Week 5: Data Visualization

5.1 Engagement: Introduction to Data Visualization

Data Visualization

Data visualization is one of the primary skills of any data scientist, hence value cannot be understated. However, it's also a large field in itself. There are a number of graduate courses at universities and design labs around the country just focused on data visualization. So, this week is really just an introduction to a much, much broader topic. For the introduction,

- we'll first spend some time looking at data visualization conceptually,
- then we'll dive right into making some visualization using Python libraries.
- Lastly, we'll take a look at some particularly effective data visualizations and their impact.
- I will also mention that if this area is interesting to you and you want to lean more, we'll be sure to provide links to textbooks and other resources on data visualization.

Role of Visualization

In this video, we'll be talking about data visualization, and its role in data science. So at the end of this video, you should be able to define data visualization, and begin to understand its important role in the data science process.

Now that we're facing the explosion of data that Ilik ai talked about on week one, what matters are the skills associated with handling all this data. Let me just read you a quote from Google's chief economist back in 2008. "The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to comminucate it, that's going to be a hugely important skill in the next decade. Because now we really do have essentially free and ubiquitous data.

- First, I like this quote because it points out well before big data and data science became buzz words, the need for data scientists.
- In his full interview, he makes a clear case for statisticians as well. As you know, statistics is one of the core areas of data science, and using statistics is one way to try and understand data. I can't understate how important statistics is, but you'll be learning more about statistics in the next course.
- So back to his quote, the other part of understanding data is visualization. In fact, visualization is part of much of the process that he describes. Processing data and extracting value from it or understanding it often requires visualizing it to gain that kind of insight. And we want to communicate your results, you'll often want to use data visualization to do that too. So visualization is central to the skills that he describes.

总结:

Defining Visualization

"The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." [Card et al., 1999]

"The representation and presentation of data to facilitate understanding." [Kirk, 2016]

- Card, S. and Mackinlay, J. and Shneiderman, B., Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers, 1999.
- Kirk, A. Data Visualisation: A handbook for Data Driven Design. SAGE publications, 2016.

There are two definitions of defining visualization that I particularly appreciate.

- The first, "the use of computer supported, interactive, visual representations of abstract data to amplify cognition."
- And the second, "the representation and presentation of data to facilitate understanding."

Both discuss the importance of representing and presenting data. Each of these is central to data visualization. But what I like the most about these two quotes is actually that end in the same key part of data visualization. Improving how we understand and think about the data.

If we think about it, a pretty large chunk of our brain is dedicated to vision. And we've got some really powerful processing associated with that vision in our heads.

In contrast, we're not very good at deciphering raw data. If you give my 100 raw XY coordinates, it could take me hours working by hand to see what they mean. But if you give me a plot with 1000 points on a trendline, I can often tell you in seconds what the data is telling us. That's the power of data visualization. In fact, let's do a couple of quick experiments.

Example 1 – raw data

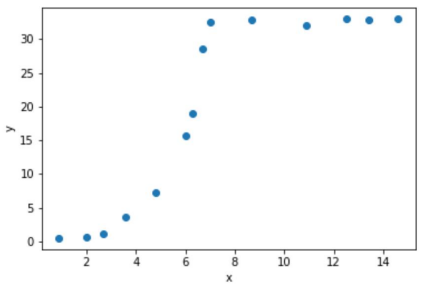
X	Y
0.9	0.5
2.7	1.1
6.7	28.6
10.9	32.8
6.0	15.7
6.3	19
7 .0	32.6

X	Y
8.7	32.3
4.8	7.3
12.5	33.1
13.4	32.9
2.0	0.75
3.6	3.6
14.6	33

	X	Y
Median	6.5	23.8
Mean	7.2	19.5
STD.DEV	4.2	13.6

Correlation = 0.88

Example 1 – Visualized



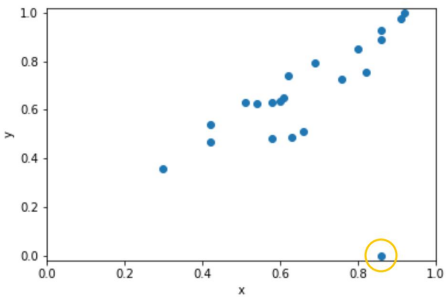
All these data points are from one data set. Can you see the relationship between X and Y? Take a few seconds. Here's some stats about X and Y, do they help? Not really. Correlation might help a bit, as 0.88 implies these values are correlated, specifically that they may have a linear relationship. We've done a lot of work and we still don't completely understand what's going on here. But now we do. It looks like there's exponential growth for Y with respect to X, from the start, zero, until it hits around X's seven. At that point it hits some ceiling where Y is 33, and increasing X no longer increases Y. Notice that correlation was actually a touch misleading, as it implied a linear relationship when the data on appearance doesn't apply that. But the key point is that human visual cortex is really talented.

Example 2 – raw data

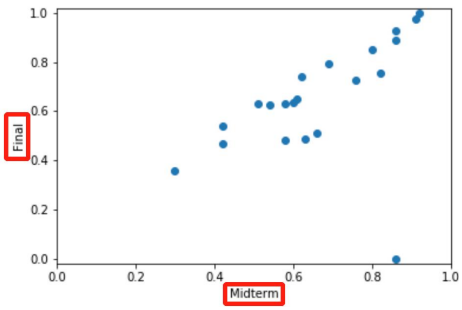
X	Y
0.42	0.46750
0.54	0.62500
0.42	0.53750
0.86	0.92750
0.60	0.63750
0.51	0.63125
0.30	0.35625
0.61	0.65000
0.58	0.63125
0.76	0.72500
0.58	0.48125

X	Y
0.66	0.51250
0.63	0.48750
0.92	1.00000
0.86	0.88750
0.91	0.97500
0.82	0.75625
0.86	0.00000
0.80	0.85000
0.69	0.79375
0.62	0.74000

Example 2 – Visualized



Example 2 – Visualized



Let's do this one more time. Here's a new set of data, all these points are from one data set. If you focus on these for a while, you may begin to notice that the X and Y values seem pretty similar with some noise. But again, let's see if we can understand it better in a graph. Ah ha. The scatter plot shows that these are really correlated, higher X values imply higher Y and vice versa. In fact, correlation as a statistic would have been helpful here, and would have given us the value 0.5. But the value is lower than you might expect **because of the outlier**. And visually, you'll probably notice that there's a strange outlier right here. But I don't know what to do with that information right now. With just the raw data points, I don't know if this is an error in the data or what caused that outlier. I'd like to dive into that value more, but I'm stuck without context. And that's the other point I wanna make here. Without a context, we don't know what to think about the data, or this outlier. In fact, without context, I'm not sure if I even care about this data. So we need context. So turns out that these are pseudo-hypothetical midterm and final exam scores for each student in a class. Now that we know that, we can recognize that performance in the midterm is pretty closely tied to performance on the final. For instructors, this information might lead them to try to intervene for students who perform poorly on the midterm, and for the outlier, context helps as well. What happened with the student? You could figure out which student supplied this data point and let's say after looking into that student, you find out that they didn't take the final. What it turns out then, is that when you're cleaning your data, the missing entry for this student was made to be a zero. But a zero on the exam and not taking the exam are two different things. If you remove the outlier due to missing data, you gain a better understanding of the data, and by the way, the correlation will now be a 0.89, without that outlier. Thinking more about that outlier, if we consider the correlation for

总结:

the other students, we could infer that if this student were to take the final, they would have likely done quite well. But we'll get more into predictions at a later week. I also want to point out that not taking a final exam is odd for a strong performing student. So if I were their instructor, this data would cause me to inquire about their wellbeing, and I'd actually do that for any student missing their final exam. So these results tell me what I might be able to predict for an outcome, and to take action regarding that outlier.

The key point is that visualization of data means nothing without the context of the data. When we didn't know what X and Y were, we didn't really care about the data, context is what makes us value the visualization, and gives us the ability to dig deeper into the data, looking to explain outliers, and also to be able to cone to conclusions about the data itself.

Now that we have these working definitions of data visualization, let's talk about the ways that we use visualization to conduct data science.

Types of Visualizations

By the end of this video, you should be able to:

- Understand the different ways data visualization is used in data science

Two key categories*

Conceptual or data-driven

Declarative or exploratory

Visualizations play a numberof different roles in data science. And the role that they play changes how we'll handle the visualization. By the end of this video, you should be able to understand the different ways that data visualization is used in data science.

There are two key ways to categorize data visualization.

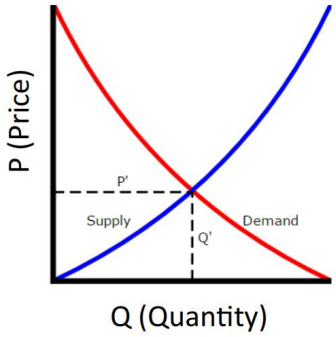
- The first is whether this is conceptual or **data-driven**. Most of what we'll focus on this week and in this course is data-driven. But visualization of **concepts** is important as well. Particularly when we aim to explain how things work conceptually. For example, economists may seek to visualize the notion of the classic supply and demand curve without using real data. And then back it up with data supporting the concept from, say, Uber surge pricing.

- Because data-driven is where we spend much of our time as data scientists, let's look at the second categorization in that context.
    - **Declarative**. There's a point when we've analyzed the data, and we have data supported conclusion we wish to articulate to our audience. At the point of presenting, we want our visualization to convey this conclusion to the observer in the most straight forward possible manner.
    - **Exploratory** (探究的，勘探的). We spend much of our time exploring data, and visualization plays a key role in that. Visualizations often encourage us, and enable us to look deeper into the data.

Let's take a look at some examples from these categories.

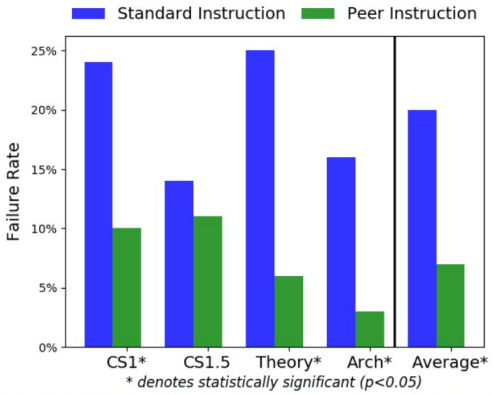
总结：

Conceptual: Declarative



https://commons.wikimedia.org/wiki/File:Supply_and_demand_curves.svg

Data-Driven: Declarative

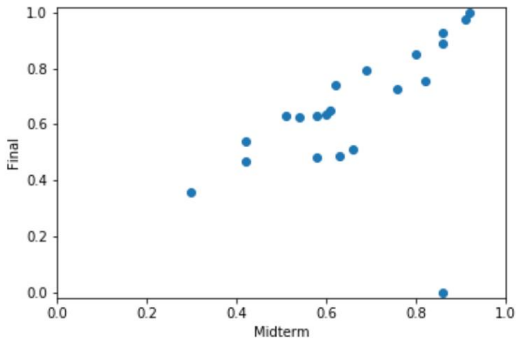


Porter, L., Bailey Lee, C. and Simon, B. Halving fail rates using peer instruction: a study of four computer science courses. In *Proceeding of the 44th ACM technical symposium on Computer science education*. March, 2013.

So this is the classic supply-demand curve from economics that I just mentioned. It explains the relationship between supply, or the quantity of goods available, with demand, or the amount someone is willing to pay for the item. When quantity is low, demand is high. As quantity goes up, demand will decline. This is conceptual, and these lines are hypothetical. But since much of our work is going to be data-driven, let's focus on that for now.

Here I'm taking a declarative example from a research paper of mine where we show the peer instruction, that the peer instruction pedagogy (教育，教育学) resulted in a significant decrease in failure rates for students, relative to standard instruction, in computer science classes at UC San Diego. My goal in presenting these results was to convey to the reader the impact peer instruction had on failure rates in our classes. Without diving into the details too much, we had failure rates from a number of classes at UC San Diego, where instructors either taught using the active learning pedagogy of peer instruction, or using standard lecture-style pedagogy. If you're interested, please read the full paper for a complete explanation of our experimental mental methods, and possible limitations of these results. As good science will always detail potential limitations of results, along with study methods, I hesitate to use the word declarative here, as it doesn't quite seem right. Rather, I prefer the term presenting. We knew the result that we had, and we were aiming to present it to the audience in a clear and useful manner.

Data-Driven: Explorative



The example of the correlation between midterm and final exam scores from our previous video is likely explorative. I might create a plot like this when trying to better understand the relationship between these exam scores in my class. And its results would likely cause me to explore the data more.
With explorative data visualizations, I'm not gonna spend as much time polishing the appearance, so long as I can interpret it, that's fine. Likewise, I'll often want to be able to quickly plug in different parts of my data set into the figure to explore different relationships. Say HOMER scores against the final exam, rather than midterm against the final exam. Exploration is really at the heart of the data science process, which we talked about in week one. When we're finding outliers or trends, we're often using visualization tools. And those visualizations lead us to dig deeper into the data. As we dig deeper, we do the same thing. Looking at data distributions using histograms. Exploring relationships between variables, or seeking other trends. Ultimately, we find ourselves zooming in and out of various parts of the data set as we try to gain a better understanding. And this process of zooming in and out of the data is almost always accompanied by and facilitated by data visualization.

So now that we understand the different uses for data-driven visualization, let's explore metrics of success for those visualizations.

总结：

Key Design Principles

By the end of this video, you should be able to:

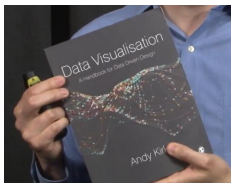
- Recognize qualities of good data visualizations

Principles of Good Design

"Good data visualization is:

1. Trustworthy
2. Accessible
3. Elegant"

- Andy Kirk*

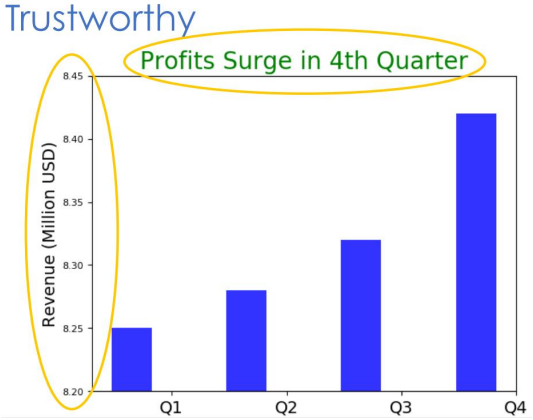


* Kirk, A. Data Visualisation: A handbook for Data Driven Design. SAGE publications, 2016.

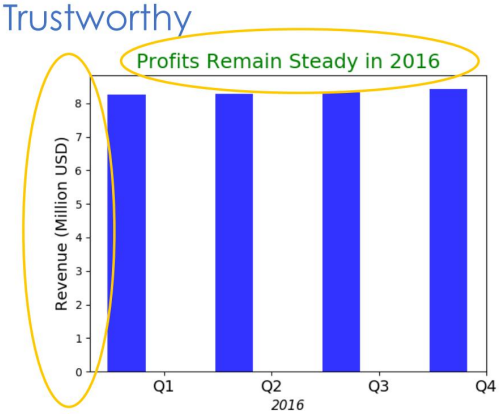
Now that we know what we mean by data visualization, let's talk about what makes a visualization useful.
By the end of this video, you should be able to recognize the qualities of good data visualization.

You know, there're lots of opinions about criteria for good data visualization. By I really like these three put forth by a leading data scientist visualization educator, Andy Kirk. **Data visualization needs to be trustworthy, accessible, and elegant**. Let's talk through all three of these in more detail but, before I do, just a quick word about Andy Kirk's book which I've mentioned a few times already. This book is a great reference for all things data visualization, and I highly recommend it if you wish to dig deeper into this area of study.

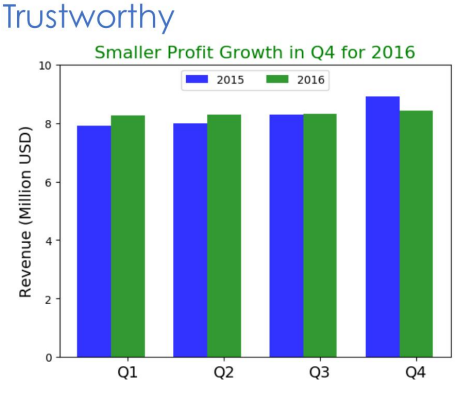
Trustworthy



Trustworthy



Trustworthy



By saying a visualization is **trustworthy**, we are saying that the data presented is honestly portrayed.
For example, if you're displaying something in such a way that it implies a relationship or a trend or a correlation, there should be evidence for such relationship in the data, otherwise you're just misleading your audience.
Let's look at a quick example. Here's a hypothetical figure you might be shown at a business meeting. Take a little, and see if you can spot what's wrong with it. If you noticed the y-axis, you realize this isn't a surge at all. The y-axis is zoomed into just a small part of the graph. We're looking at a roughly 2% increase between Q1 and Q4. That's hardly surging. I suspect the author of such a chart of trying to be dishonest or, at the very least, misleading. They're trying to convince me of something which isn't true. Let's just look at all the flaws here. Surge? Two percent is a surge? The title's simply misleading. We already noticed that the y-axis is being zoomed in on to exaggerate the growth between Q1 and Q4, but did you notice how the font size on the y-axis seems to have been made intentionally small just so it's harder to read. Lastly, the fourth quarter's when holidays occur. I'd expect to see a rise in profits for many industries during the fourth quarter.

A more honest graph might also plot prior years' profits to compare against. So, yikes. By the time I'm done reading this graph, I simply do not believe the author anymore. **All credibility's gone**. So let's see how we could've done this better. Okay, in this figure, we setup the y-axis honestly. And this actually looks like pretty steady revenue over each quarter of 2016, so I fixed the title to reflect that. If this is the only year for which we have data, **the story might be over**, but, if you have data from prior years, we should plot them to see how things look. Supposing we have prior year data, if I plot last year's quarterly revenue, I see that there was a surge in profits in 2015 which is absent in 2016. In 2015, the company had a 12% boost between Q1 and Q4 which is less than 2% in 2015. We could've titled this in a different ways. We could've just said that our revenue was steady in 2016, and that's not misleading. Or we could focus in on the fact that we didn't see a surge in Q4. Either way, we want to dive into the data more to figure out why we didn't see the revenue boost in Q4 of 2016 that we saw in Q4 of 2015. And we do that prior to presenting these results, so we have good answers of why this happened. So after fixing the presentation and pulling in more hypothetical data, we've actually reversed our initial misleading conclusion.

总结：

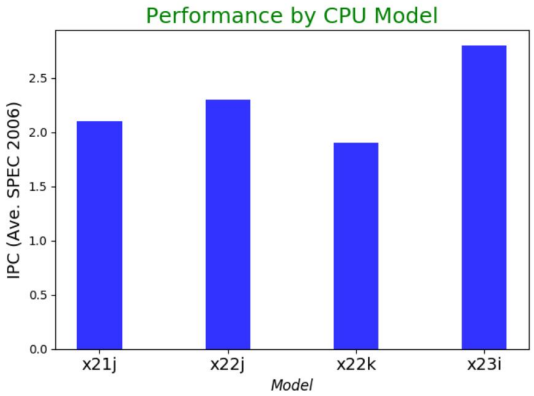
Trustworthy

- Trust is hard to earn, easy to lose.
- Honesty and integrity should be everywhere in the data science process

So there are two key takeaways here which help echo some of the messages you heard in week one.

- First, you need to take trust seriously. People who are looking at your results are trusting you not to doctor the data or misrepresent results. They want to take action based on your findings. To make this even more concrete, as a scientist, I often review papers for publication. If I find one seemingly intentionally misleading claim or figure, I call into doubt everything the author said. And if someone similarly tries to mislead you with poor data visualizations, you shouldn't trust them either.
- The second point is just a reminder that honesty isn't limited to the visualization stage. Honesty has to be everywhere in data science. And this can be hard at times. You may have your own beliefs about what you'll find in the data. In fact, we could have an entire separate course on how to combat human psychological bias. But for now, it suffices to say that you should recognize the biases you have before you look at the data and do your best to have the data itself, not your bias, drive your inquiry.

Accessible



Accessible

- Know your audience
- Understand the purpose of the visualization

Our second key principle is **accessibility**. Accessibility to me is about focusing on your audience and their ability to use your visualization. Let me give you another example. This is a hypothetical example from one of my areas of research, computer architecture. This graph could be useless or great depending on your audience.

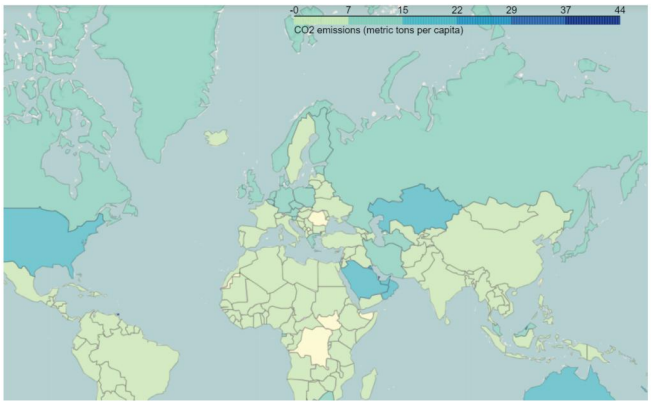
- For a computer architect, the y-axis is reporting instructions for cycle, or IPC, which is one of three major components of execution time, and the inverse of execution time is often used a measure of performance. We're reporting average IPC for an accepted benchmark suite, SPEC 2006, which is used by researchers in computer architecture. If these processor models were real, it's possible I'd know a lot about them and know how they differ architecturally. That'd give me intuition about why these results differ by model. So to someone who knows these processor models, this figure could be honest as there may be a compelling reason to report only IPC, ignoring the other two elements of execution time and performance. To an expert, this figure could be both honest and accessible depending on their use. Notice I said depending on their use.
    - If the goal of the figure was to show me the IPCs by processor model, this works great.
    - But as an expert, I might want a more complex figure which helps to explain why the x23i has such higher IPC than the others. So this figure fails if the intent to is not only report IPC but to also explain why we found these differences. Of course, that might be a separate figure, or it could just be part of the knowledge of how these processors or models differ.
- But for someone not in computer architecture, this graph means virtually nothing. You don't know what these models of processors are, so you don't know why they may perform differently. You likely don't know what IPC is or even what SPEC 2006 is. So for a non-expert, this figure is clearly not accessible. It's unreadable. It has no relevance to them. And if a viewer doesn't know that IPC is just one component of performance, this could be unintentionally misleading.

So the two main things I'd stress are to know your audience and how they perceive the information. Understand what they understand, and know how they might interpret the result. Also be sure to know what the purpose is of your visualization. Are you exploring the data? Or presenting it? That helps you craft it in the appropriate way.

One last point. Take into account the expected time for the audience to read and understand the result. That depends on whether this is a slide you might show for one minute at a presentation or a figure that you're sharing with a colleague who might spend some time really diving into it. These questions about your audience and how the visualization will be used allow you to craft it appropriately to make it accessible.

总结：

Elegant



Elegant

- Focus on the relevant
- Be stylish if possible
- Think about decorations

What do you think of when you hear the word **elegant**? You might think about style or grace. You might think about something clear and aesthetically beautiful. These all reasonably apply to good data visualizations. I would note that in practice, I put a lot more time into elegant visualizations when I'm presenting results. When I'm exploring data, it's nice to have, but it is no means critical. I like to think of it this way. If my graphic is gonna be front and center on the New York Times, it better be perfect. If it's in a slide deck for teaching, it should be really solid. And if it's for me when I'm exploring the data, it should be acceptable. At the very least, the lack of elegance shouldn't get in the way of my interpreting the data. For example, later this week you'll be creating an overlay like this. Take a moment to take it in. There are elements of this which are elegant.

- The use of an overlay on a map helps an observer see the different countries quickly.
- The color coding to show numeric data helps the viewer quickly interpret results.
- The color scheme goes towards darker blue with more CO2 emissions per capita, so that might not actually be perfect. More CO2 emissions per capita is arguably a bad thing, so a different color scheme which conveys that CO2 emissions is bad might actually be better.
- There isn't unnecessary other data here. I didn't put the numbers over each country as that'd likely distract more than help. I didn't try to combine with some other measure like GDP per capita in the same figure as it'd likely be too much to process but might make a really nice complementary figure.
- If I added any additional decorations like maybe a smokestack on high CO2 per capita producers, I'd be sure it added to the figure rather than detracted. But this figure, I thought it actually detracted more than it added so I left it off.

Overall, style and beauty is subjective, but I think when we look into our case studies at the end of this week, you'll be able to appreciate the elegance of some of their designs. We'll also give you links to websites and talks which were exceptionally elegant visualizations as well.

For elegant visualizations, you should focus on what is relevant and remove anything that isn't adding to the figure. You're trying to make the design invisible so that the viewer can take as much from the visualization as possible without being distracted. Now this isn't the same thing as **minimalism**, but I've heard some folks argue for. I feel minimalism doesn't just remove the unnecessary but often starts remove things which are helpful. So there's a balance to be struck with how much to include. Think about your style. For those of your who read Nate Silver's website, 538, you know there's a distinct style to his presentations and graphs. To me, I really like that style, and his website's success, I think, is likely tied at least in part to how recognizable his style is and presentation. Decorations may seems contrary to honesty, and, in some cases, they kinda are. In scientific papers about blood donations, I'd be surprised to see the bars in bar chart intentionally made to resemble flowing blood. But if the graphic is put out by a charity encouraging folks to give blood the week before Halloween, decorations may bring the visualization extra attention. And this just goes back to knowing your intended audience.

If this were a data visualization course in a design lab, we could have spent much of the course just on exploring these principles in depth. Our hope with this course is to just give you some concrete goals that you can aspire toward when you're creating your visualizations.

Which is NOT a quality of good data visualization, according to Andy Kirk?

RESULTS

Trustworthy	2%
Accessible	5%
Elegant	9%
Meaningful	83%

Submit

Results gathered from 447 respondents.

总结：

5.2 Engagement: Matplotlib and Other Libraries

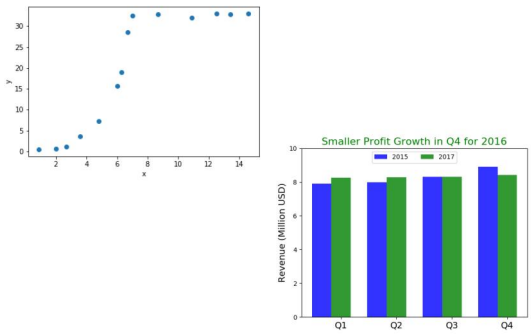
Matplotlib

By the end of this video, you should be able to:

- Explain the role of Matplotlib in python

What is matplotlib

- Plotting Library for Python



In this lesson we'll be focused on Matplotlib. So by the end of this video, you should be able to explain the role that Matplotlib has in data science. Matplotlib is a widely used plotting library for Python. It's often my go-to for creating quick plots of data and almost all the plots I showed you in the last lesson were made just using Matplotlib.

Why matplotlib

"Matplotlib tries to make easy things easy and hard things possible."

https://matplotlib.org/

Other libraries?

- Seaborn
- ggplot
- Altair
- Bokeh
- Plotly
- Folium

I value Matplotlib because it's easy to produce quick plots of data when I'm doing data exploration and because I find the plots clean and generally aesthetically pleasing. I actually really like this phrase from the Matplotlib homepage.

- The first part, make easy things easy, is what I was referring to when I said it's my go-to for most plots during data exploration. You can produce a bar chart, a line chart, a scatter plot, a box plot or histogram in just a few lines of code. This is fantastic as you can quickly check the relationship between features, plot trend lines, or just get a better feel for the distribution of data.
- The second part is why you'll find yourself often using Matplotlib even when you create complex visualizations, where you want it to customize your plots for presenting your results. One of my favorite stories about Matplotlib is that a colleague of mine took the time to set up scripts for the figures in a journal paper that we published. The setup took a bit of time, as he wanted to have an elegant design, and he wanted the fonts to mesh seemingly with the paper text. If you've read research papers, you'll undoubtedly have seen a figure which just stands out from the paper in a bad way.

One reason why figures stand out in this way is that all the fonts are completely different in the figure from the paper. Another reason is that if you're using LaTeX for your paper, the figure may seem less elegant than the paper. So his work aimed to remedy these common problems. And what was so great about his work was that because it relied on Python scripts reading CSV files and Matplotlib, we build the scripts into the building of the paper using a make file. So we'd update the CSV with some new results, we'd hit build, and we'd get an updated paper with these really clean figures. And all of this is made possible by Matplotlib.

In the larger context of Jupyter notebooks and the landscape of Python tools, you may know that Python is a, that Matplotlib is a fairly low-level tool for plotting. There are a ton of other tools out there, including Seaborn, ggplot, Altair, Bokeh, Plotly, and Folium, among many others. For us, we use a library that fits our visualization needs best, based on those needs. So we'll have a reading later in this lesson which outlines when you might wanna use one library versus another. And we'll keep updating this as new libraries come out, or as libraries add new features. Given that there are so many libraries out there, and we so frequently work with Matplotlib in our everyday work, we'll start by diving into a notebook showcasing Matplotlib. Then we'll have a short notebook later showing you how to work with more advanced libraries.

总结:

World Development Indicators

见本单元 jupyter notebook: 05a_Matplotlib_Notebook

Basic Plotting in Matplotlib Part 1

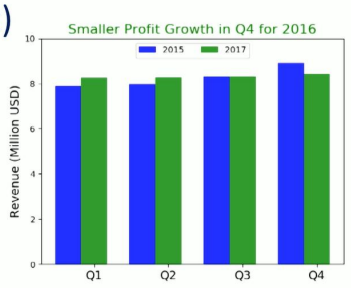
见本单元 jupyter notebook: 05a_Matplotlib_Notebook

By the end of this video, you should be able to:

- Create bar charts, line charts, and histograms using matplotlib
- Recognize the common components in a matplotlib figure

Common Components

- Chart type (bar, line, hist)
- Axes data ranges
- Axes labels
- Figure labels
- Legend
- Aesthetics
- Annotations



That you can do more advanced graphics with MatplotLib. There are more questions I have about this data for now. But I'll leave you to explore them if you want. But before I wrap up the video, I wanna point out, after just seeing a few of these figures, you've probably already picked up on the common features of a MatplotLib plot. So, let's state them explicitly. The common features are the chart type. And here, we have a bar chart. Next, we have the ranges for the x and the y values. The possible labels for the axes. We could also possibly label the figure itself and provide a legend. And then lastly, aesthetics, like font size, line size, plot size. Even more complex things like, annotations, can be added.
So, now that we've seen all these features of MatplotLib plots, I encourage you to use this notebook. Along with the others we provide this week for more examples. But before we finish with MatplotLib, I want to explore the relationship between CO2 emissions and GDP. But let's do that in our next video.

When is it acceptable to avoid axis labels in plots using matlibplot.

RESULTS

When you are simply exploring the data.	70%
When you are unsure of what the labels should be.	7%
When the labels can be easily implied.	23%

Submit

Results gathered from 404 respondents.

总结:

Basic Plotting in Matplotlib Part 2

见本单元 jupyter notebook: 05a_Matplotlib_Notebook

By the end of this video, you should be able to:

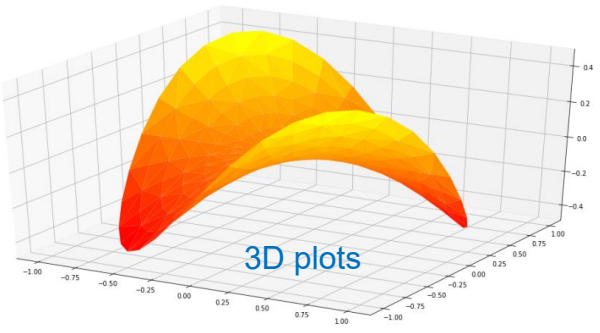
- Create line charts and scatterplots using matplotlib

Matplotlib Additional Examples

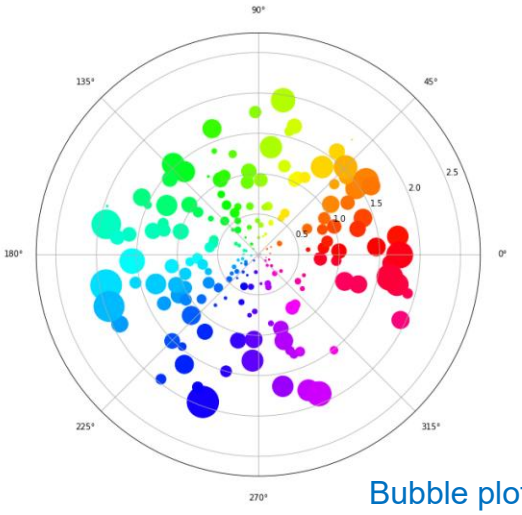
见本单元 jupyter notebook: 05b_Exploring Indicator's Across Countries

By the end of this video, you should be able to:

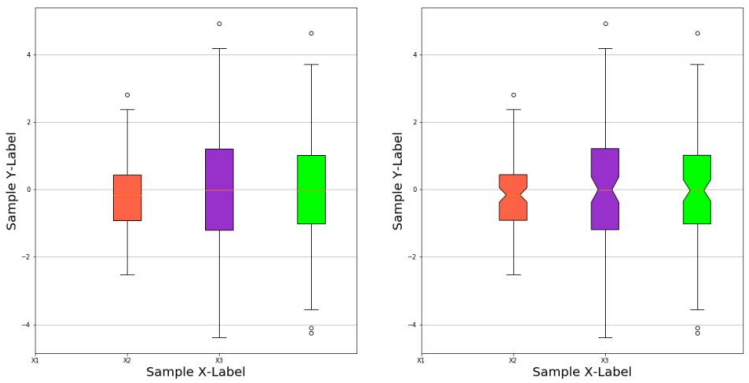
- Use the additional Jupyter notebook for examples



3D plots



Bubble plots



Boxplots

This is a short video introduction to the notebook with additional Matplotlib examples that will be provided for this week. So by the end of this video, you should be able to use additional Jupyter notebook resources as examples. In our additional notebook, we provide code for a number of examples. Still working with the rural development indicators we pick random indicators and compare them against each other using line plots and scatter plots.

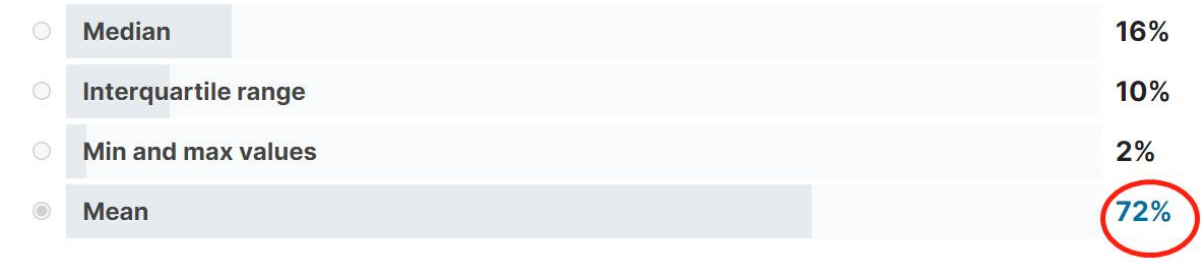
- After working with the rural development indicators, we also provide an example of how to create **3D plots** like this in Matplotlib.
- We also provide an example using bubble plots. **Bubble plots** can be really helpful when you want to graph three dimensions easily. We can have an x, a y, and a size of each point. In this example, we also use color code to provide a fourth dimension. So on this image we have: angle, distance from the center, size of the bubble, and color all coded here. Know that unlike some of our earlier figures, you need to spend time interpreting and understanding the data in a figure like this. But it could be incredibly valuable for data exploration and for conveying more complex relationships when you're presenting your data.
- In addition to histograms, I frequently use **box plots** when I'm trying to understand distributions. This example provides you with a template for boxplots and for placing figures side-by-side in Matplotlib. Boxplots tell you the median, values within the interquartile range, and elements above the third quartile, and below the first. As well as the maximum and minimum values.

That's a lot of useful information all in one figure. So please be sure to check out these additional notebooks, as well as other examples online when you're looking for ways to visualize your data using Matplotlib in your upcoming project weeks.

总结:

Which of the following does a boxplot NOT show you?

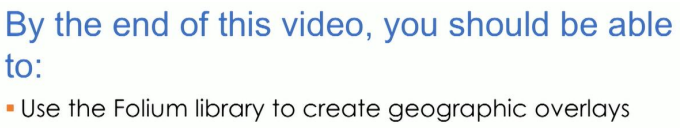
RESULTS



Results gathered from 396 respondents.

Folium Example

见本单元 jupyter notebook: 05c_Folium_Notebook



We mentioned earlier that there are a numberof useful libraries outside of Matplotlib. Particularly when working with a dataset like the world development indicators, creating geographic overlays can be a powerful way of visualizing your data. So by the end of this video, you should be able to use **the Folium library** to create geographic overlays. So jumping into our Jupyter notebook--before you're able to use Folium, you may need to install it on your system. You will also want to grab the json file listed below and place it in your folder under geo/world-countries dot json if you haven't already.

So, we provide this notebook as an example of how to do geographic overlays. But also as an example of how to use additional visualization libraries and how they can be powerful depending on your visualization needs. So please be sure to check out the readings along with our library recommendations on what visualization library might work best for your needs.

总结:

Readings: Visualization Libraries

The following list provides a few plotting libraries for you to get started based on their use case(s). This list is focused on providing a few solid options for each case rather than overwhelming you with the variety of options available.

The foundation: Matplotlib, most used plotting library, best for two-dimensional non-interactive plots. A possible replacement is `pygal`, it provides similar functionality but generates vector graphics SVG output and has a more user-friendly interface.

Specific use cases:

- Specialized **statistical plots**, like automatically fitting a linear regression with confidence interval or like scatter plots color-coded by category.
  - `seaborn`: it builds on top of Matplotlib and it can also be used as a replacement for `matplotlib` just for an easier way to specify color palettes and **plotting aestetics**
- **Grammar of graphics plotting**, if you find the interface of Matplotlib too verbose, Python provides packages based on a different paradigm of plot syntax based on R's `ggplot2`
  - `ggplot`: it provides similar functionality to Matplotlib and is also based on Matplotlib but provides a different interface.
  - `altair`: it has a simpler interface compared to `ggplot` and generates Javascript based plots easily embeddable into the Jupyter Notebook or exported as PNG.
- **Interactive plots**, i.e. pan, zoom that work in the Jupyter Notebooks but also can be exported as Javascript to work standalone on a webpage.
  - `bokeh`: maintained by Continuum Analytics, the company behind Anaconda
  - `plotly`: is both a library and a cloud service where you can store and share your visualizations (it has free/paid accounts)
- **Interactive map visualization**

  *`folium`: Creates HTML pages that include the Leaflet.js javascript plotting library to display data on top of maps. *`plotly`: it supports color-coded country/world maps embedded in the Jupyter Notebook.
- **Realtime plots** that update with streaming data, even integrated in a dashboard with user interaction.
  - `bokeh plot server`: it is part of Bokeh but requires to launch a separate Python process that takes care of responding to events from User Interface or from streaming data updates.
- **3D plots** are not easy to interpret, it is worth first consider if a combination of 2D plots could provide a better insight into the data
  - `mplot3d`: Matplotlib tookit for 3D visualization

总结：

5.3 Case Study

Case Study 1, Cholera

By the end of this video, you should be able to:

- Describe the role played by data visualization in understanding how cholera spreads

Cholera

- Bacterial infection
- Causes severe diarrhea, possibly leading to death by dehydration
- Remains public health threat
  - 1.3-4.0 million cases worldwide
  - 21,000-143,000 deaths worldwide

This lesson will focus on case studies of exemplar data visualizations. In this video, we'll focus on a data visualization which helped the world understand the true cause of cholera. So by the end of this video, you should be able to describe the role played by data visualization in understanding how cholera spread.
As background, cholera is a bacterial infection which can cause severe diarrhea, possibly leading to death through dehydration. After exposure to the bacteria, a person shows symptoms 12 hours to five days later, and acute cases can kill within hours if untreated. Unfortunately, cholera is still a danger today, with millions of cases worldwide and deaths in the tens to hundreds of thousands.

How Cholera Spreads

Photo by Ashley Wheaton, April 2009.

How they thought Cholera spreads

"Doktor Schnabel von Rom"
Artwork by Paulus Fürst, 1656

For those of you who are a bit squeamish, you may want to skip ahead to the video where you see the map graphic. Cholera is particularly dangerous, as it can easily cause outbreaks in areas with poor sanitation. People with cholera have diarrhea, which contains the bacteria for up to 10 days after the infection. If someone ingests that bacteria, they contract the disease. So if sewage contaminates food or water supplies, this can cause an outbreak. Today, we know how cholera is spread because of a scientist who had a theory and a terrible outbreak which occurred in 1854 in London. Now, in London in the 1800s, sanitation struggled, particularly in poor neighborhoods. With the rising population and lack of proper sewage and sanitation guidelines, many folks lived in terrible conditions. They often got their water from wells, but well water was susceptible to contamination.

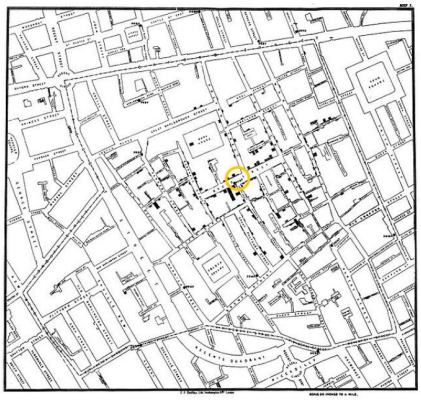
If they didn't know it was spread through water, what exactly did they think caused it? They actually thought it was a combination of smelling something bad and poor constitution. The first was what led to doctors wearing masks, like in this image. The artwork here actually relates to doctors trying to avoid the **Black Death**, but the same theory was still present in the 1800s. Bad smells caused disease, and if you could put some sort of perfume in the mask with you, you could avoid those smells, although we know now that this thinking is flawed. We have the benefit of germ theory. I suspect back in the past, the origin of this idea was likely just a mistaking correlation for causation. Areas where there were people who were ill or had poor sanitation likely had worse odors. Since people got sick in areas with poor odors, they falsely attributed the smell to the source of the disease.

The second idea was a bit more repugnant. At the time, the poor tended to suffer more from the disease. Not surprising, given what we know now about how diseases are spread and the conditions in which they lived at the time. But at the time, some people believed poor people were just somehow weaker people, and by their own weaknesses, were susceptible to disease. Again, this is mistaking correlation for causation. However, this time it's combined with bigotry.

总结:

The outbreak

The Cholera Map



So now that you have a decent idea of what they think caused the disease, that leads us to John Snow, who was a scientist living in London in the 19th century. His background was as an anesthesiologist, where he'd applied scientific principles to make the practice far safer for patients. He'd also published his theory that cholera was water born in 1849, but his theory gained little traction. So John Snow is living in London in 1854, near where the cholera outbreak occurred.
The outbreak was caused by contaminated water entering the well sourcing the Broad Street water pump. After the outbreak, John Snow went door to door in the vicinity of the outbreak and followed up with every **fatal case** to find out where they lived and where they got their water. He plotted all these deaths onto a map.

This is the full map. It's a bit harder to see, but the Broad Street pump is right there in the middle of all those blocks. Each of those little rectangular blocks was a death from cholera. If we zoom in, we see that there are a large number of deaths in the neighborhood around the pump, and away from the pump, the deaths declined. By talking with the families of the deceased, he was able to learn where victims got their water, and in a perfect example of data science, the outliers were critical for finding convincing answers. There was a workhouse or a prison nearby the Broad Street pump with many inmates but almost no fatalities. It turns out the workhouse had its own well and shipped water in from a different source, so they weren't using the pump. There was a brewery nearby, which had also escaped cholera. Again, they had their own water supply or they drank the liquor that was produced in the facility, so they weren't drinking from the pump. Lastly, there was a woman and her niece who lived some distance away from the pump and had died of cholera. With some investigation, he found out the aunt had previously lived in the area and liked the water from the Broad Street pump so much that she had it delivered to her and often shared it with her niece. The combination of this figure and the gathered data that John Snow had put together led town officials to shut down the well by removing its handle.

Despite being skeptical that the pump was the cause, the number of new infections almost instantly declined and were quickly stopped. So even though this outbreak was stemmed by the action of taking off the handle from the pump, it took some time before John Snow's ideas were widely accepted. One key piece to this puzzle was provided later by a reverend in the area named Henry Whitehead. He had sought out to prove John Snow wrong, but in the process, he stumbled on the answer to the mystery of how the pump got contaminated. A baby had become ill from cholera in a house near the pump shortly before the outbreak. The parents had cleaned the diapers for the baby into a cesspool near the well, which then seeped into the well, contaminating it. Another key find related to cholera came some years later when a German physician, Robert Koch, isolated the bacteria which causes cholera. He did that in 1883. These epidemics and the realization of its cause led to Europe and the United States adopting water sanitation.

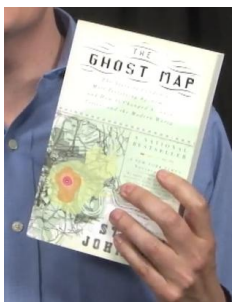
Snow, for his work, is often regarded as one of the pioneers of the field of epidemiology. I would note that well much of the developed world has access to clean water, the story of cholera is, sadly, not over, as my numbers at the start of this video told you. The World Health Organization has estimated that almost 80% of people in third-world countries lack access to clean water supplies, and in many of those areas, there's no sewage treatment, either. Despite our knowing how to prevent cholera, it remains a global health concern.

To end things on a positive note, I want to share a picture of me next to the Broad Street pump in London in 2013. I was there for a conference and appreciating the science and the data visualization in the story around the Broad Street pump. This was actually one of my first stops of that trip. I also want to end with a quick plug for the book The Ghost Map, the story of London's most terrifying epidemic, how it changed science, cities, and the modern world, by Steven Johnson. Much of my knowledge about this story came from reading this book, and when I was teaching the interdisciplinary course on scientific literacy and disease at Skidmore College, we used this book as required reading. So I really can't recommend it enough.

True or False: Outliers can be critical to finding convincing answers when analyzing data.

RESULTS

True	92%
False	8%



总结：

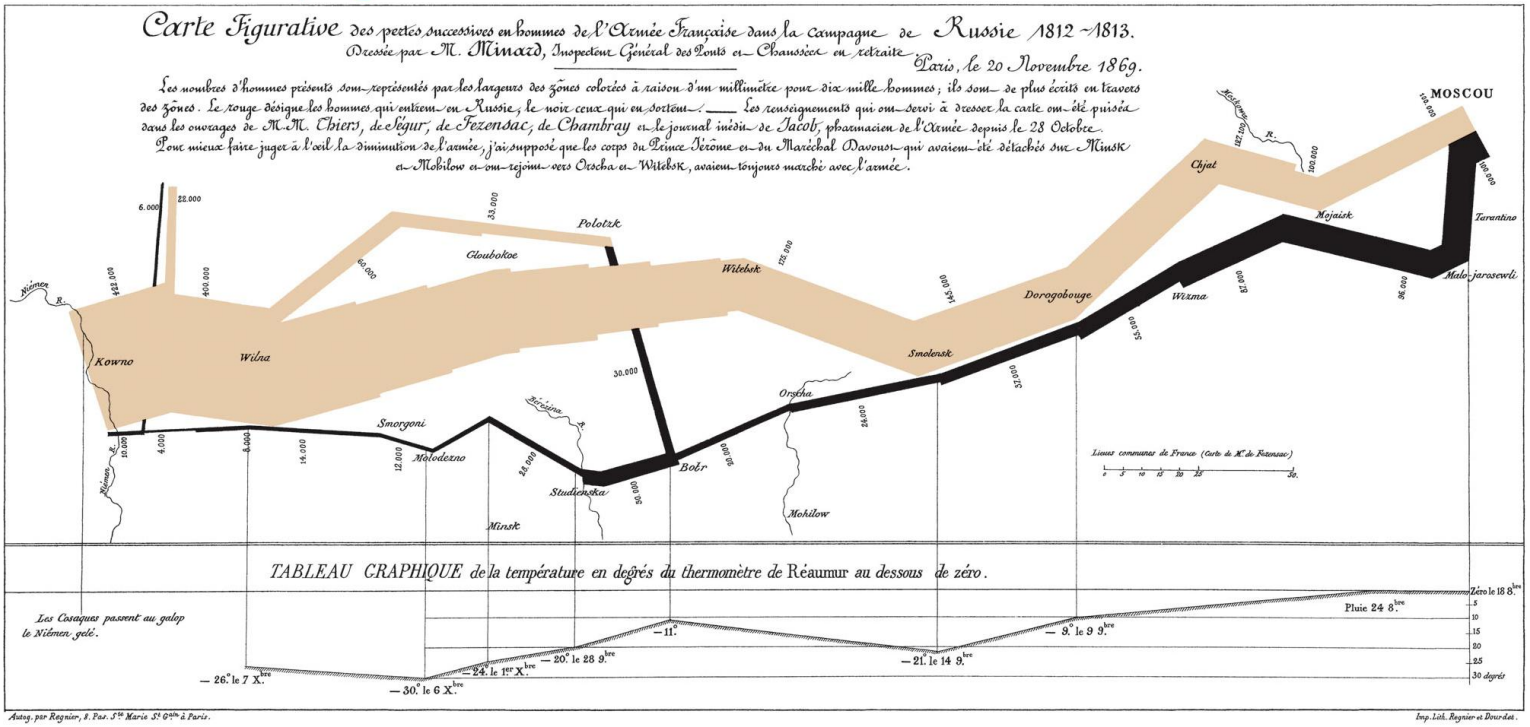
Case Study 2, Napoleon's March

By the end of this video, you should be able to:

- Explain why the data visualization of Napoleon's Russian campaign in 1812 is particularly effective

In this video, we'll be looking at another famous visualization. This one was used to depict France's Russian campaign in 1812. So by the end of this video, you should be able to explain why the data visualization of Napoleon's Russian campaign in 1812 is considered so effective.

The Russian Campaign of 1812

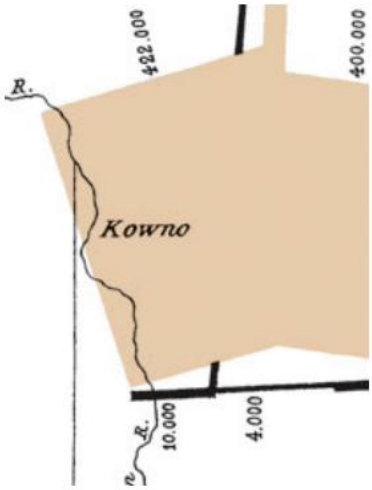


https://upload.wikimedia.org/wikipedia/commons/2/29/Minard.png

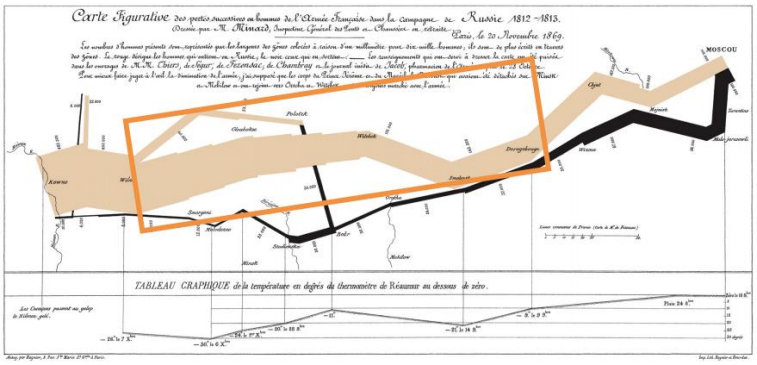
Charles Joseph Minard created this figure in 1869. The title of the figure is, when translated to English, Figurative Map of the Successive Losses in Men for the French Army in the Russian Campaign, 1812 through 1813. Minard was a successful civil engineer in France. He was considered an early leader in using visualizations in engineering and statistics. He is most well-known for creating this figure, which charted that Russian campaign.

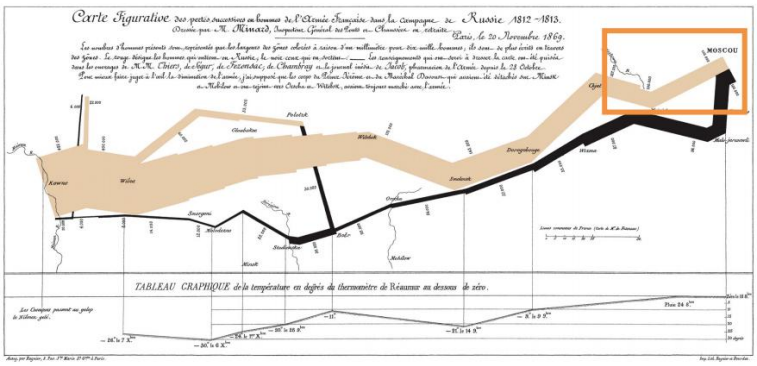
- In orange, you have the French Army, as it advanced from the crossing of the Neman River towards Moscow.
- In black, you have the French Army in retreat. The width of the line depicts the size of the army. You could zoom in quickly to the Neman River, where you'll find the army's size at the start and the end of the campaign. You can see almost instantly how disastrous the invasion of Russia was for the French. They started with around 450,000 troops, although some estimates are actually higher, and they ended it with barely 10,000.

So how are those losses suffered? Let's briefly talk through this campaign using the figure.



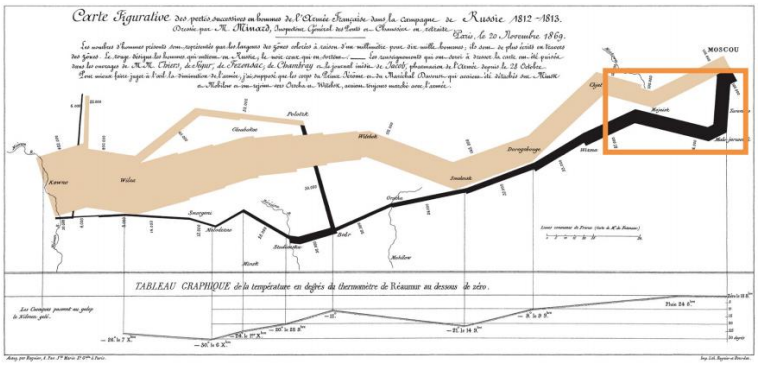
总结:

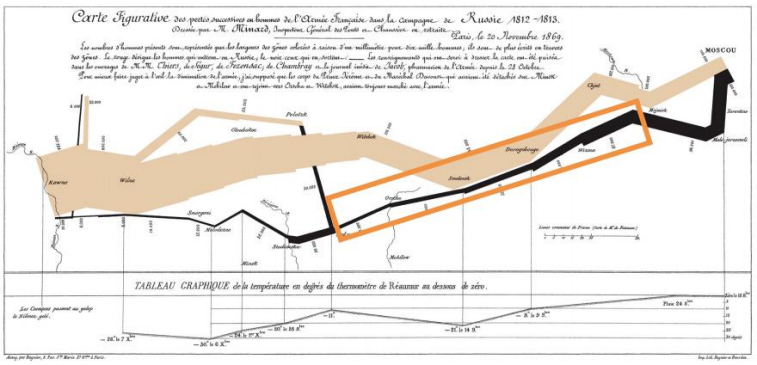




So first, what started all this? In 1812, Russia, under Czar Alexander the First, had been blockading British goods as allies of France. They were blockading British goods under Napoleon's edict that barred trade with England. However, the banning of trade with England was having a disastrous effect on the Russian economy, so Russia decided to end its participation in the embargo. Napoleon, after negotiations with Russia to restore the embargo failed, began his campaign to invade Russia in 1812. The plan was to attack the smaller Russian Army near the Russian border, crush it in battle, and end the war. That was the plan, however the Russian Army refused to engage in Napoleon's huge army in direct battle. As the Russian Army fled back to Moscow, they would destroy the countryside, causing the French Army to fall short on supplies. You can see that in the highlighted region of this figure here. In addition, Russian Cossacks would perform small hit and run attacks on the French Army, both killing troops and also lowering morale. Without engaging in any battle, the French Army was suffering huge losses mostly due to sickness and desertion.

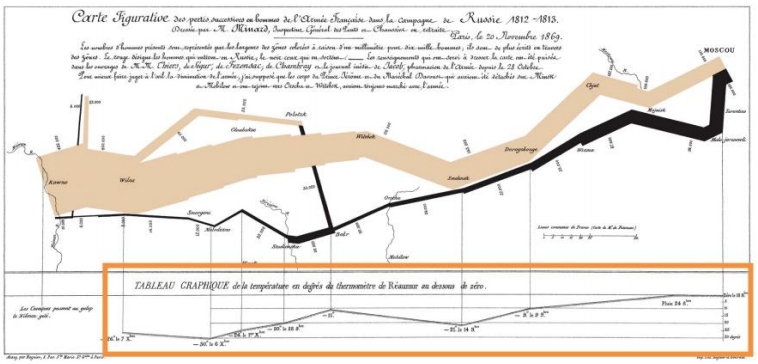
Napoleon finally got the battle he desired at the village of Borodino outside of Moscow. The battle was brutal and ultimately France prevailed, but it did so without the decisive victory desired. On September 14th, the French Army moved into Moscow, only to find it deserted and shortly, much of it was on fire. Napoleon had expected a Russian surrender when he entered Moscow, but he would never receive one. After waiting a month in Moscow to negotiate surrender, snow flurries began to fall. Napoleon realized his army could not **weather the winter** in Moscow.

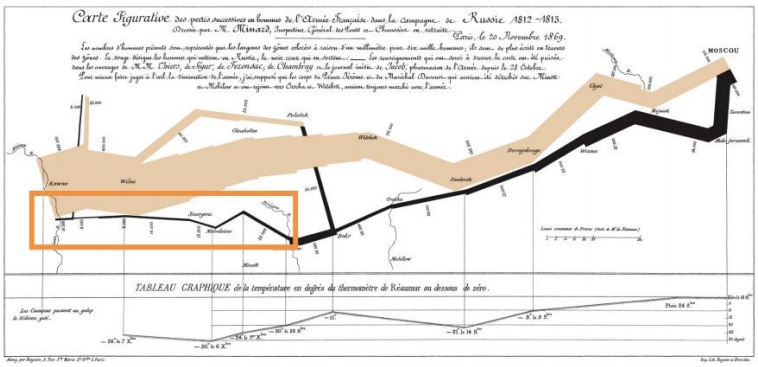




After leaving Moscow, the French Army was attacked by the Russian Army, which had received reserves and was now able to actually go on the offensive. Their presence forced Napoleon to flee using a northern route, which was problematic because winter continued to set in. Moreover, the Russian forces were able to attack the French Army as they continued to flee.

Conditions during the retreat were terrible.The troops lacked supplies, they were being attacked while they were fleeing, and they had to endure an early and harsh winter.





On this figure are the temperatures the troops faced during their retreat. Temperatures were often below zero, and the army encountered a great deal of snow.

So the time Napoleon's army returned to their start, it was a tiny fraction of its former strength. Without the might of his former army, Napoleon's power in Europe would now come under challenge.

So what makes this visualization so special? Most notably, it is six types of data all being visualized, which we've discussed. We have the latitude and the longitude for the armies. We have the direction the army was traveling, the size of the army, the distance the army traveled, and the temperature that the army faced during their retreat. By encoding all these types of data, one figure is capable of presenting the story of how French losses, French losses during the Russian campaign in 1812.

总结:

Sources

- Jesse Greenspan. Napoleon's Disastrous Invasion of Russia. History.com, June 2012.

http://www.history.com/news/napoleons-disastrous-invasion-of-russia-200-years-ago

- G.D. Sankey or Harness. The Economist. June 2011.

http://www.economist.com/blogs/dailychart/2011/07/data-visualisation

- The Russian Campaign, 1812. PBS.org.

http://www.pbs.org/empires/napoleon/n_war/campaign/page_12.html

To conclude, I want to acknowledge some of the key sources I used when assembling this video. Thank you to Jesse Greenspan, The Economist, and PBS for their excellent articles on the Russian campaign of 1812. I encourage you to explore these articles more if you wish to learn more.

Case Study 3 - Interactive Visualization of World Data

One of the pioneering educators of Data Visualization, Professor Hans Rosling's videos on world development have inspired a whole generation of data scientists. Please feel free to watch this short clip from BBC below:

200 Countries, 200 Years, 4 Minutes - The Joy of Stats

Hans Rosing passed away February 7, 2017. We encourage you to read more about his life and achievements in his obituary.

200 Countries, 200 Years, 4 Minutes - The Joy of Stats:
https://www.youtube.com/watch?v=jbkSRLYSojo

Hans Rosing's obituary
https://www.theguardian.com/global-development/2017/feb/07/hans-rosling-obituary

Week 6: Mini Project (旁听生看不到更具体的内容，跳过)

总结: