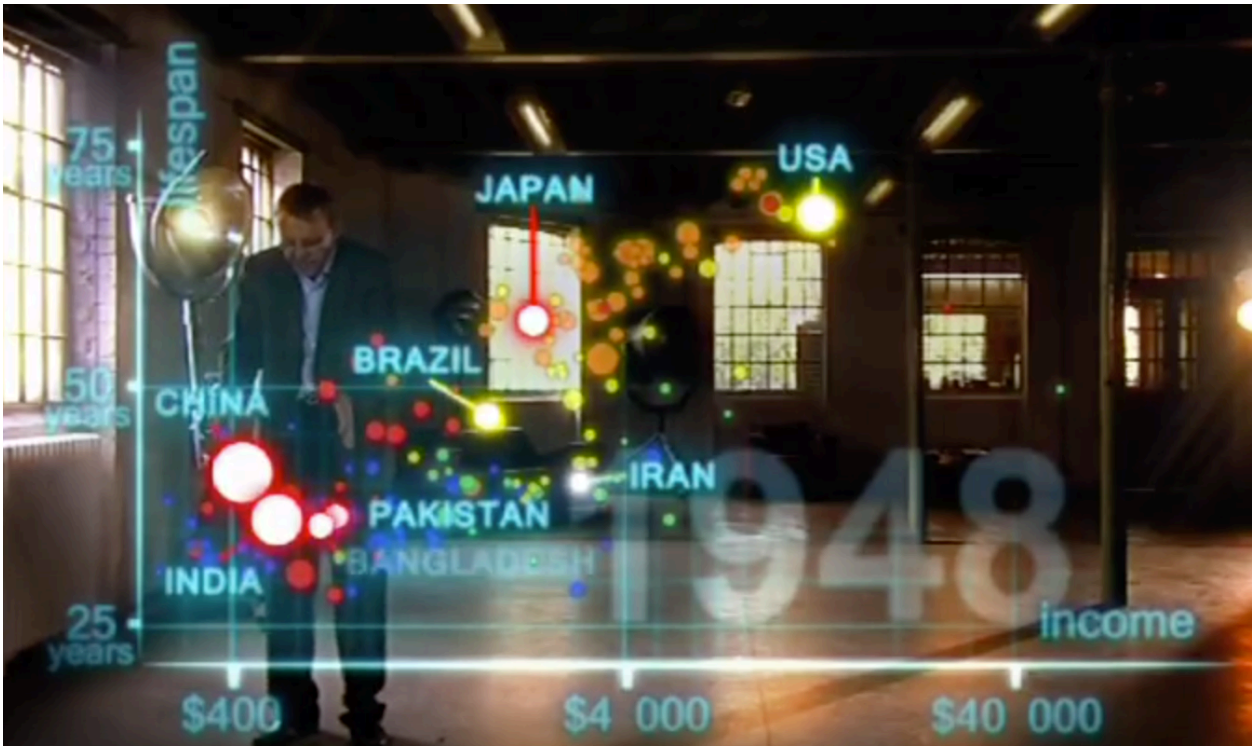
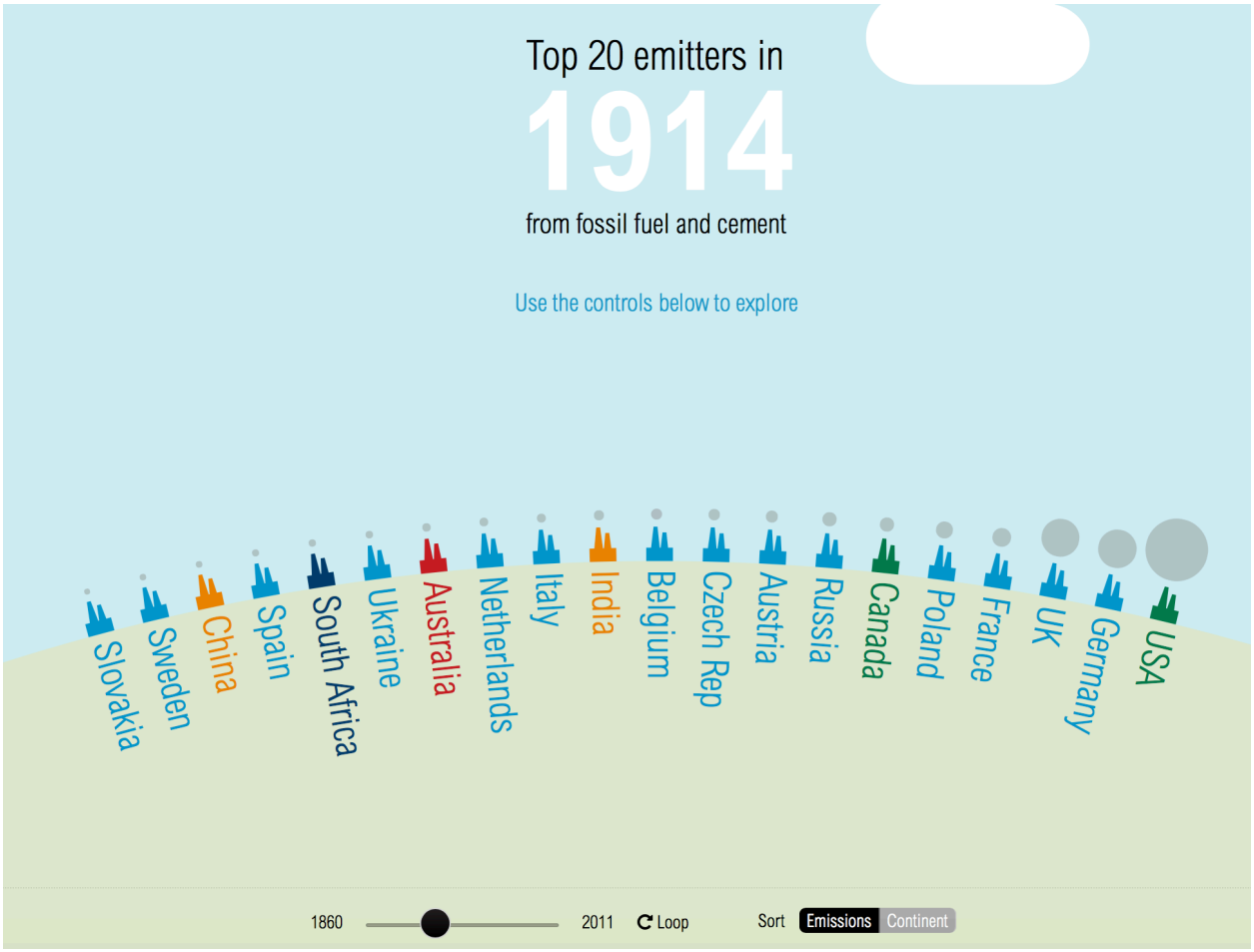


Figure 4.10: Screenshot from "200 Countries, 200 Years, 4 Minutes"

4.6.1 Global Carbon Emissions

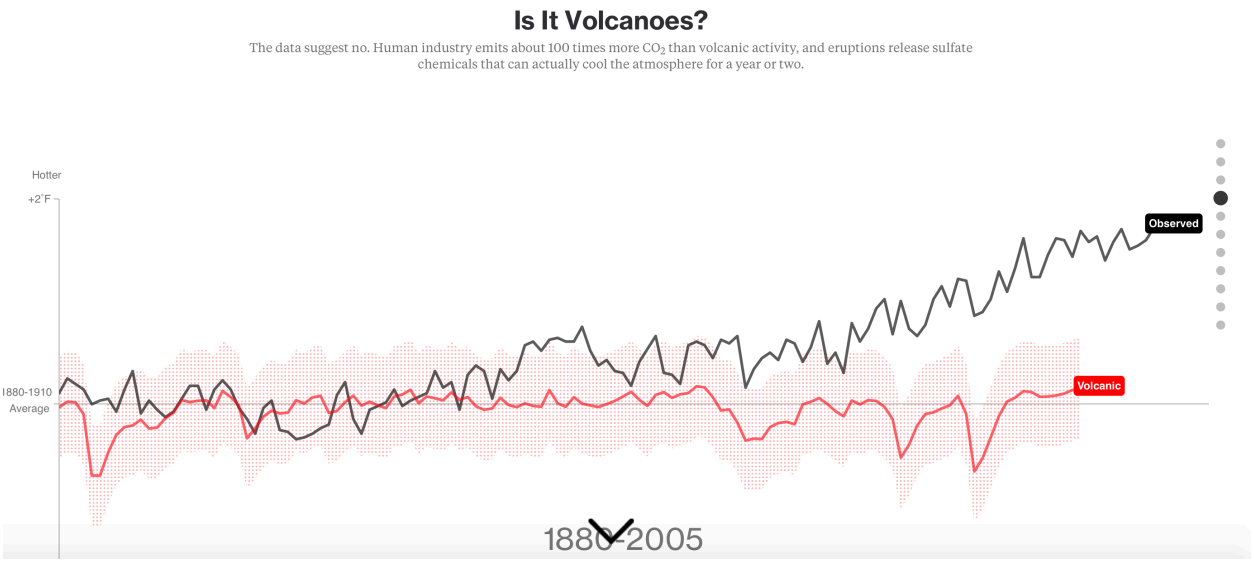
This data visualization, based on data from the World Resource Institute's Climate Analysis Indicators Tool and the Intergovernmental Panel on Climate Change, shows how national CO emissions have transformed over the last 150 years and what the future might hold. It also allows the audience to explore emissions by country for a range of different scenarios (World Resources Institute 2014).



Source: (World Resources Institute 2014)

4.6.2 What's really warming the world?

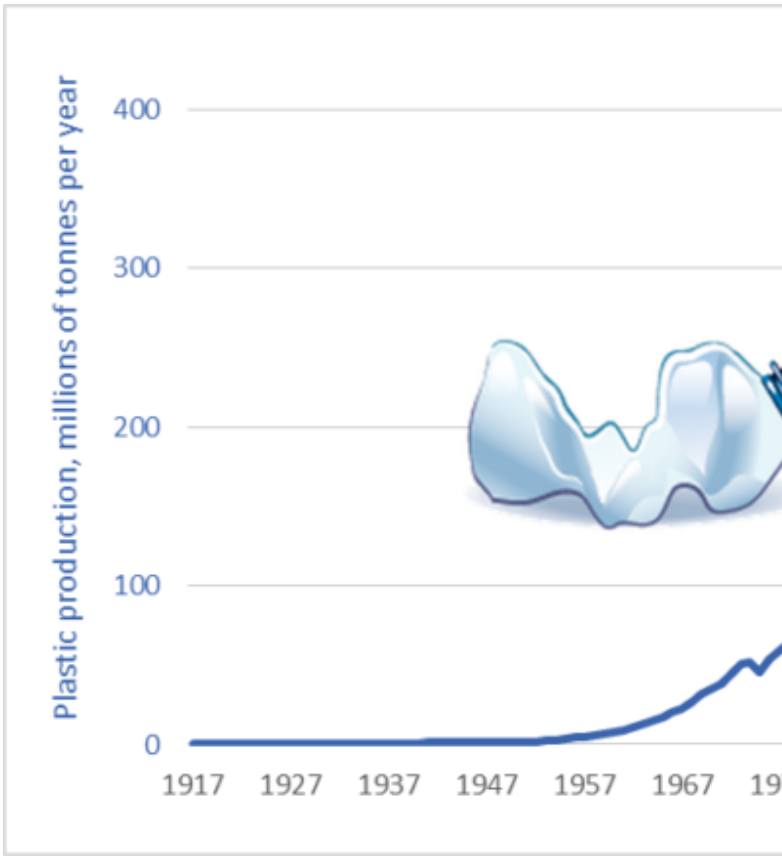
This case study begins by clearly explaining necessary background information and the analytic questions it seeks to answer. Next, it analyzes each factor separately using both verbal explanations and dynamic graphics to compare the observed temperature movements, and then categorizes related factors into "natural factors" or "human factors." After that, it combines all the dynamic graphics into one, which makes the results more accessible and more straightforward to compare. Lastly, the authors provide further detailed explanations of dataset sources to support their results. Overall, this case study is straightforward, easy to understand and informative (Roston and Migliozzi 2015) (Crooks 2017).



Source:(Roston and Migliozzi 2015)

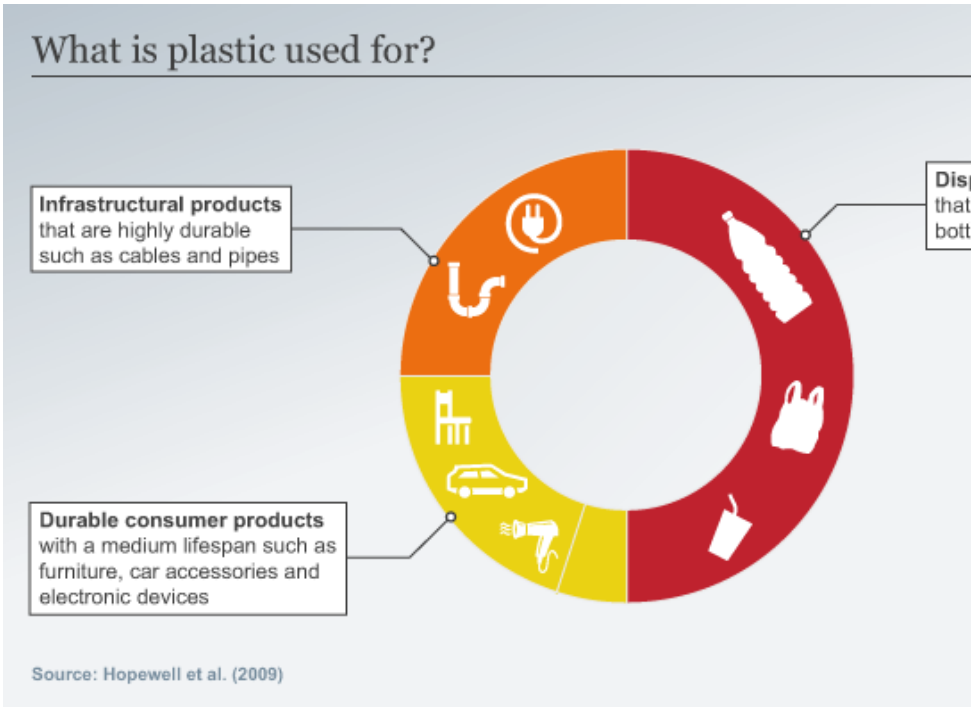
4.6.3 Understanding Plastic pollution using visualization

Plastic pollution is the accumulation of plastic products in the environment that adversely affects wildlife, wildlife habitat, or humans. Human usage of plastic has increased manifolds in last few decades. Since plastic is inexpensive and durable, it has a wide variety of uses in our everyday life. Since the 1950's, an estimated 6.3 billion tons of plastic has been produced, of which only about 9% is recycled (contributors 2019b).



Usage of plastic in last few decades (Qualman 2017):

Plastic has become part of our daily life, and human dependence on plastic has increased over time. The visualization below shows some common plastic products undermining environmental health. (Grün 2016)



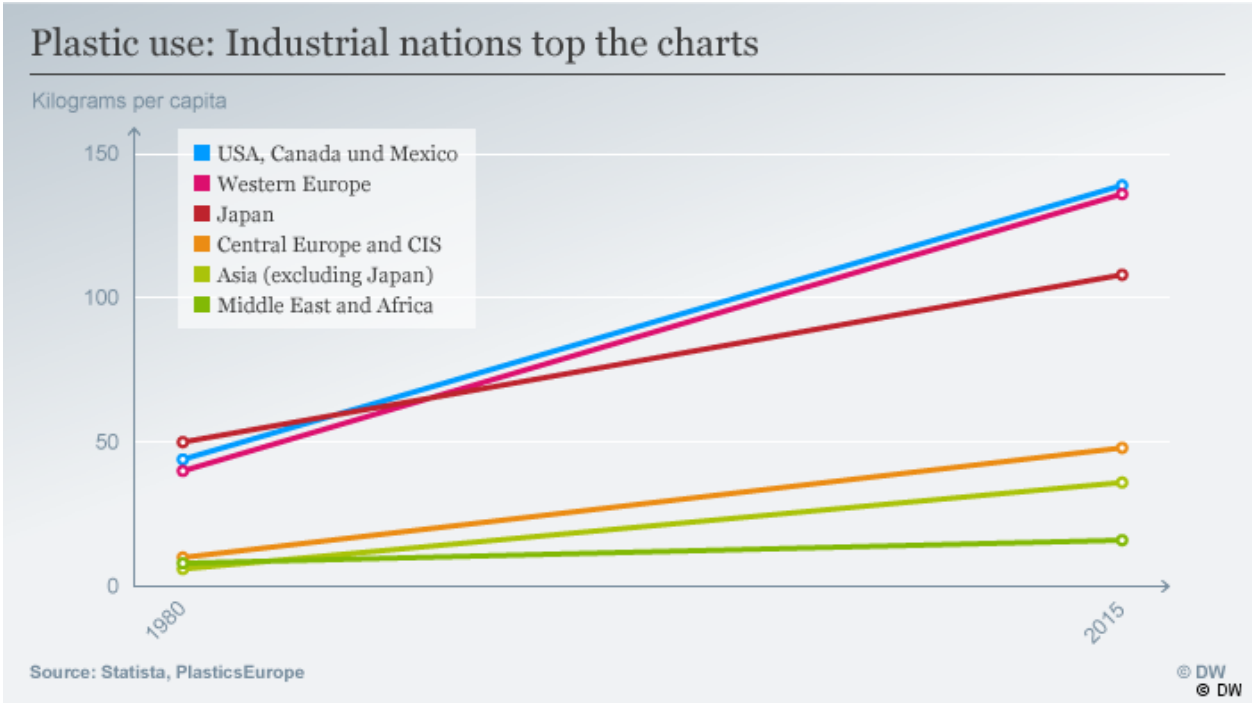
What is plastic used for. (Grün 2016)

With a share of 26 percent, China may be the largest plastic producer in the world; yet the largest plastic

consumer is neighboring Japan. The people living in the island nation have consumption that exceeds that of Africa and the rest of Asia combined.

Donut chart is a modern version of pie-chart which looks cleaner, and embedded visual imagery makes the distribution easy to understand. (Grün 2016)

Plastic Use: Industrial nations top the charts (Grün 2016)

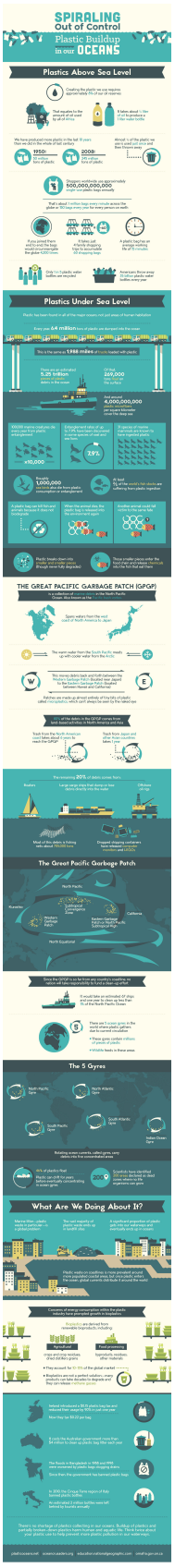


This visualization uses a simple line chart to show increasing trends. A positive aspect of this chart is the removal of the vertical grid which creates noise in the visualization when its objective is to show the trend, rather than the numbers.

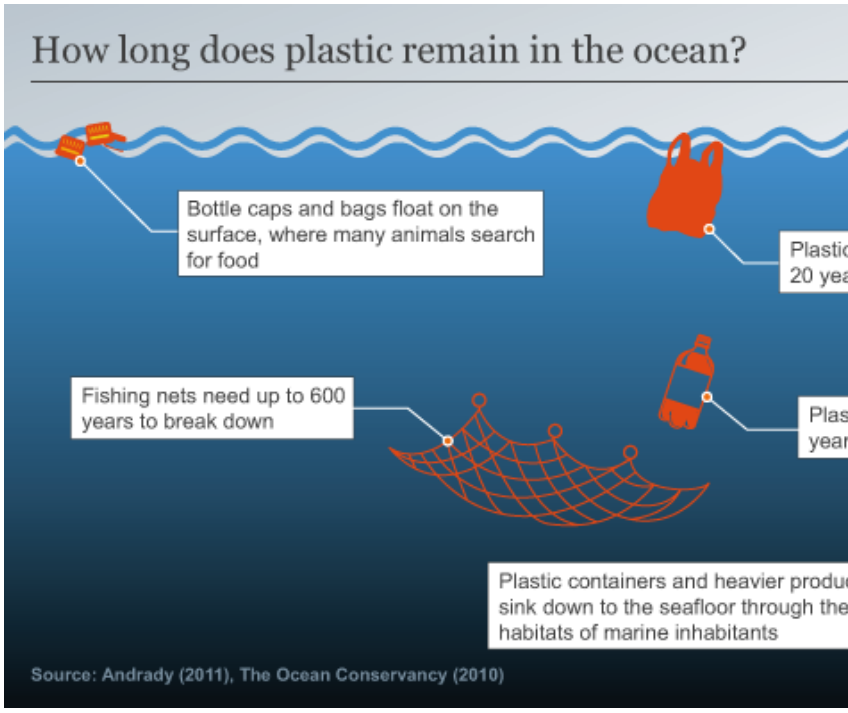
Visualization of Ocean Plastic collection: This worldview visualization shows how much plastic

is in our oceans.(Moret 2014)

"Plastic where it shouldn't be" combines four large-scale plastic marine pollution datasets, each published in a different scientific journal over the last five years, totaling 9,490 surface net tows. It is a symbol map shows the amounts of plastic wastes distribute in oceans. Please note: just because there is no plastic displayed in a certain region does not mean that it isn't there. The open ocean is vast and pollution research is both time- and cost-intensive.(Moret 2014)



Infographic plastic pollution (ROUTLEY 2018)



Infographic plastic pollution (ROUTLEY 2018)

How long does plastic remain in the ocean? (Grün 2016)

Overall, this visualization is useful in the following ways:

- It provides content: those plots serve one of the primary purposes of data visualization - storytelling. It naturally leads the audience to understand the effects of plastic pollution.
- Effective use of charts: the correct use of different types of plots makes the visualization both effective and exciting.
- Efficient use of color: this visualization is a good example of color playing an essential role in a data visualization by guiding the reader to grasp the relationships in the data. There is no redundant color, and no primary color is missing.

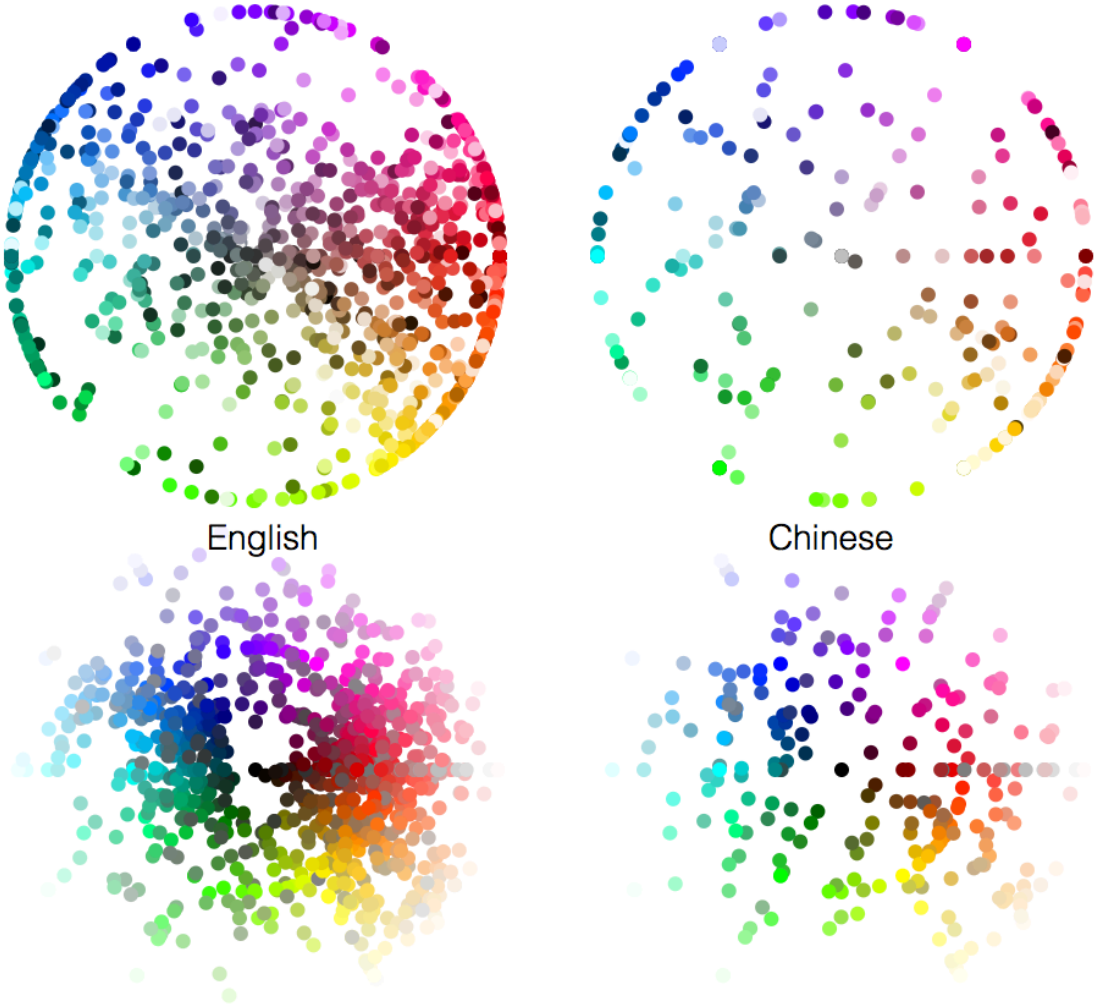
4.7 Language

4.7.1 Green Honey

> Language shapes the way we view the world. Different languages may have vastly different ways of describing things—including color.

Muyueh Lee takes this idea and expands upon it, examining the differences in describing color in Chinese and English through a helpful visualization. The visualization spans a webpage (Lee 2016) referenced in (Kayla Darling 2017). 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 famous "base" colors for English and Chinese. The continuous flow of this visualization helps bring it together, allowing users to scroll through the information at their own pace, but also creating a seamless, creative work. Using data from the English and Chinese versions of the Wikipedia entry on color, the visualization shows the differences in how English speakers and Chinese speakers describe color. Looking at the infographic, it's clear that English (or at least the English Wikipedia article) has more words for color than Chinese does. Additionally, the most popular "base color words" in Chinese are red, blue and green. In English, it's blue, green and pink. English also

differs from Chinese in using place names to distinguish between colors, like in “Persian Blue.”(Kroulek, n.d.)



Source:(Lee 2016)

4.7.2 Linguistic Concepts

This case study is about the use of linguistic concepts; it discusses how the data is being used and how visual graphics are used to deliver the central insights. It presents an educational tool that integrates computational linguistics resources for use in non-technical undergraduate language science courses. By using the tool in conjunction with case studies, it provides opportunities for students to gain an understanding of linguistic concepts and analysis through the lens of practical problems in feasible ways. (Alm, Meyers, and Prud'hommeaux 2017).

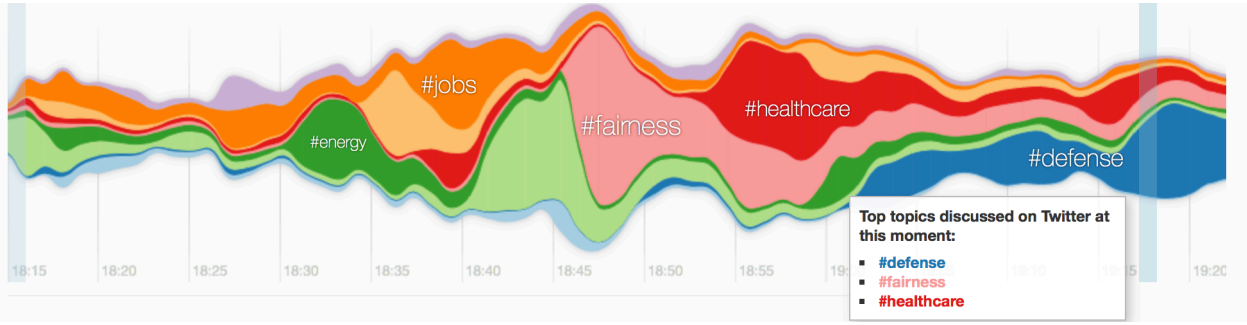
HistoBankVis is a novel visualization system designed for the interactive analysis of complex, multidimensional data to facilitate historical linguistic work (Michael Hund 2015). In this paper, the visualization's efficacy and power are illustrated utilizing a concrete case study investigating the diachronic interaction of word order and subject case in Icelandic.

Much of what computational linguists(CL) fall back upon to improve natural language processing and model

language "understanding" is the structure that has, at best, only an indirect attestation in observable data. The sheer complexity of these structures and the visible patterns on which they are based, however, usually limit their accessibility, often even to the researchers creating or studying them. Traditional statistical graphs and custom-designed data illustrations fill the pages of CL papers, providing insight into linguistic and algorithmic structures, but visual 'externalizations' such as these are almost exclusively used in CL for presentation and explanation. There are particular statistical methods, falling under the rubric of "exploratory data analysis," and visualization techniques just for this purpose are available. However, these are not widely used. These novel data visualization techniques offer the potential for creating new methods that reveal structure and detail in data. Visualization can provide new ways for interacting with large corpora, complex linguistic structures, and can lead to a better understanding of the states of stochastic processes.

4.7.3 State of the Union 2014 Minute by Minute on Twitter

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 (Belmonte 2014). You can slice and dice the data by topic hashtag (for example, #budget, #defense, or #education) and state, resulting in a powerful detailed and cluttered visualization. Since the visualization is about the topic density in a specific time frame, maybe it's a good idea for us to use this kind of format when we encounter the expression of a poisson distribution.



Source:(Belmonte 2014)

4.8 Political Relationships

4.8.1 Connecting the Dots Behind the Election

This article in the New York Times lists several different candidates and creates compelling visuals that link their campaigns to previous ones (Aisch and Yourish 2015)(Kayla Darling 2017). 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 is then connected by arrows. Hovering over a specific dot highlights the connections between the groups. This visual is a great way to summarize what would otherwise require a long slog through years of information into an easily accessible and viewable format so that voters can figure out where the candidates' experiences lie.

4.8.2 A Guide to Who is Fighting Whom in Syria

One of the charts shown in the link (Crooks 2017), the visualization of 'A Guide to Who is Fighting Whom in Syria' is an exciting graphic to study. The visualization and its report can be seen at (Keating and Kirk 2015).

Source:(Keating and Kirk 2015)

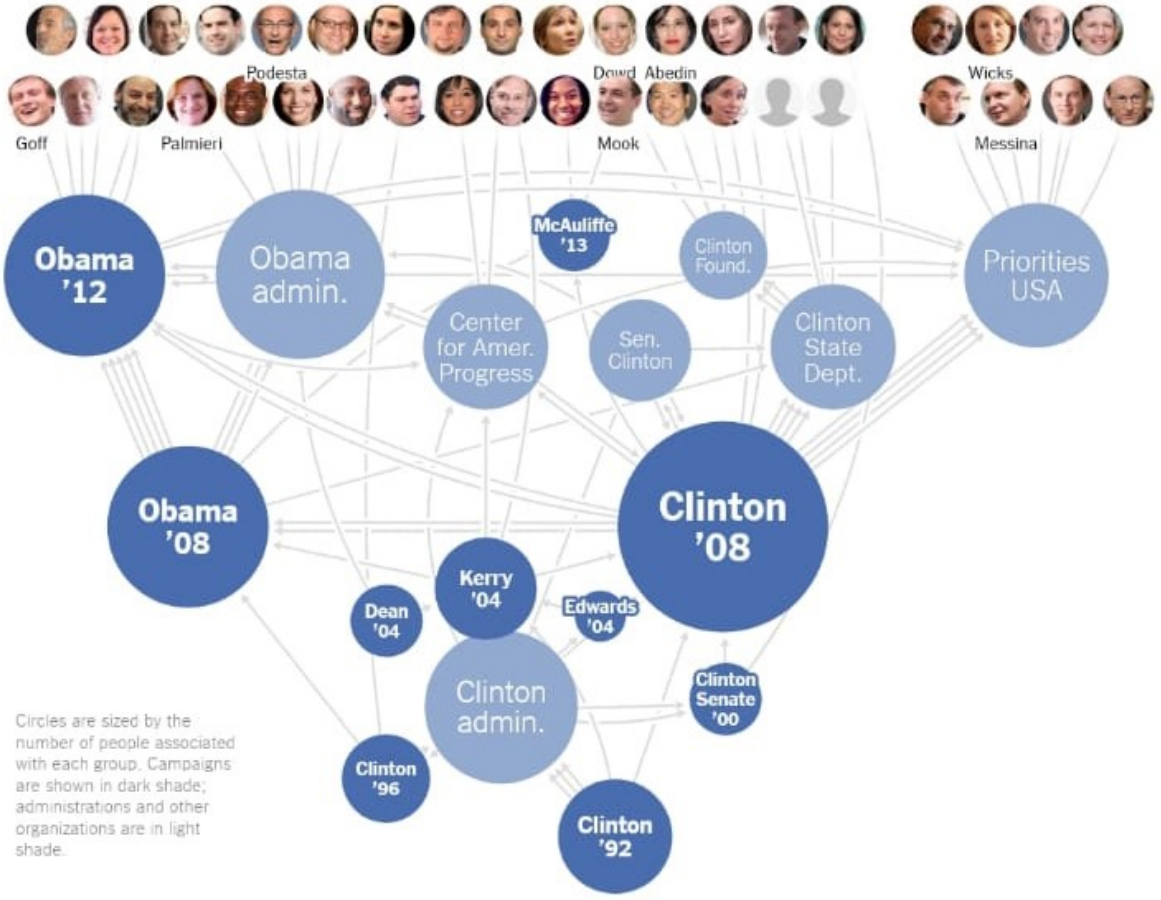


Figure 4.11: Clinton 2016 Campaign Staff

Figure 4.12: Who is Fighting Whom in Syria

This visualization helps elucidate an extremely complicated topic like the Syrian War. It consists of 3 different emojis in three different colors, with each color and facial expression combination showing the ties and conflicts between the various groups involved in the Syrian War. When you click on each emoji, a small dialogue box pops up that explains the relationships between the various countries and rebel groups involved in the war. This is not only easy to understand but is also pleasing to the eyes.

On the other hand, the inherent complexity of relationships between different groups make it difficult to understand the complete picture. If the list of involved parties could be sorted by simplified "sides" (such as Syrian Government on one end with Syrian Rebels on the other) or ranked by how liked they are, then it may be easier for a trend to emerge at first glance. Also, the table format of the visualization means that the data is duplicated, making it appear even more complicated. Instead, one side of the diagonal divide could be greyed-out to simplify the audience's experience with this visualization.

4.9 Uncategorized

4.9.1 Simpson's Paradox

The Visualizing Urban Data Idealab (VUDlab) out of the University of California-Berkeley put together this visual representation of 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 a small but statistically significant bias in favor of women. This is called a Simpson's Paradox.

By "properly pooled," the investigators meant broken down by the department. For instance, men were more inclined towards science and women towards humanities. When compared to each other, the science

Figure 4.13: Green emoji shows 'Friendly' relationship

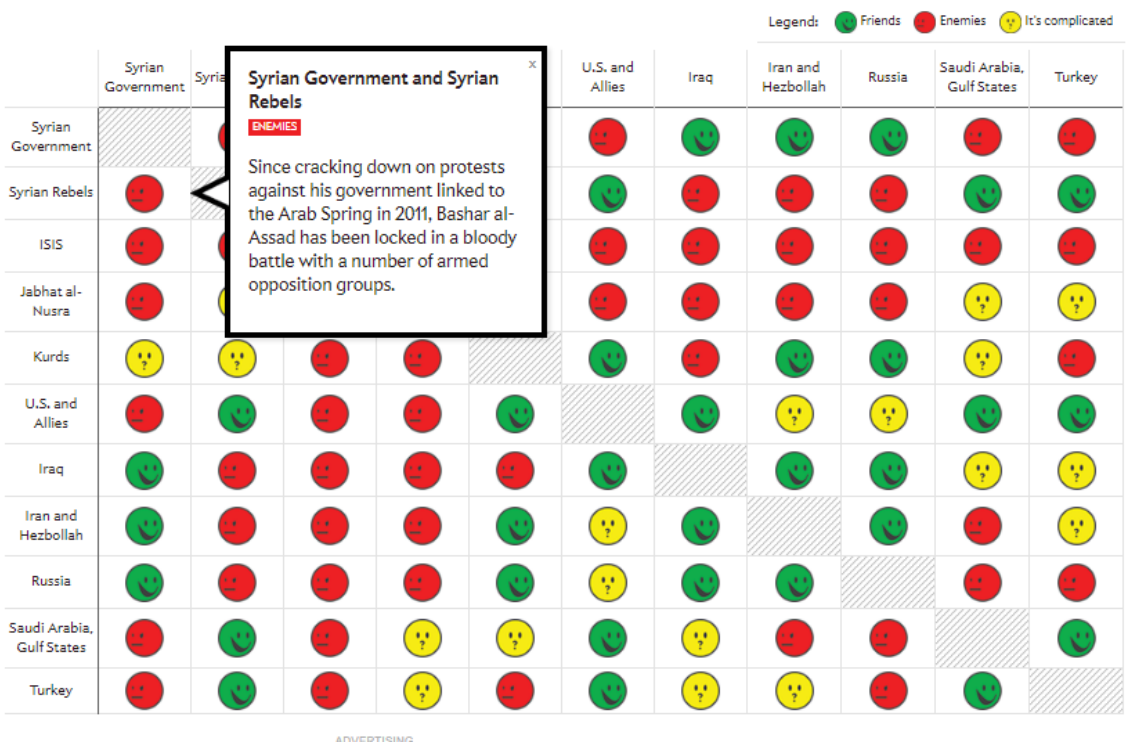


Figure 4.14: Red emoji shows the 'Enemies' relationship

Figure 4.15: Yellow emoji shows 'Complicated' relationship

departments required more specialized skills while the humanities would accept applicants with a more standard undergrad curriculum, thus creating the Simpson's Paradox.

Source:(Lewis Lehe 2013)

4.9.2 Every 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 currently orbiting the Earth. Each satellite is represented by a circular icon, color-coded by country and sized according to launch mass (Yanofsky and Fernholz 2015).

Source:(Yanofsky and Fernholz 2015)

Interactive graph have its own specific advantages. It helps bridge the gap between programmers and non-programmers. This plot is a good example why using interactive graph is a good idea: - It provides an intuitive way for anyone to understand the data regardless of their technical knowledge. - It helps to identifying causes and trends more quickly - It tells a consistent story through data - It improves efficiency of representing data

4.9.3 Malaria

The authors of Vizwiz redesigned "The Seasonality of Confirmed Malaria Cases in Zambia Southern Province" by pointing out what works well, what could be improved, and why their new visualization will be better (Andy 2009).

Figure 4.16: Simpson's Paradox originally from vudlab.com

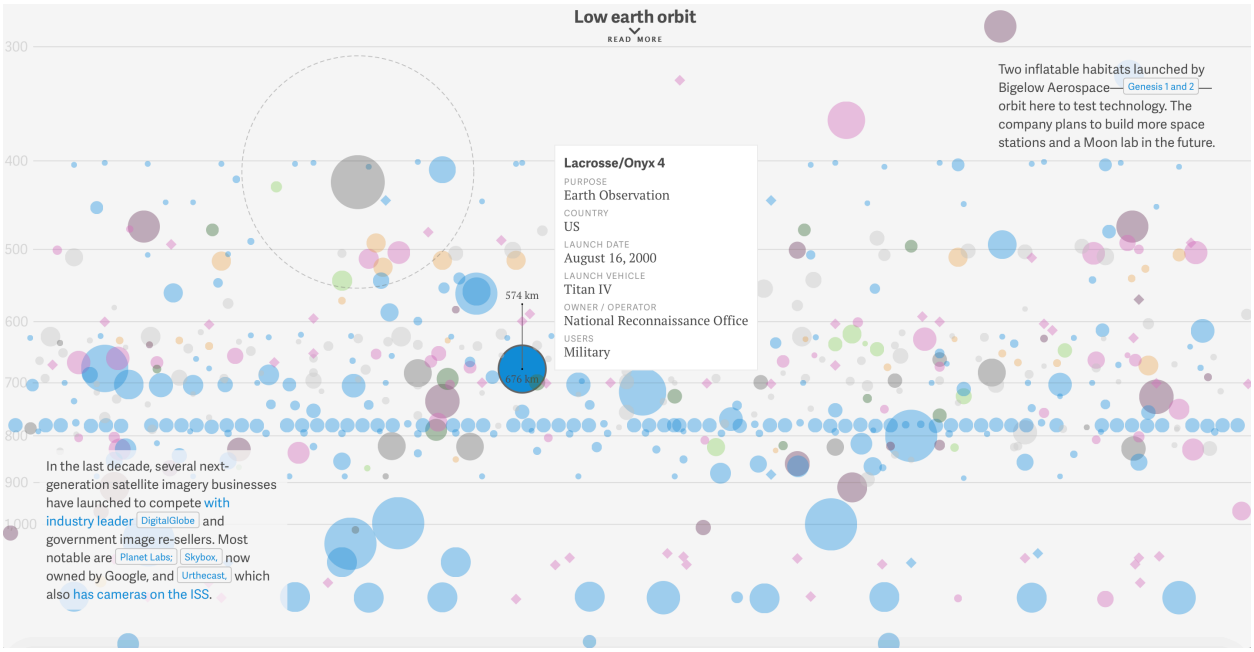
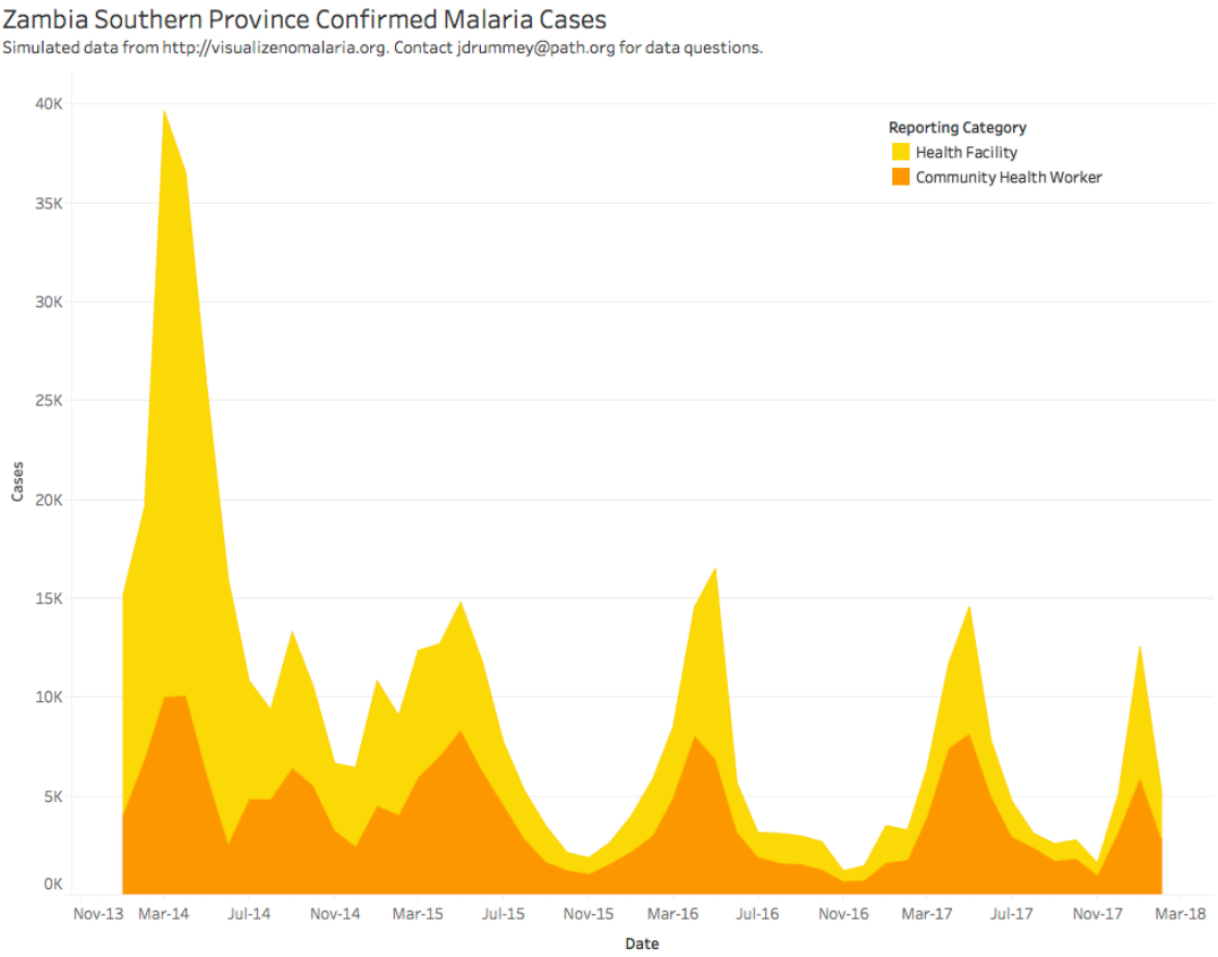


Figure 4.17: Low Earth Orbit Satellites



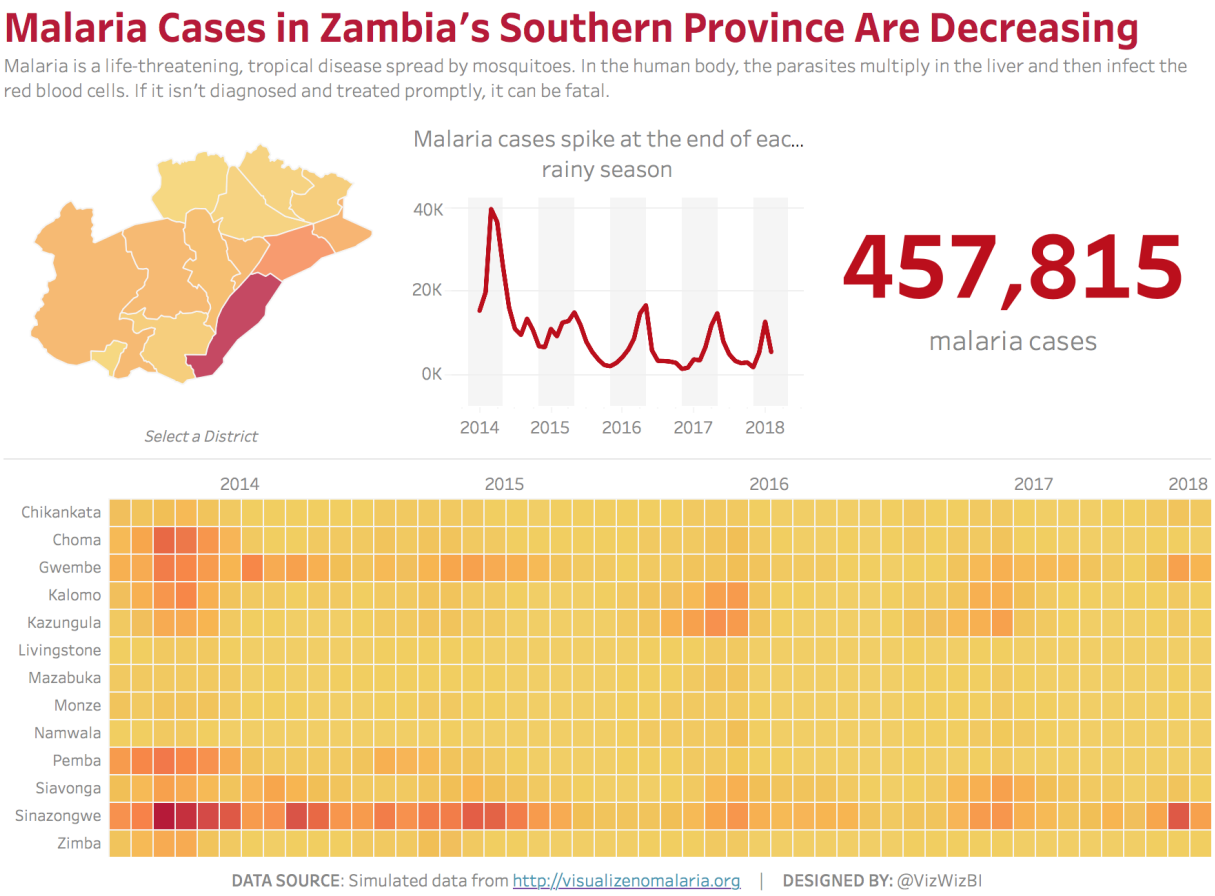
Original Version:

(Source)

This chart below shows number of malaria cases reported for health facilities and community health workers and a comparison between the two over the years. From this chart we can clearly see that as summer approaches, cases of malaria increase indicating a seasonality. The colors are also distinct from each other.

The original visualization effectively shows the seasonality of malaria cases but is unclear if the two reporting categories are stacked or one behind the other and is rather garish. The creator of the redesign made the seasonality more obvious by combining the reporting categories and explaining the spikes better.

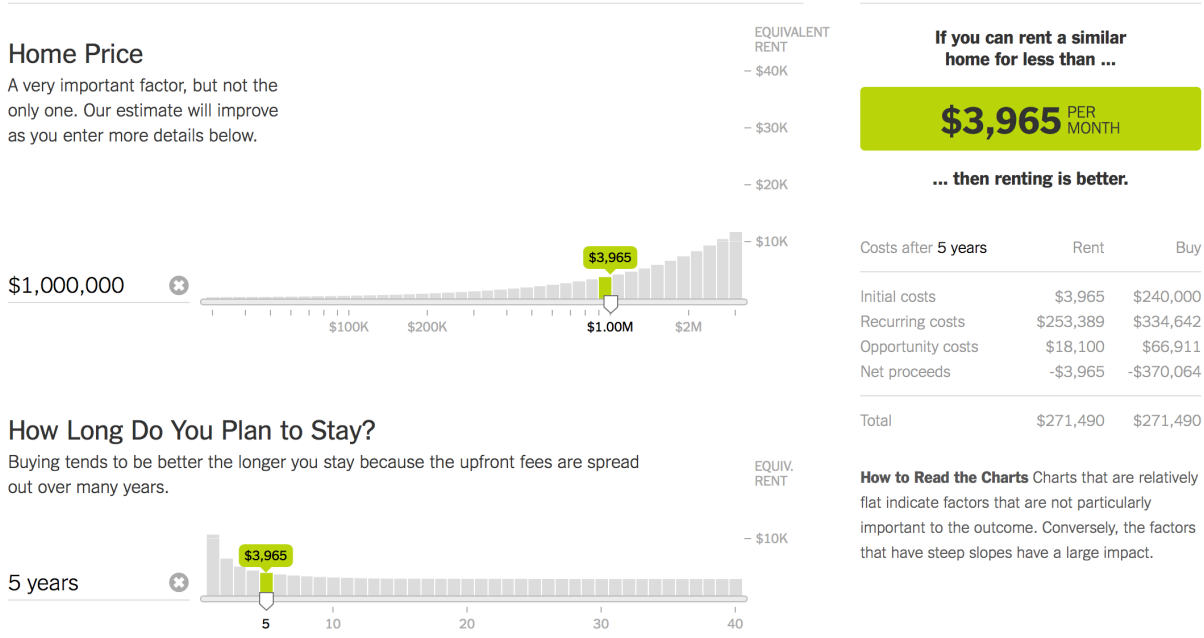
Furthermore, by adding the yearly data split by districts, we can lead to a possible actionable solution to the study of malaria cases in Zambia which is an important objective of visualization. The author has combined the data to find out what the data looks like when combined with health facilities and health workers. And the usage of the color scheme is much more effective than the previous version which makes seasonality more evident.



New Version:

4.9.4 Is it Better to Rent or Buy?

There are many factors involved in deciding to rent or buy a house which has led to many calculators that are supposed to simplify this decision. This calculator includes several sloping charts, each including a factor that will affect how much you will have to pay, such as the individual cost of your home and your mortgage rates (Bostock, Carter, and Tse 2014). A movable scale along the bottom of each chart allows you to enter different data, such as changing the "cost of rent per month" on the side. This can be useful for price comparison: if you can find a similar house to rent for that much per month or less, it is more cost effective just to rent the home. This visualization is incredibly thorough and a useful tool for homeowners of any age and status.



Source:(Bostock, Carter, and Tse 2014)

4.9.5 An Interactive Visualization of NYC Street Trees

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 (Zapata 2014). As the reader hovers over a tree or bar segment, the connected sections light up, making it easier for the reader to look at what otherwise could have been a very dense chart.

We can see what some of the familiar and uncommon trees planted in the five boroughs of New York City are. This visualization allows one to see the distribution quickly. One can make inferences based on the distribution, such as trees in the Bronx and Manhattan seem to be distributed more uniformly compared to the other three boroughs. It gives a direct comparison between the five boroughs which could be used to make a compelling decision by the audience.

Source:(Zapata 2014)

The interactive visualization is an advantage that enables the display, and intuitive understanding of multidimensional data provides a variety of visualization chart types and enables the audience to accomplish traditional data exploration tasks by making charts interactive. Moreover, this visualization provides a good example: it enables the audience to explore on their own and finds exciting facts about NYC street trees.

4.9.6 Adding up the White Oscars Winners

A visualization of all previous winners of the Best Actor/Actress Oscar winners can be seen in an article by Bloomberg ("Adding up the White Oscar Winners" 2016). From the attributes of past Oscars winners, the authors have developed a set of attributes that they believe will continue to be prevalent in future Oscar winners. It is fascinating 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 the features of the actors and actresses who are likely to win the Oscars. Based on different attributes selected, the visualization changes to give you the data specific to the attributes. For each attribute selected,

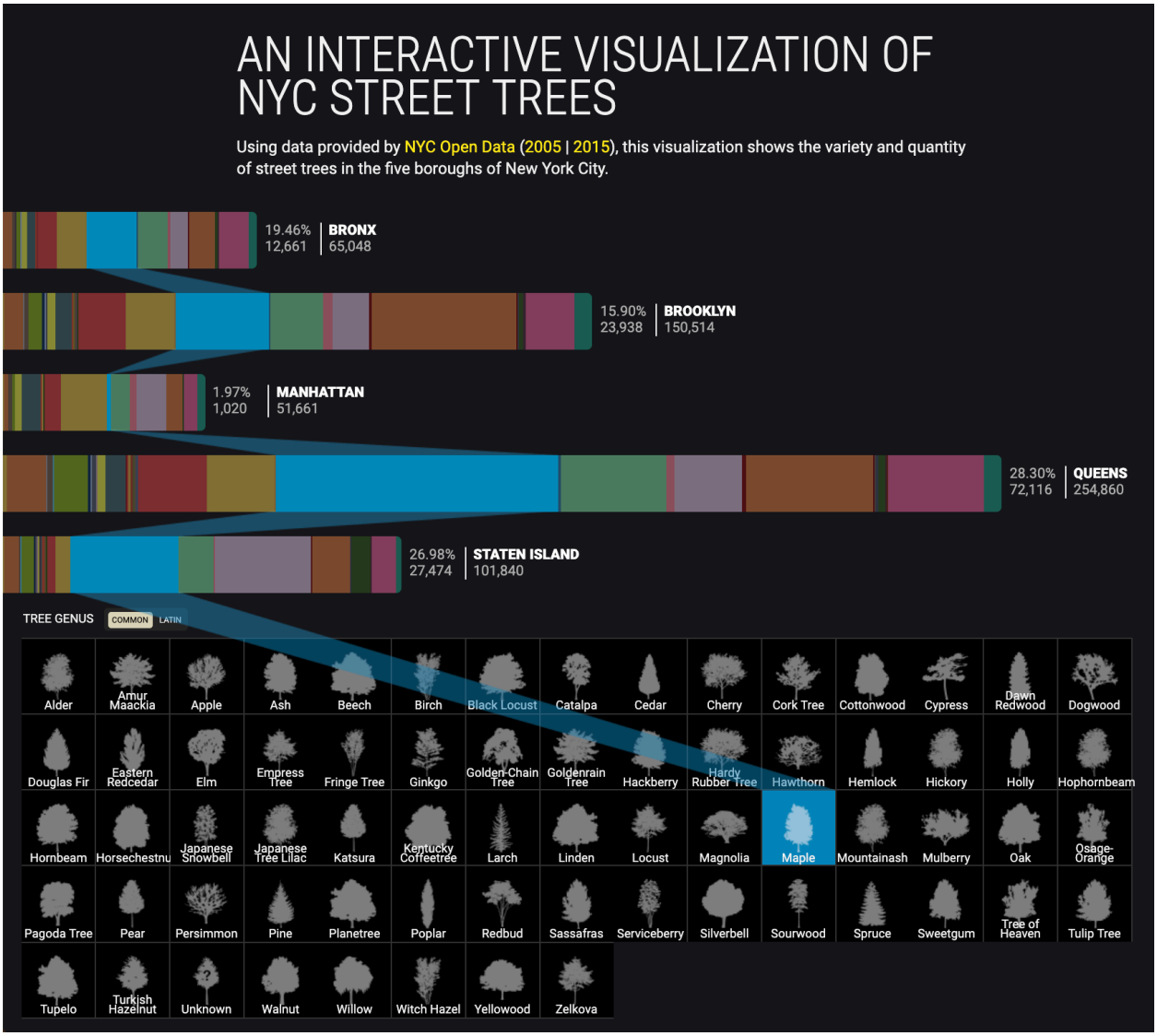


Figure 4.18: NYC Street Trees

Figure 4.19: Best Actor and Best Actress

it gives you a fact about the selected attribute related to the Oscar Winner. For instance, when you select the race, it states "In the entire history of the Oscars all but 8 of the Best Actors and Best Actresses have been white". Similarly, the visualization also gives information about the different aspects of movies that are more likely to win, like 'Length,' 'Month,' 'Budget,' etc., and also predict about the future nominees who are likely to win Oscar.

Source:("How to Build an Oscar Winner" 2015)

4.9.7 Kissmetrics blog: visualization of metrics

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

This example from the Kissmetrics blog is about Facebook's organic reach (Patel 2018). One crucial point discussed in the blog is whether the Facebook's organic reach is decreasing drastically.

The general trend shows that there is a considerable decline in Facebook's page organic reach.

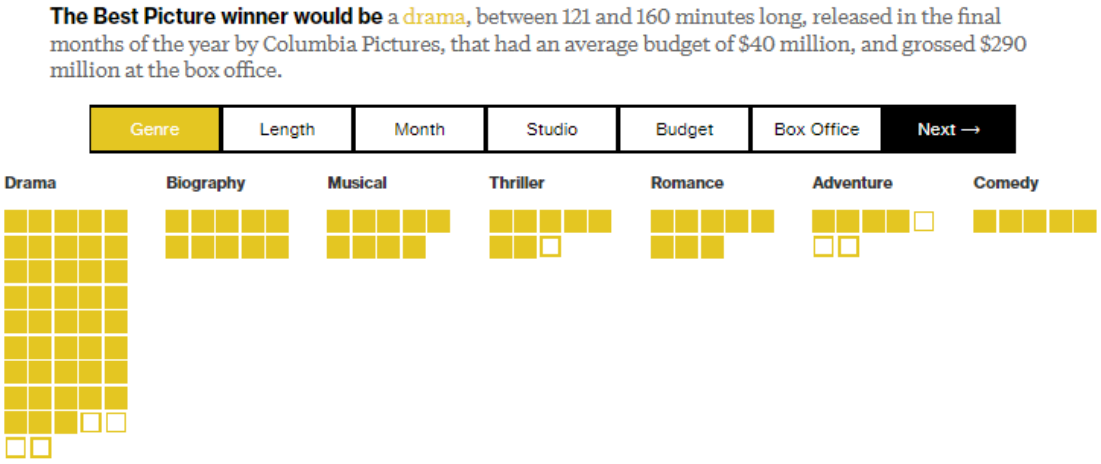
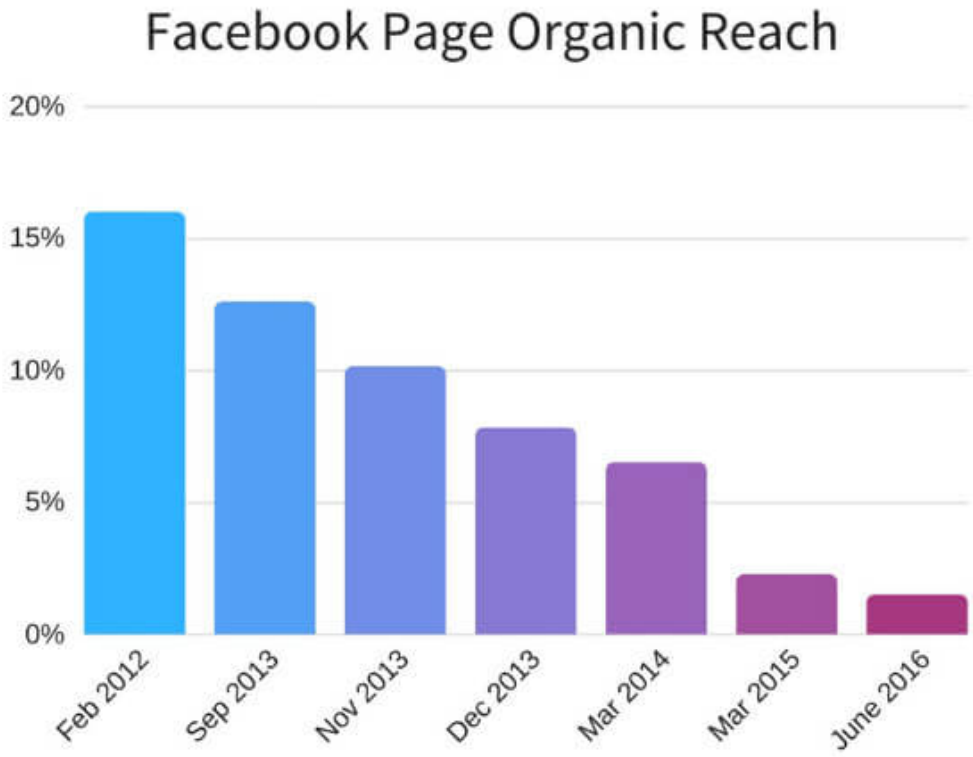
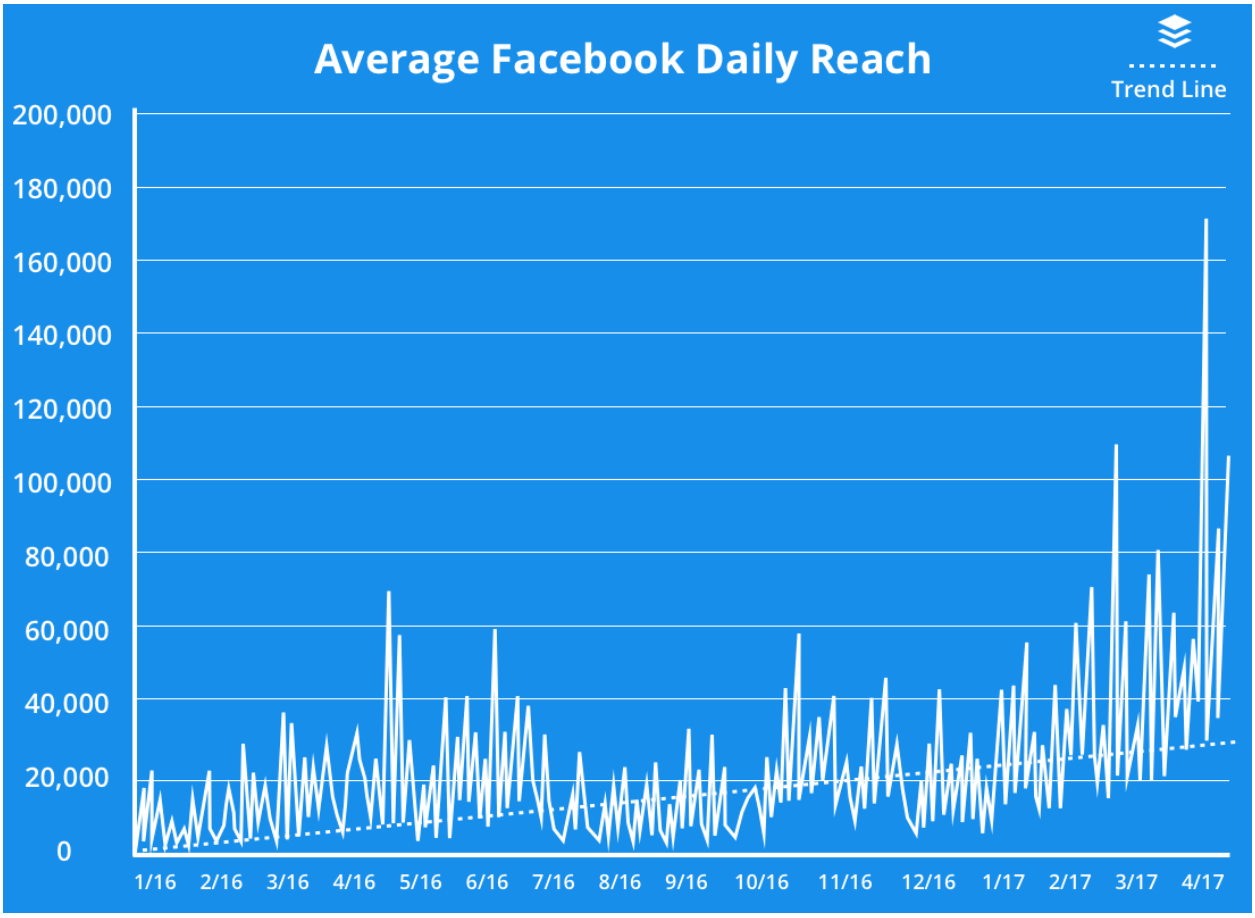
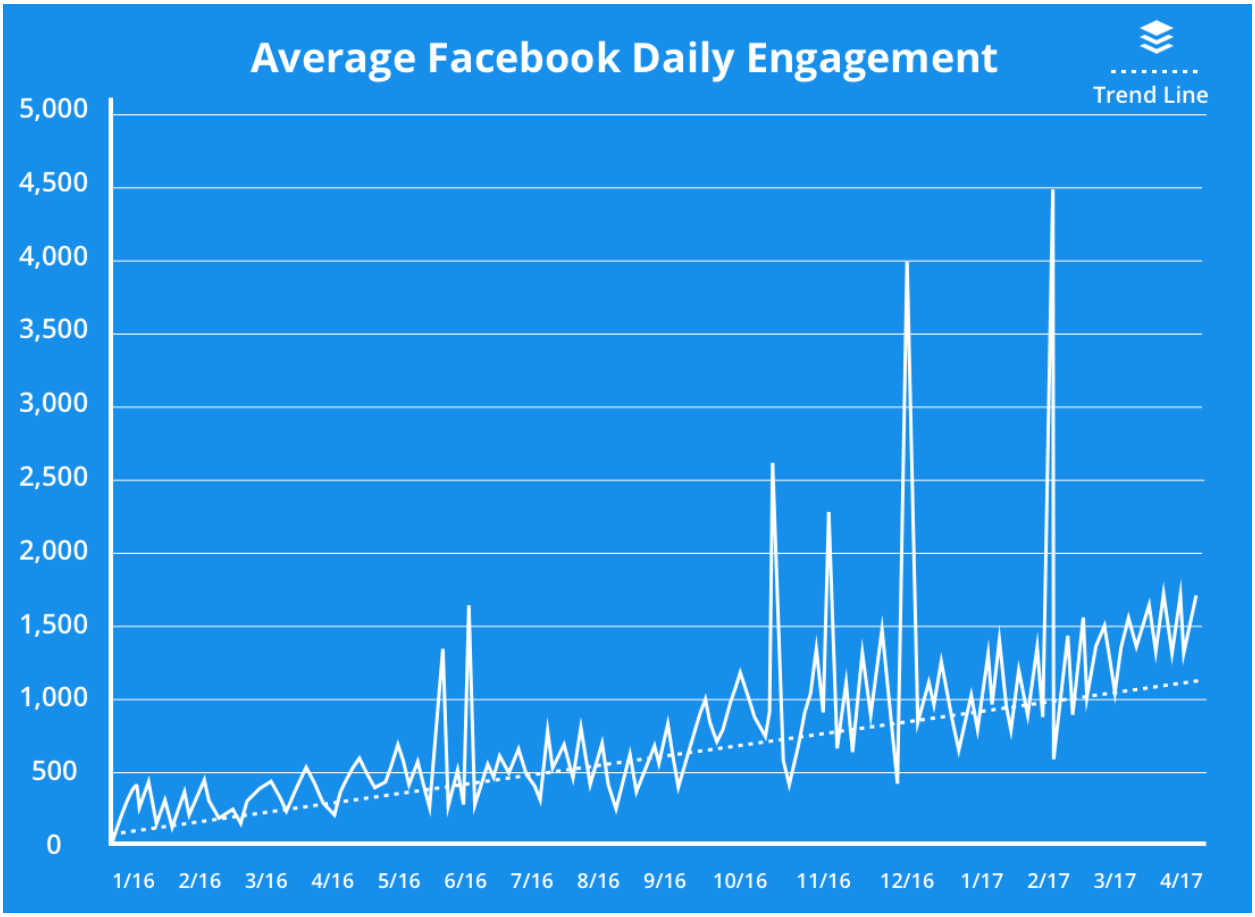


Figure 4.20: Best Picture



The following graphs show that the engagement is increasing; that is, while the quantity of content is decreasing, the quantity is increasing.


Average Facebook Daily Reach
Trend Line
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
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

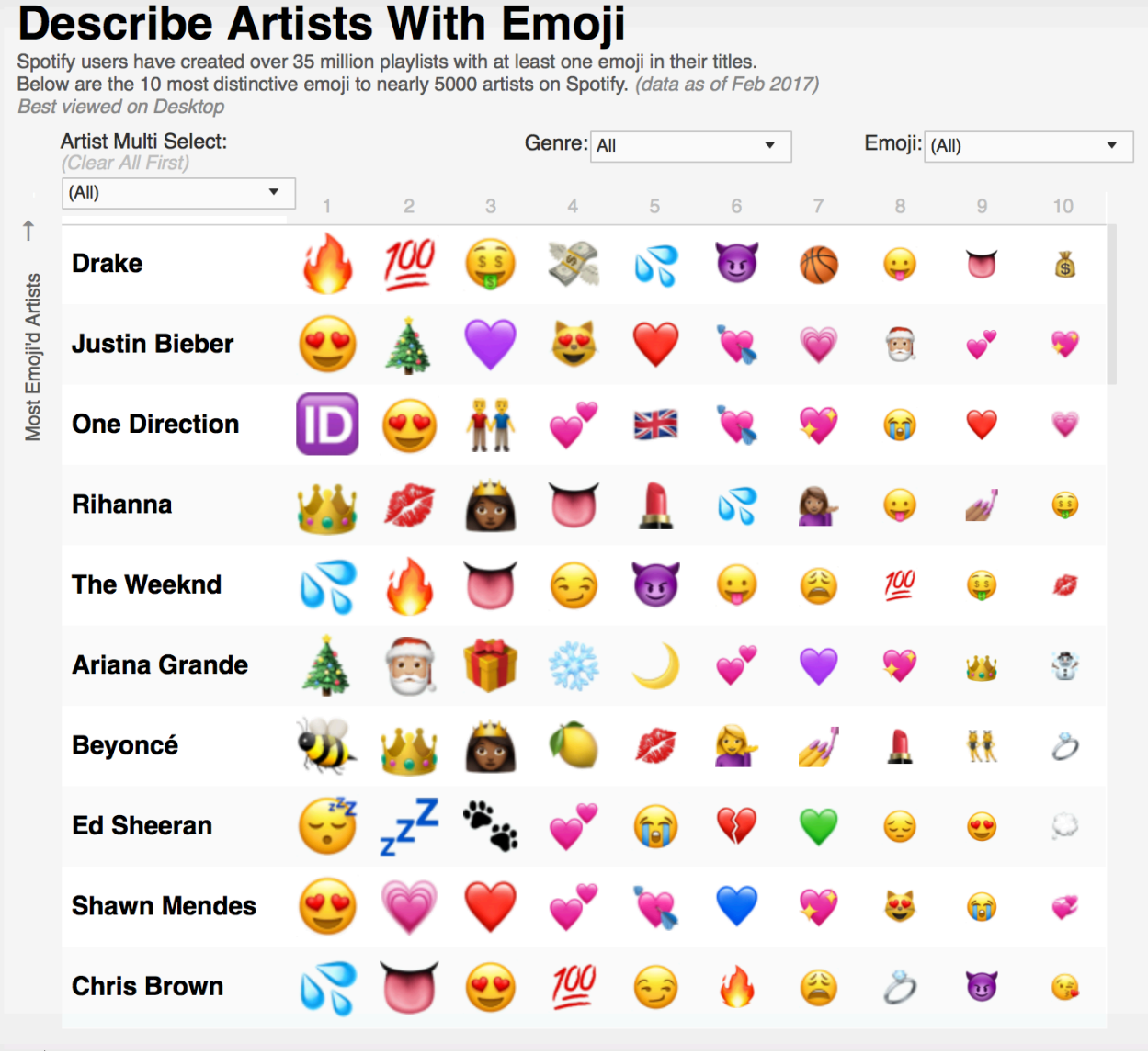


Source:(Patel 2018)

This resonates with what we have learned at class regarding how different perspectives of interpreting data can lead to different conclusions.

4.9.8 Describe Artists with Emoji

Using the data from Spotify, the author listed the ten most distinctive emoji used in the playlists related to favorite artists (Insights 2017). The table being used in this visual is very straightforward to link the artist to the emojis and is very easy to compare among artists. When you hover over the emoji, further information is presented.



Source:(Insights 2017)

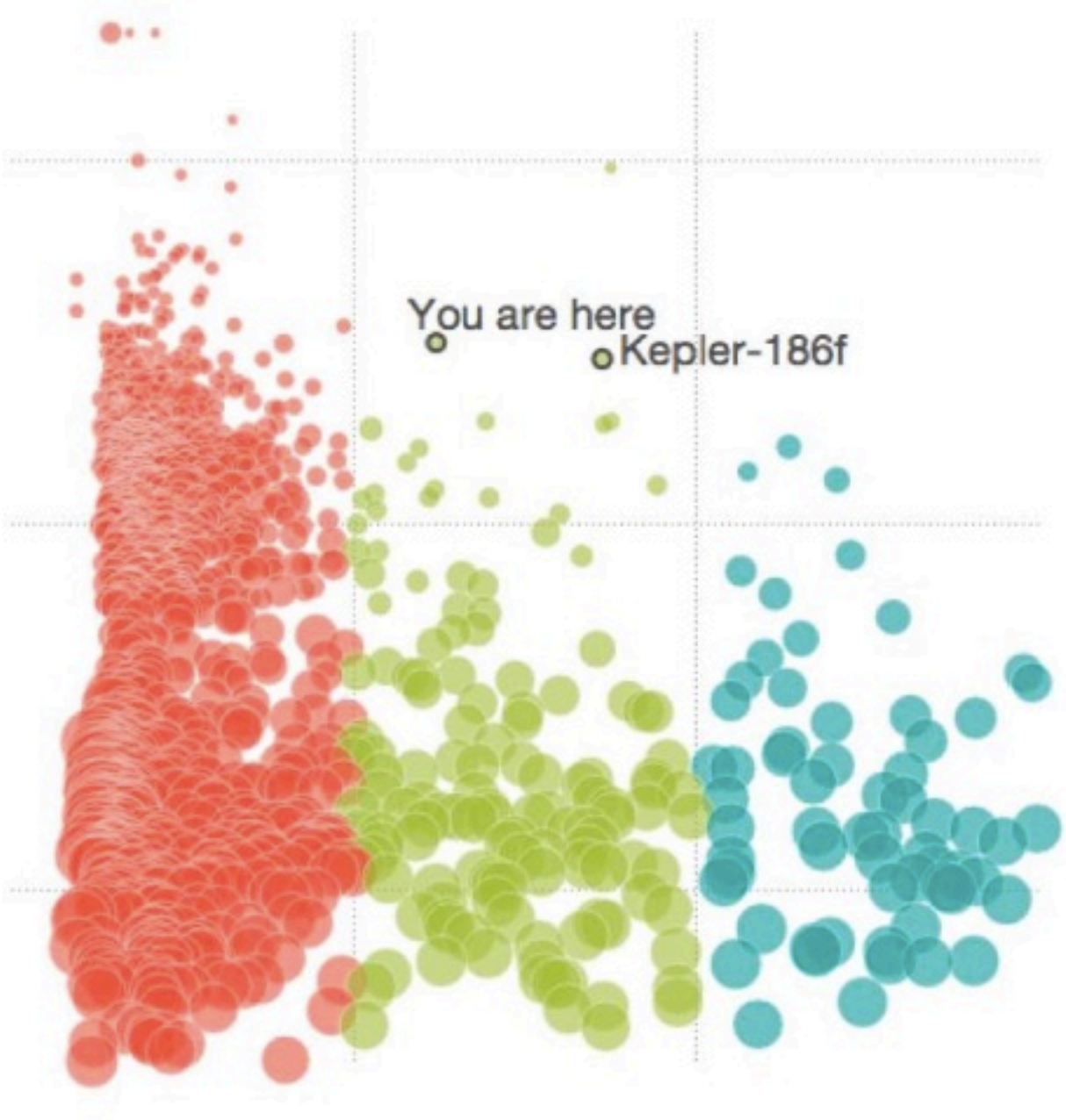
4.9.9 Goldilocks Exoplanets

Using data from the Planetary Habitability Laboratory at the University of Puerto Rico, the interactive graph on Astrobiology plots planetary mass, atmospheric pressure, and temperature to determine what exoplanets might be home, or have been home at one point, to living beings (Tomanio and Gonzalez Veira 2014).

One highlight of the graph is how color has been used. The red dots represent planets that are too hot, the blue dots mean too cold, and the green ones mean just the right temperature. This is very intuitive for people to understand without the necessity to read through the notes. The dots are semi-transparent so the overlapping of planets does not detract from the audience's ability to read the graph.(VERGANO 2014)

Additionally, the size of each dot represents the radius of each planet. At first glance, one might assume that most planets are much larger than Eath, but the visualization includes a note explaining that larger planets are easier to find. This is a good example of how much explanation to include in a visualization, not

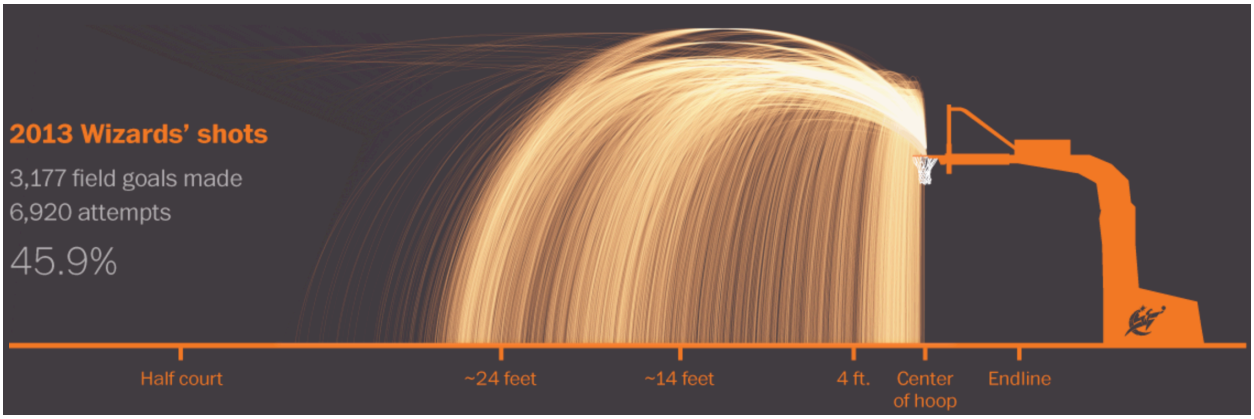
so much that the audience is distracted from the graph but enough that they have the information needed to interpret it.



Source:[Astrobiology]

4.9.10 Washington Wizards' Shooting Stars

This detailed data visualization demonstrates D.C.'s basketball team's shooting success during the 2013 season (Lindeman and Gamio 2014). Using statistics released by the NBA, the visualization allows viewers to examine data for each of 15 players. For example, viewers can see how successful each player was at a variety of types of shots from a range of spots on the court, compared to others in the league.



Source:(Lindeman and Gamio 2014)

Generally this is a data visualization for following reasons because it demonstrates complex infomation in a simple and topic-related format. It highlights fact numbers to tell important information. The use of colr is retrained but efficient. However, it is undefined that what is targeted audience. It can also reduce cognitive overload for lines.

4.9.11 Visualization of big data security: a case study on the KDD99 cup data set

This paper utilized a visualization algorithm together with significant data analysis to gain better insights into the KDD99 dataset:

Abstract
Cybersecurity has been thrust into the limelight in the modern technological era because of an array of attacks often bypassing new intrusion detection systems (IDSs). Therefore, deciphering better methods for identifying attack types to train IDSs more effectively has become a field of great interest. Critical cyber-attack insights exist in big data; however, an efficient approach is required to determine strong attack types to train IDSs to become more active in critical areas. Despite the rising growth in IDS research, there is a lack of studies involving big data visualization, which is crucial. The KDD99 dataset has served as a reliable benchmark since 1999; therefore, this dataset was utilized in the experiment. This study utilized a hash algorithm, a weight table, and sampling method to deal with the inherent problems caused by analyzing big data: volume, variety, and velocity. By utilizing a visualization algorithm, the researchers were able to gain insights into the KDD99 dataset with precise identification of "normal" clusters and described distinct clusters of possible attacks.

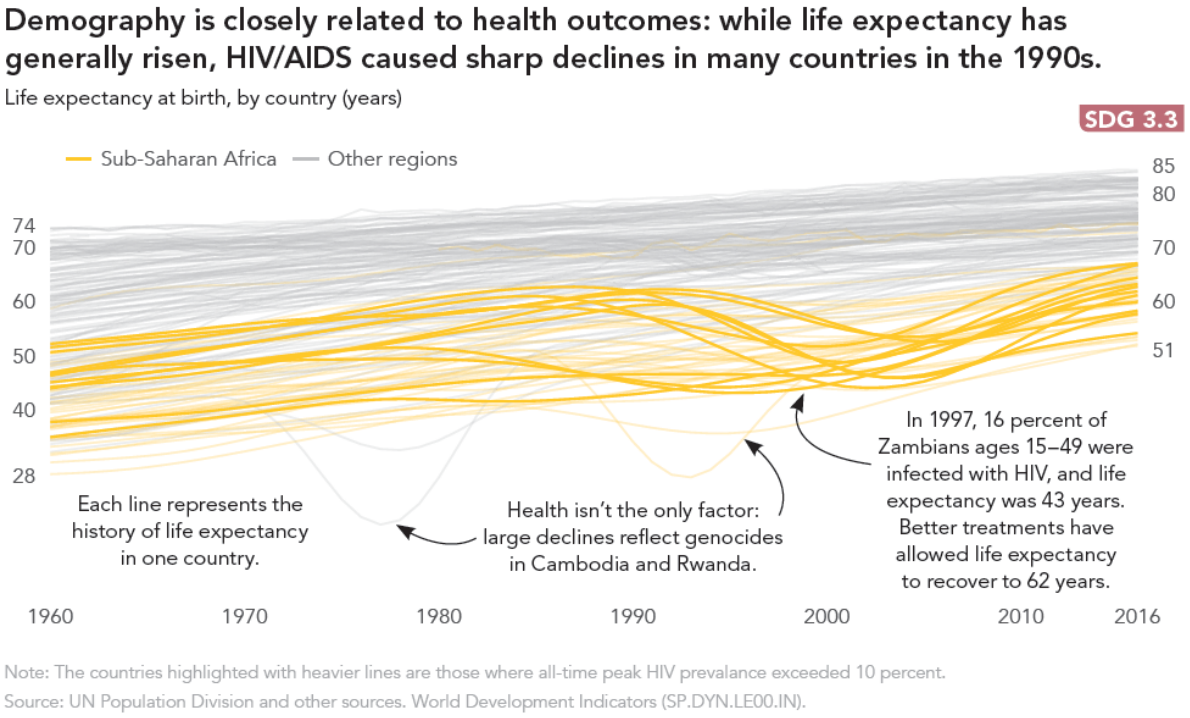
To read the full paper, please follow the reference link:

(Ruan et al. 2017)

4.9.12 The Atlas of Sustainable Development Goals 2018 - Data Visualization of World Development

(TEAM 2018)

This is an exciting source and an excellent visual guide to data and development. It discusses trends, comparisons, and measurement issues using accessible and shareable data visualizations. As the graphs cite below, they are informative and clean:

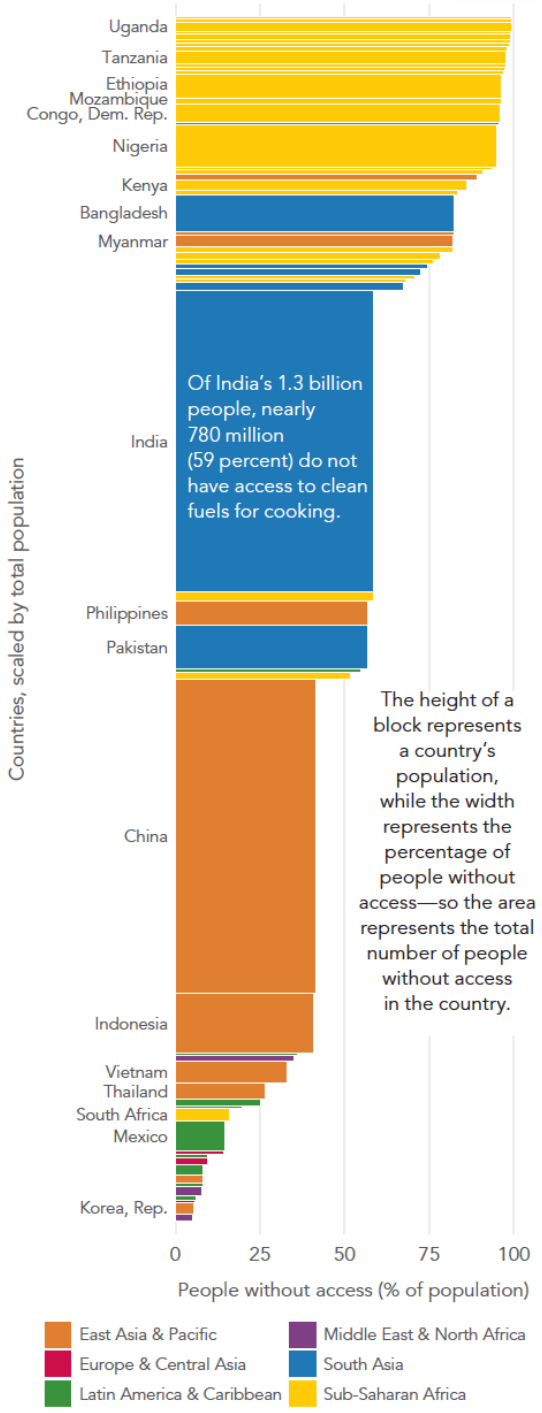


1 2

Worldwide, 3 billion people lack access to clean cooking fuels and instead use fuels that create health risks.

People without access to clean fuels and technologies for cooking, 2016

SDG 7.1

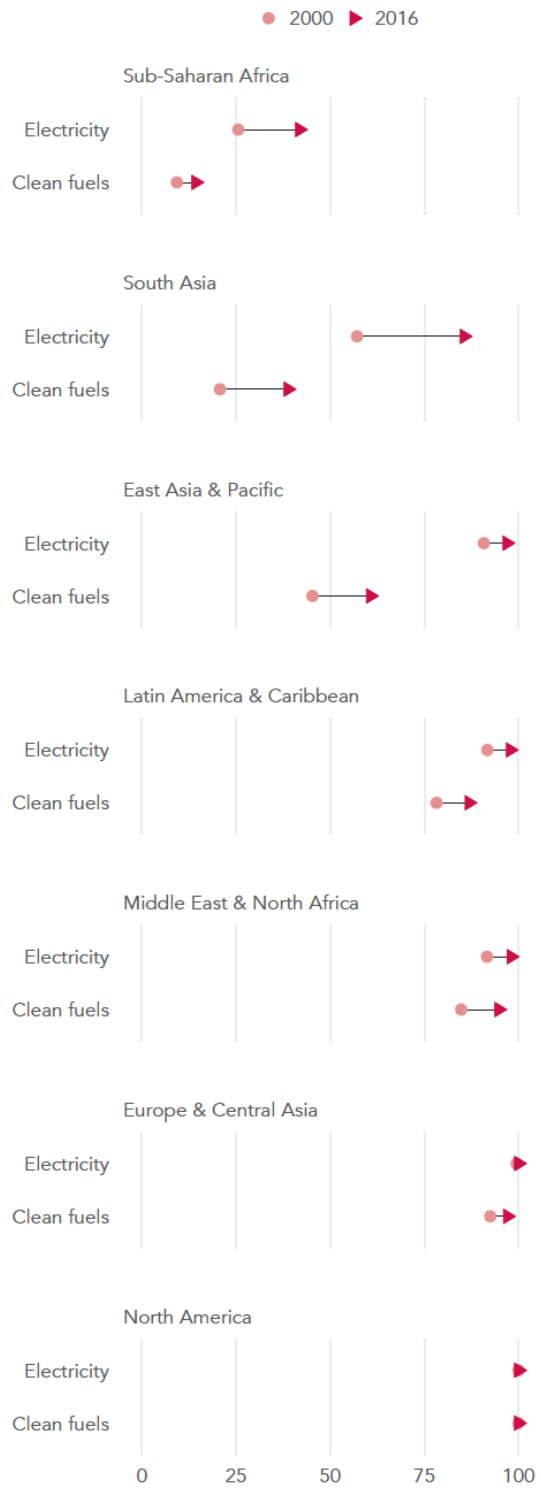


Note: Excludes countries with a population of less than 10 million or an access rate above 95 percent

Source: WHO. WDI (EG.CFT.ACCS.ZS; SP.POP.TOTL).

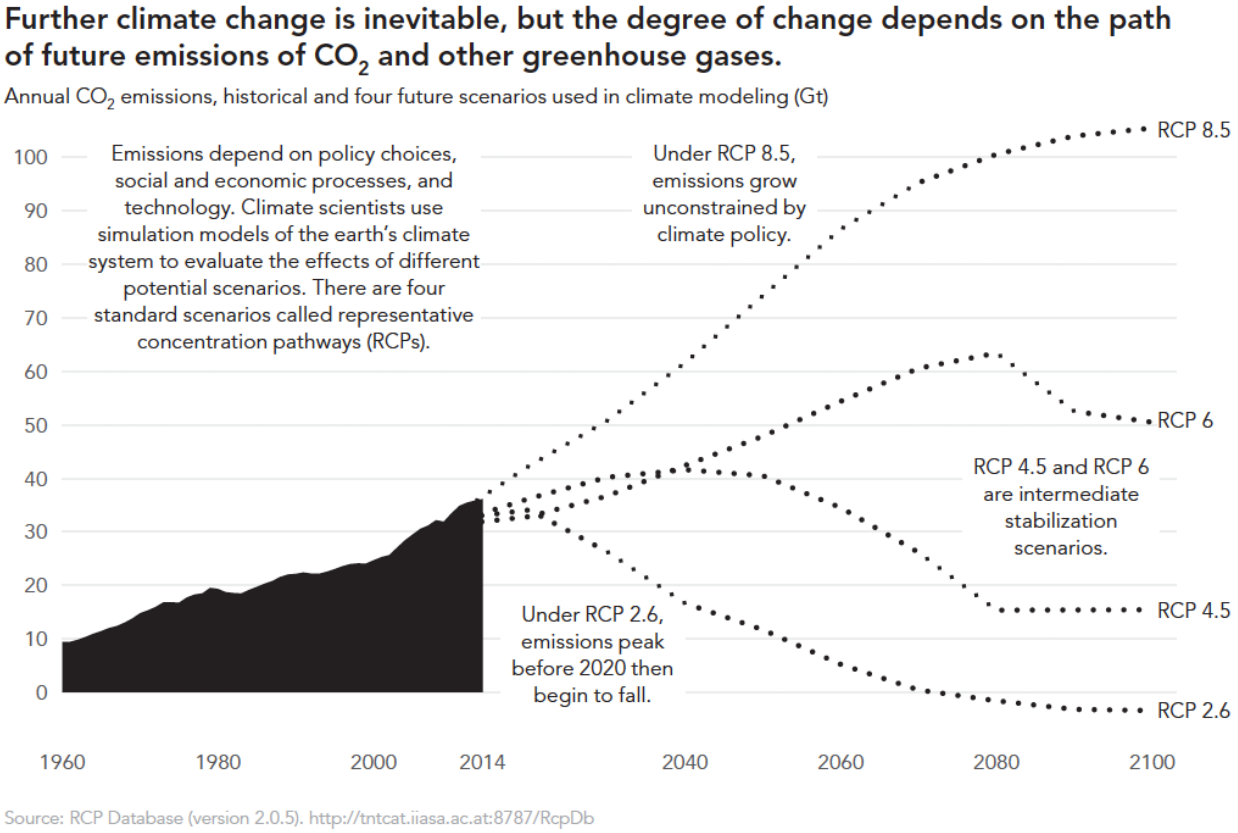
In South Asia and Sub-Saharan Africa gains in access to clean fuels have not kept up with those in access to electricity

Access rates, 2000 and 2016 (% of population)



Source: World Bank, WHO. WDI (EG.ELC.ACCS.ZS; EG.CFT.ACCS.ZS).

1 2



The data draws on the World Development Indicators- the World Bank's compilation of internationally comparable statistics about global development and the quality of people's lives. For each of the SDGs, relevant indicators have been chosen to illustrate important ideas. The Atlas features maps and data visualizations, primarily drawn from World Development Indicators (WDI) - the World Bank's compilation of internationally comparable statistics about global development and the quality of people's lives.

The editors have been selected to emphasize on essential issues by experts in the World Bank's Global Practices. The Atlas aims to reflect the breadth of the Goals themselves and presents national and regional trends and snapshots of progress towards the UN's seventeen Sustainable Development Goals related to: poverty, hunger, health, education, gender, water, energy, jobs, infrastructure, inequalities, cities, consumption, climate, oceans, the environment, peace, institutions, and partnerships.

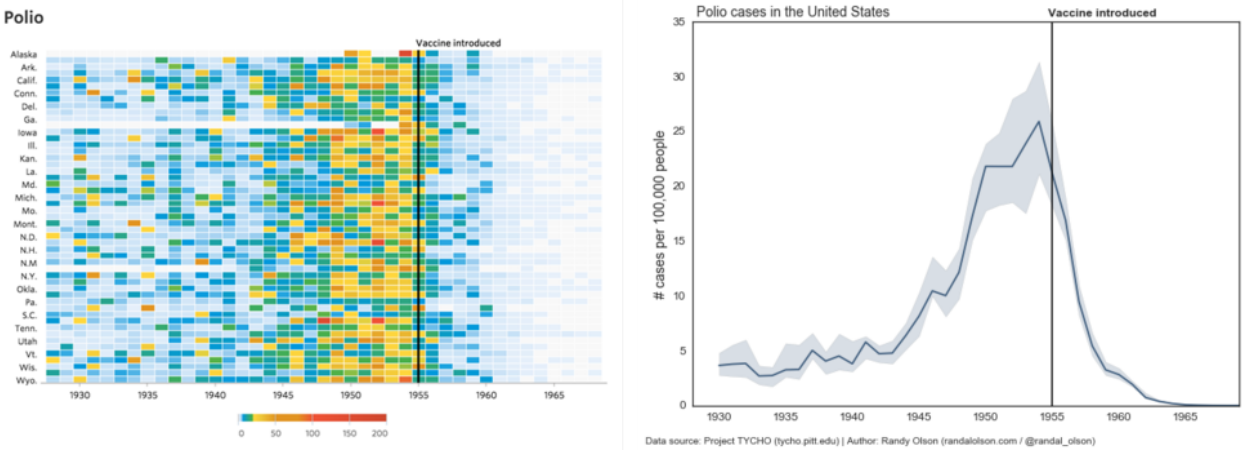
Contents of this publication: (Group 2018a). The data is available at (Group 2018b). The code used to generate the majority of figures is available at (Whitby 2018).

4.9.13 Is Beauty Important?

This case study is about this article: https://www.infoworld.com/article/3048315/the-inevitability-of-data-visualization-criti
html

Andy Cotgreave is the current Senior Technical Evangelist at Tableau. In the above article he defends the use of elaborate visualizations and argues that beauty is a quality worth pursuing when making data visualizations. One visualization that he focuses on is a heat map that shows the effect of introducing vaccines

on the number of polio cases in the US made by the Wall Street Journal. This particular visualization received a great deal of attention, and was sent around the internet to demonstrate the positive effects of vaccination. After spending some time on the internet, another author named Randy Olson responded with his own article where he remade the heat map as a simple line graph. Both versions are shown below.



In his article, Cotgreave argues that the heat map was visually striking, and its novelty made him more likely to interact with it. As someone involved in visualizations, he seen hundreds, if not thousands of line graphs, and would've likely skipped over the line graph version. Cotgreave doubts that the line version would have won awards, or been virally shared as the heat map was. While Cotgreave acknowledges the readability of the line graph, he ultimately feels that there is a place for visualizations to be beautiful.

The takeaway then, is that the visualization you choose to present should be tailored to your situation. In other words, think of your audience. If you were presenting your visualization to the internet at large, then being beautiful and novel is important. If your visualization becomes viral, then it will advance and promote your message to exponentially more people. On the other hand, if you have a more limited audience, like a team of managers, that wants visualizations that can be read quickly, then the line chart will be more suitable.

Chapter 5

Patterns

This chapter is a practical guide to a plethora of data visualizations; it explores different types of visualizations and tools and provides helpful tips for using them effectively.

In general, there are two basic types of data visualisation: exploration, which helps find a story the data is telling you, and explanation, which tells a story to an audience. Both types of data visualisation must take into account the audience's expectations.

5.1 Data Exploration (Like, Outlier Detection)

(Arribas-Gil and Romo 2014) We can use data visualization for outlier detection in a data set. Different methods for outlier detection in functional data have been developed over the years. Several of these methods rely on different notions of functional depth, robust principal components, or random projections of infinite-dimensional data into R. Some distributional approaches have also been considered (Gervini 2012). 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. Other types of outliers include:

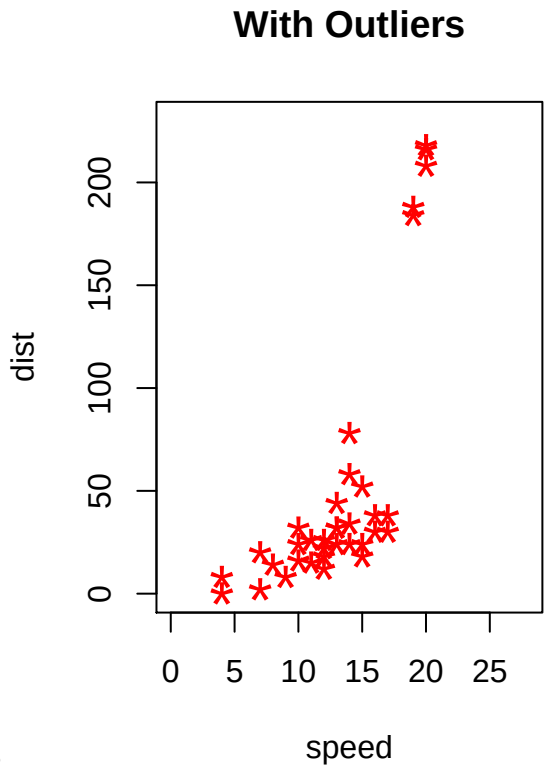
Outlier	Description
Global Outliers (or "point anomalies")	A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found.
Contextual (Conditional) Outliers	A data point is considered a contextual outlier if its value significantly deviates from the rest of the data points in the same context. Note that this means that the same value may not be considered an outlier if it occurred in a different context. If we limit our discussion to time series data, the "context" is almost always temporal, because time series data are records of a specific quantity over time. Contextual outliers are common in time series data.
Collective outliers	A subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense. In time series data, one way this can manifest is as normal peaks and valleys occurring outside of a time frame when that seasonal sequence is normal or as a combination of time series data that is in an outlier as a group.

Below is a simple example. Outlier treatment is important because it can drastically bias/change the fit

estimates and predictions.

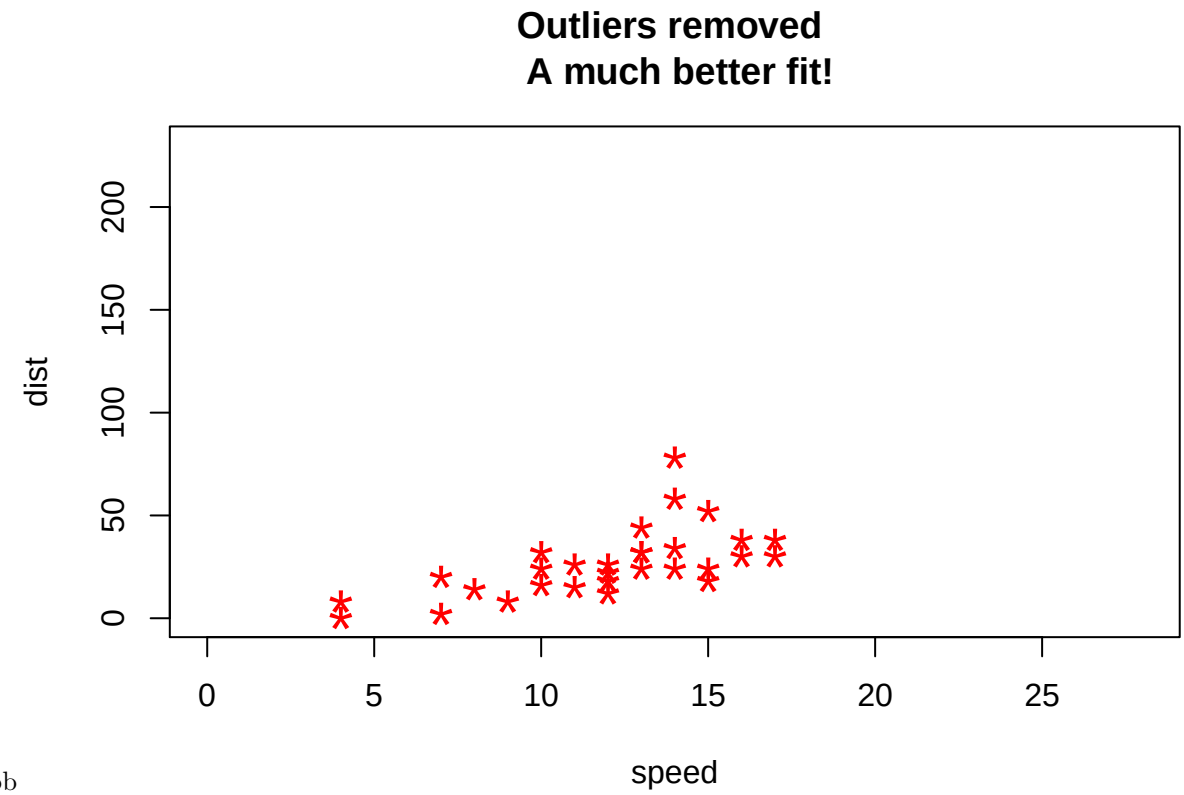
```
# Inject outliers into data.
cars1 <- cars[1:30, ]  # original data
cars_outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218))  # introduce outl
cars2 <- rbind(cars1, cars_outliers)  # data with outliers.

# Plot of data with outliers.
par(mfrow=c(1, 2))
plot(cars2$speed, cars2$dist, xlim=c(0, 28), ylim=c(0, 230), main="With Outliers", xlab="speed", ylab="
```



Detection-1.bb

```
# Plot of original data without outliers. Note the change in slope (angle) of best fit line.
plot(cars1$speed, cars1$dist, xlim=c(0, 28), ylim=c(0, 230), main="Outliers removed \n A much better fi
```



Detection2-1.bb

Detection of Outliers is performed using:

- Univariate Approach
- Multivariate Approach
- Multivariate Model Approach

5.2 Data Explanation (Like, Storytelling)

We can use visualizations to communicate impactful stories to our audiences, which might lead to some desired action. But, creating such a visualization requires some serious thought and careful consideration. Some key questions to think about are:

5.2.1 What makes a chart effective?

Data visualization is a combination of art and science. When it comes to the artistic aspect, there are no correct answers for doing the visualization. There are many ways to present the data. However, when making sense of facts, numbers, and measurements, a better understanding and effectiveness is promoted by a logical path to follow. To determine the best type of chart is hard for those new to data visualization. Most people learn it by referring to other people's work without understanding the underlying logic, so they don't have the theory in their mind to make the judgment.

Therefore, before we begin visualizing our data, we need to start with the following:

- **Know the purpose**(Kosara 2016) - (Analytical or Presentation): It is important to know the purpose of designing a visualization. In many cases, it is designed to explore or analyze data to enable readers to find insights in data themselves. But there are also cases when its purpose is to present and create awareness about certain findings or even to make a decision. For example, when a journalist creates a visualization for reporting on the current weather situation, the goal there is to mainly present the key

trends and create awareness among the general public. When climate scientists create visualizations for communicating their results to policy makers on climate change, they are mainly calling for actions.

- **Know your audience**(Mekhatria 2017): After we know the why we are designing a visualization, it is important to know who are we targeting with that visual. No matter who your intended audience is, it is important to customize it to their needs, interest, level of expertise and analytical ability. Certain factors like their cultural preferences, expertise level, etc., also play a key role in designing an effective visualization. For eg., colors have a special significance in Chinese culture. They use red to represent a dynamic or/and a positive event, such as growing sales in a region, while in most of the western world blue or green represents positive trends, such as sales revenue, etc. Similarly, a visualization designed for a finance analyst will be different from a visual designed for a marketing manager. Therefore, customization is key in ensuring effectiveness of a visualization.

- **Know the right chart type**(Mekhatria 2017): Once you know the purpose and have identified the target audience, it is important to choose the right chart type. Choosing the right visual, which could be a chart, map, table, dashboard or infographic, ensures that it resonates well with your audience. Also, it empowers the readers to explore the data, identify insights and make decisions after evaluating different scenarios.

After answering these questions, you should be able to get a better image of your ideal graph. The simple guidance for using the different types of the chart is - line charts for tracking trends over time, bar charts to compare quantities, scatter plots for a joint variation of two data items, bubble charts showing the joint variation of three data items, and pie charts to compare parts of a whole. However, let's delve deeper into the various presentation styles and types of common charts.

5.2.2 How to decide which chart type to use?

While it is possible that data can be visualized using multiple charts, however, it is important to choose the 'right' chart type that clearly and accurately communicates the key message by separating the noise from the data. Remember, data is only valuable if you know how to visualize it and give context. (Infogram, n.d.)

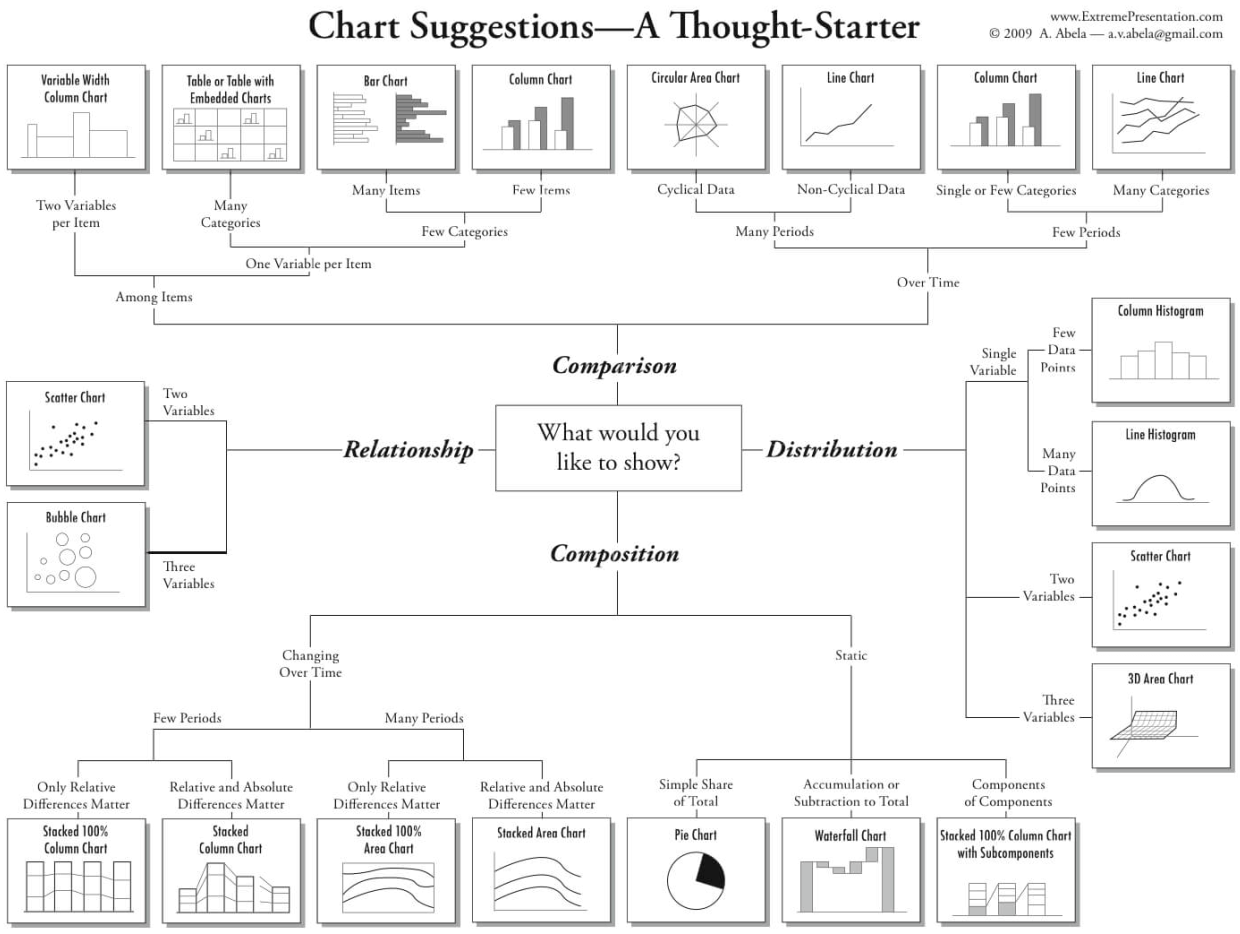
There are four basic presentation types that you can use to present your data:

- A **Comparison** chart sets two variables against each other and displays the interaction between those two variables. For eg., a line chart displaying the variation of online sales across different months during a given time period.
- A **Composition** chart displays how individual parts make up the whole of something. For eg., a pie chart displaying the market share of a phone company by region.
- A **Distribution** chart helps to understand outliers, the normal tendency, and the range of values in the dataset. For eg., a column histogram displaying the distribution of grades on a school exam.
- A **Relationship** chart tries to show a connection or correlation between two or more variables. For eg., a scatter plot displaying the relationship between marketing spends and sales revenue.

To determine which chart is best suited for each of those presentation types, first you must answer a few questions(Jānis Gulbis 2016a):

- How many variables do you want to show in a single chart? One, two, three, many?
- How many items (data points) will you display for each variable? Only a few or many?
- Will you display values over a period of time, or among items or groups?

After you have answers to these questions, you can refer to a chart selection diagram created by Dr. Andrew Abela that should help you pick the right chart for your data type.



5.3 Tips to Improve Data Visualization

(French 2017), (Steier et al. 2012), (Evergreen, Stephanie;Metzner, Chris 2013) In order to design impactful visualizations, it is important to keep in mind certain rules which create a stark divide between - **effective and ineffective** visualizations. Some of these rules are:

5.3.1 Comparison

Include a zero baseline if possible. Although 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.

5.3.2 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 the chart and graph headers simple and to the point. There's no need to get clever,

verbose, or puntastic. 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.

5.3.3 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 saturation 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.

- Tips for Color in Visuals

Use Case	Tip	Rationale
Numerical Scales	Color for numerical scales should be used with caution.	The way you interpret a shade depends on the colors around it and sometimes it can lead to false conclusions.
Color Associations	Color can be used to leverage long-term memory very quickly.	We automatically associated strawberries with red. If we can leverage the how people associate different colors with different things, we will not even need a legend to explicitly match color to meaning.
Highlights	Bright colors can be used to highlight a certain part of the data.	Alarming colors draw the eye quickly to areas that need attention.
Color Combinations	Use contrasting dark and light colors. Combinations such as red–green or blue– yellow should be avoided.	This will cause difficulty for people with color blindness.
Choice of Colors	Match the content of a color with the meaning in readers' culture.	This helps readers understand the graphs quicker and easier. For example, green for forest and blue for lake, and red for Repulicans and blue for Democrats. But do avoid stereotypical colors, such as pink for women and blue for men.
Number of Colors	Do not use more than 6 colors in a single layout.	Too many colors do not help readers distinguish categories easily, instead, readers will be confused. If more than six colors are needed, we should consider using another type of chart or categorize groups together (Rost, 2018).

(Jager 2019)

5.3.4 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).