Data Visualization and Decision-Making

2024/2025

Anscombe’s Quartet

Anscombe’s Quartet (II)

- Are both data visualizations or only the last one?
- What was the difference between them **as user**?
- Which one is better?
- That one is always better?

Anscombe’s Quartet (III)

Anscombe was not lucky. You can find many datasets like those one:

Source: Autodesk Research: Same Stats, Different Graphs

What do you see here?

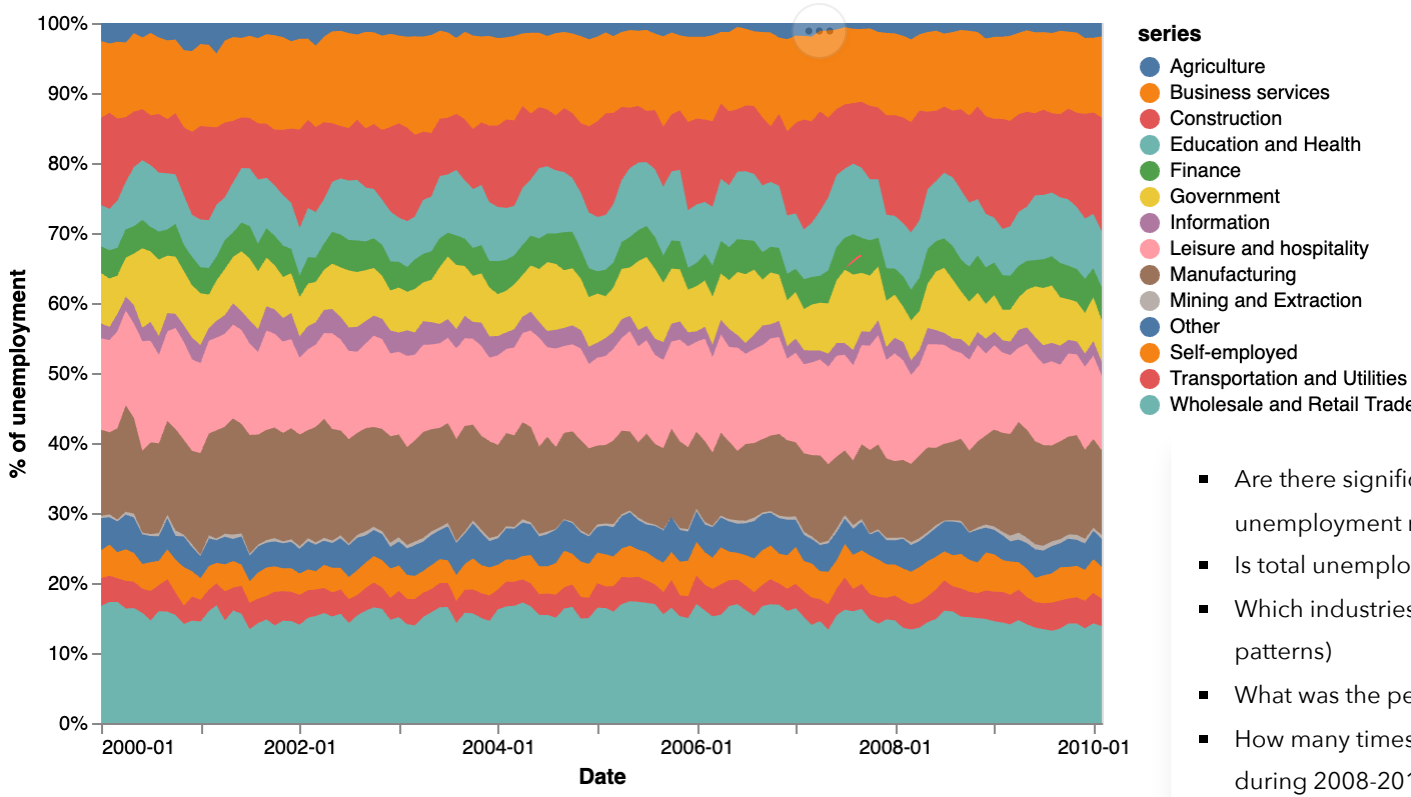
Exercise!

We'll use a dataset about unemployment across industries from 2000 to 2010.

This dataset is represented in different ways in the following slides.

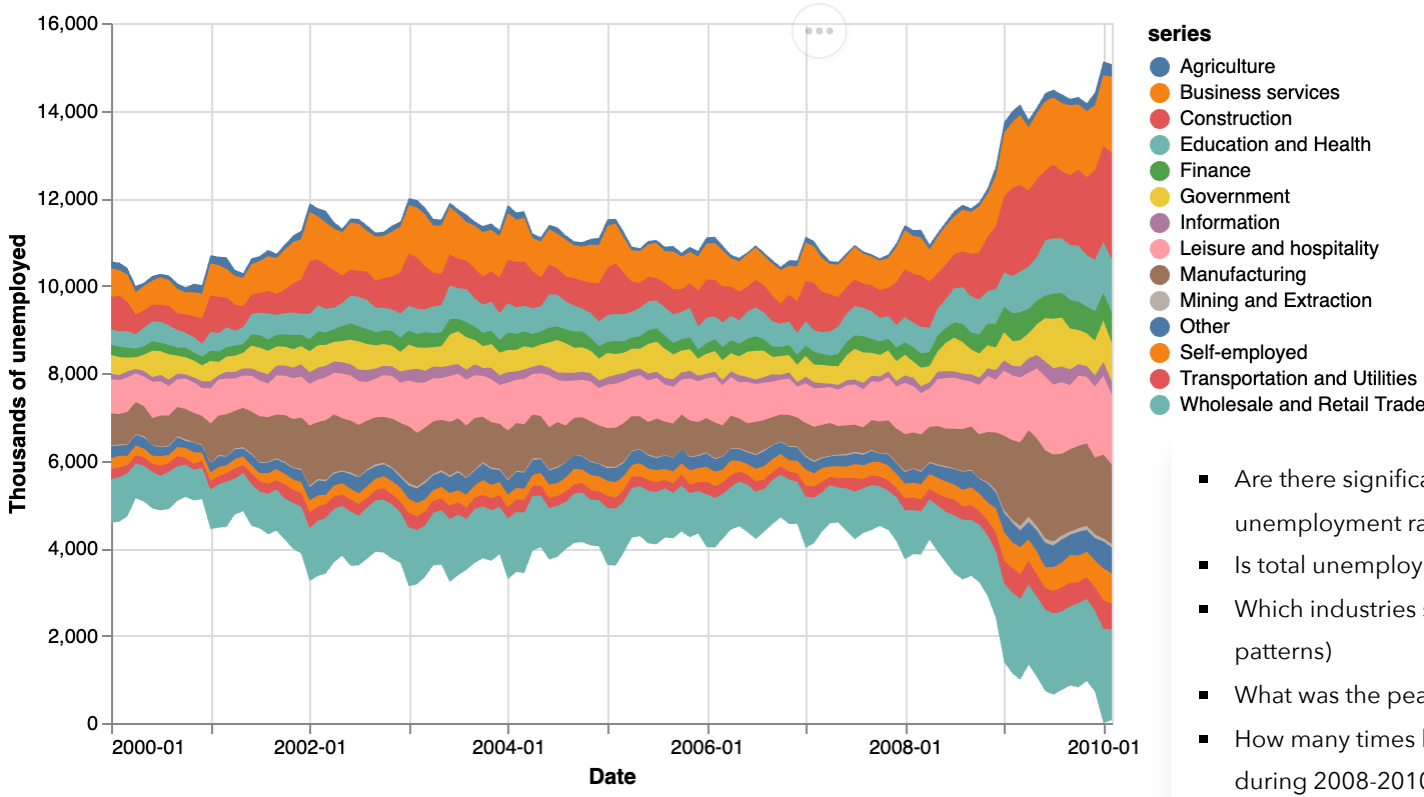
Your task is to answer the following questions using each of the visualizations:

- Are there significant changes in the relative unemployment rates among industries?
- Is total unemployment going up?
- Which industries show yearly patterns? (wave-like patterns)
- What was the peak unemployment before 2008?
- How many times bigger did unemployment get during 2008-2010?
- What is the industry with the highest unemployment rate at 2010?



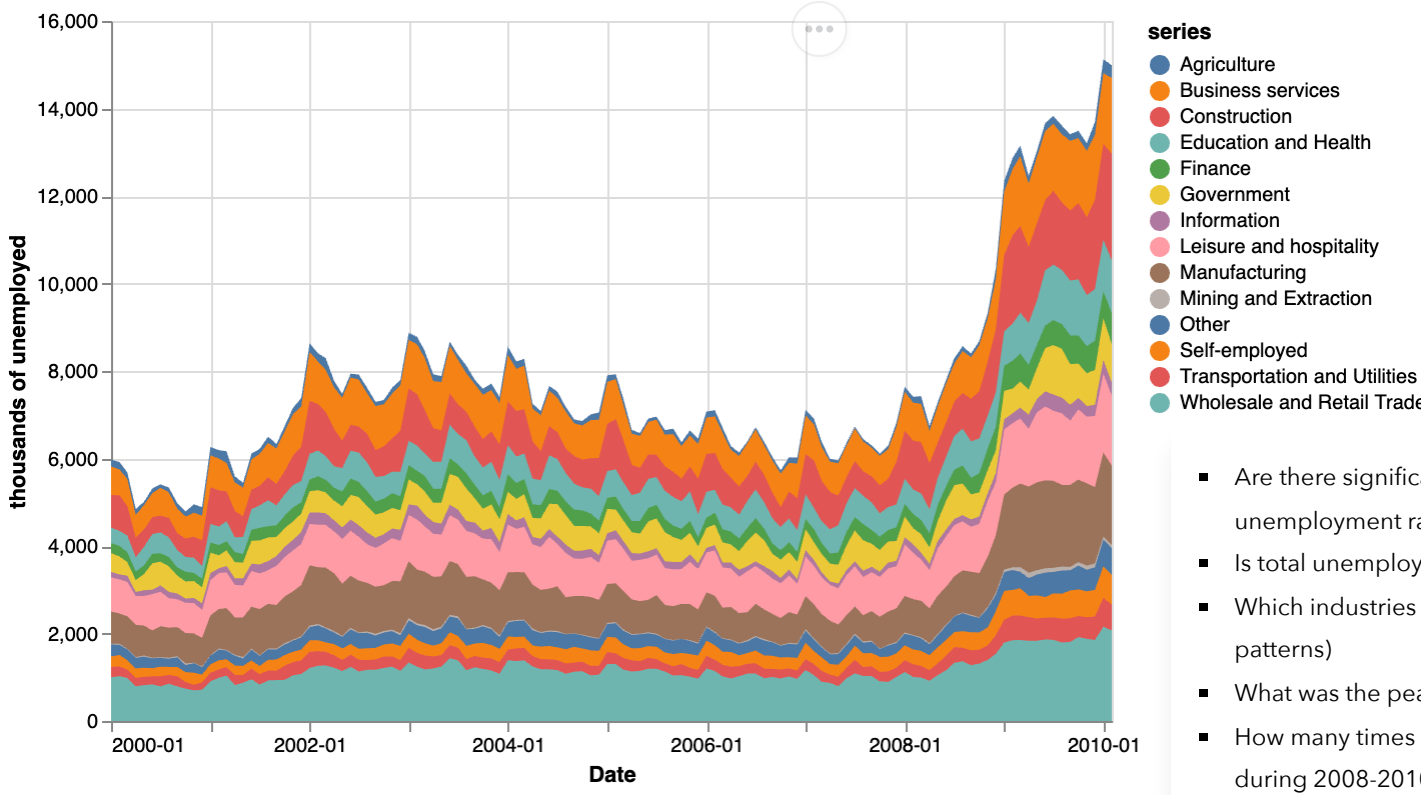
- Are there significant changes in the relative unemployment rates among industries?
- Is total unemployment going up?
- Which industries show yearly patterns? (wave-like patterns)
- What was the peak unemployment before 2008?
- How many times bigger did unemployment get during 2008-2010?
- What is the industry with the highest unemployment rate at 2010?

Violin Plot



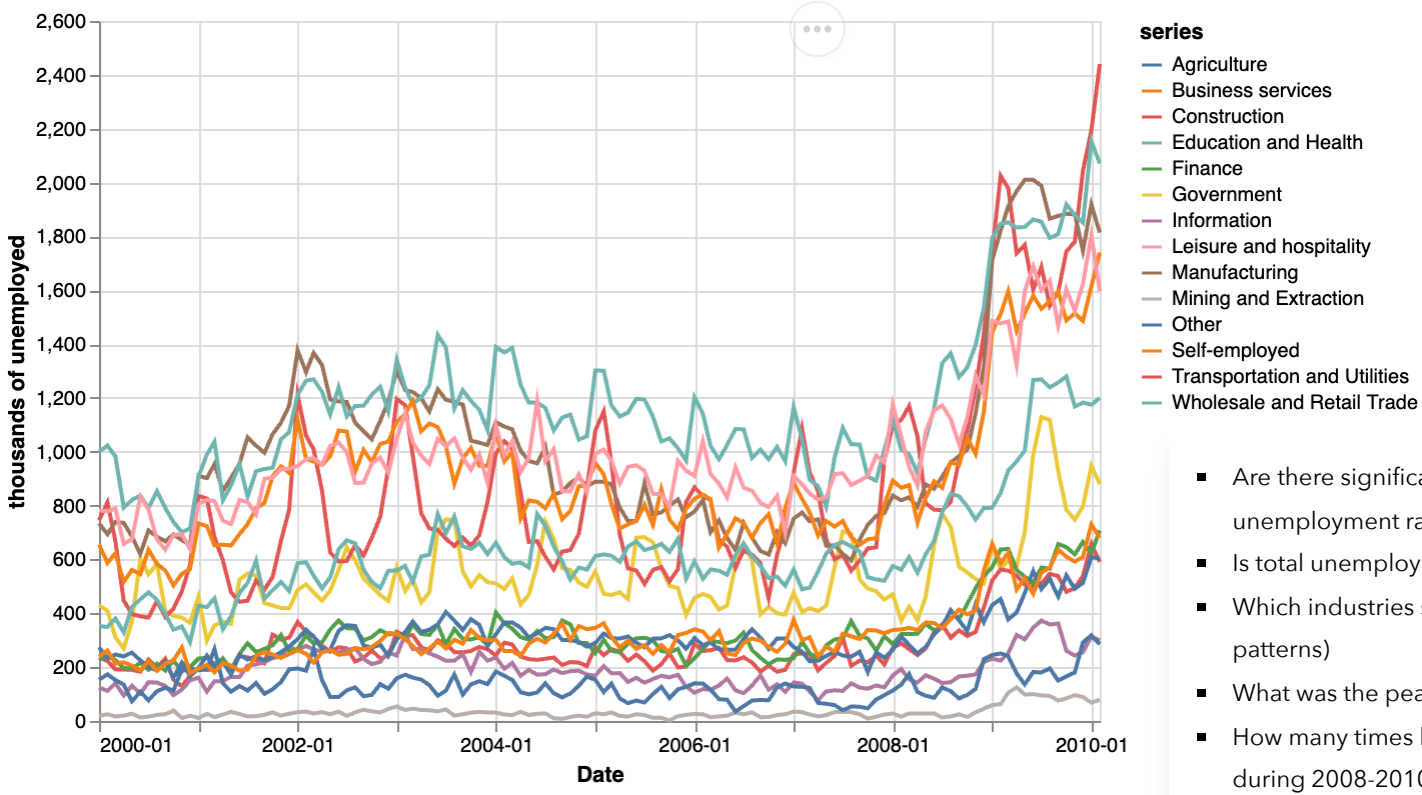
- Are there significant changes in the relative unemployment rates among industries?
- Is total unemployment going up?
- Which industries show yearly patterns? (wave-like patterns)
- What was the peak unemployment before 2008?
- How many times bigger did unemployment get during 2008-2010?
- What is the industry with the highest unemployment rate at 2010?

Absolute numbers



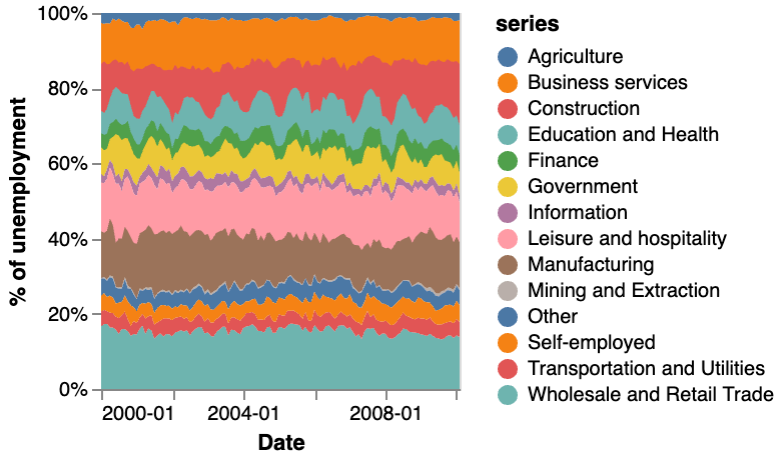
- Are there significant changes in the relative unemployment rates among industries?
- Is total unemployment going up?
- Which industries show yearly patterns? (wave-like patterns)
- What was the peak unemployment before 2008?
- How many times bigger did unemployment get during 2008-2010?
- What is the industry with the highest unemployment rate at 2010?

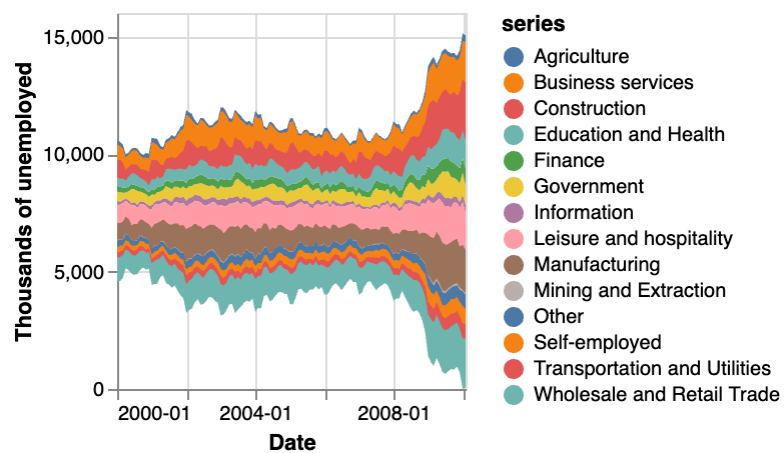
Lines non-stacked



- Are there significant changes in the relative unemployment rates among industries?
- Is total unemployment going up?
- Which industries show yearly patterns? (wave-like patterns)
- What was the peak unemployment before 2008?
- How many times bigger did unemployment get during 2008-2010?
- What is the industry with the highest unemployment rate at 2010?

What is the industry with the highest unemployment rate at 2010?





Unemployment example: conclusions

- To evaluate and desing visualizations, we need to consider the task
- Functionality comes first, aesthetics second
  - We are not looking to "paint pretty pictures" or make "data less boring"
- We are creating a product with a user and a problem to solve

Unemployment example: conclusions (II)

To create good visaulization we have to understand human perception. Since we are using its properties to communicate information we need to squeeze the most out of it and avoid errors in our perception:

- Alignment helps with comparations. Our eyes are evolved to compare objects that are aligned (same ground)
  - Violin is worse than normal area because **we lost the y-axis reference**
- Stack is a visual addition: it helps if we need to sum up everything
  - E.g. unemployment rate is rising? Stacked area is better than independent lines
  - Downsides: it creates patterns depending of the order of the data. Sometimes makes not sense at all (the variable has no order)
- Simple aggregations (like %) could hide important details
  - E.g. unemployment rate is rising? The % graph does not show it
- Eye movement is a anti-sign of good design. Typically if we have to move our eyes to understand the graph we are doing something wrong.
  - In this case the # of colors forces us to move our eyes to the legend to understand the graph

What about the wave-like patterns?

Let’s design a visualization that helps to answer: "Which industries show yearly patterns?"

For that we are going to use Voyager which is a open source, browser-based tool for creating and sharing interactive visualizations.

Add dataset > From URL. Use `https://vega.github.io/vega-lite/data/unemployment-across-industries.json`

Solution #1

Solution #2

Aligment is really important.

If you **align by month** on the x-axis now the **yearly patterns are easier to notice**.

Our eyes catch some patterns. We want to exploit that to create visualizations that triggers "aha moments".

We will call this Level 1 and Level 2.

Level 1 and Level 2

We design for "two users": our eyes and our brain.

Level 1: Visual Level (The Squint Test)

- It’s what you see when you squint your eyes at a visualization
- Automatic and intuitive - like noticing a bright color in a dark room
- Processes visual patterns without reading any labels or context
- Universal: all humans share similar visual processing systems
- Examples: seeing clusters, trends, outliers, or dominant colors
- To evaluate: look at the graph without reading axes or knowing the data context
- Think is modern art and you are lookin at it for the first time

Level 1 and Level 2

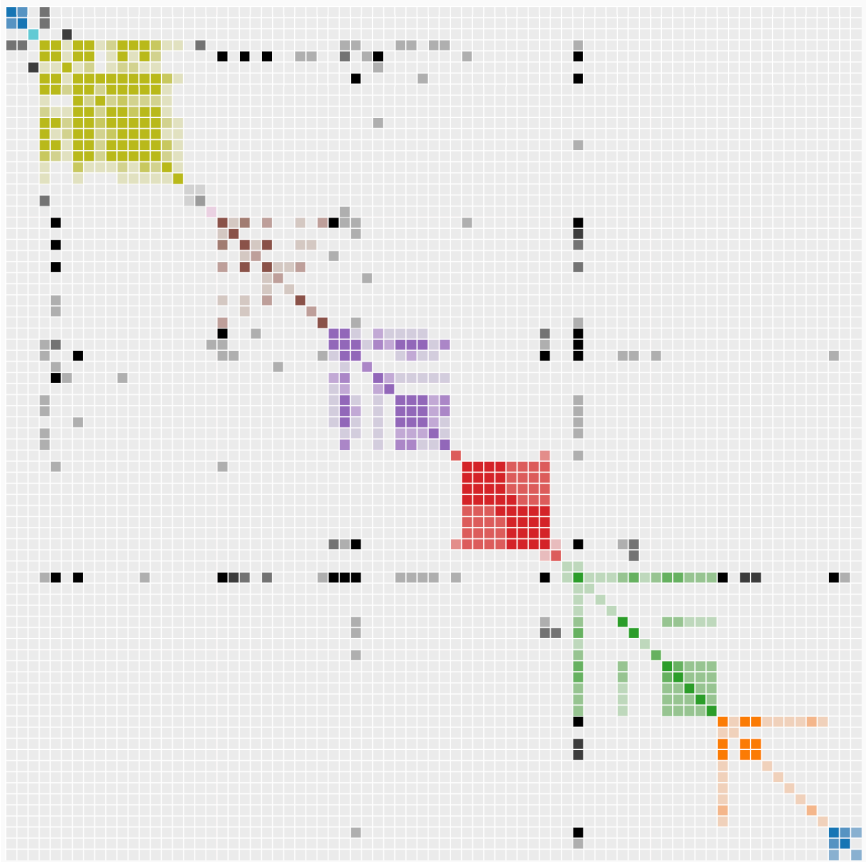
We design for "two users": our eyes and our brain.

Level 2: Cognitive Level (The Meaning Layer)

- Takes Level 1's visual input and adds understanding
- Like translating visual morse code into a meaningful message
- Combines visual patterns with context to generate insights
- Uses cognitive processing to solve problems or answer questions
- Examples: understanding market trends, identifying correlations, comparing data points

When we design visualizations we need to consider both levels:

- Design visualizations that in some situations generate patterns that are meaningful for the visual system (Level 1)
- Ensure these patterns translate to meaningful insights (Level 2)



I removed the axis, labels and titles.

What do you see here?

That's level 1. Without knowing the data, you can still see patterns

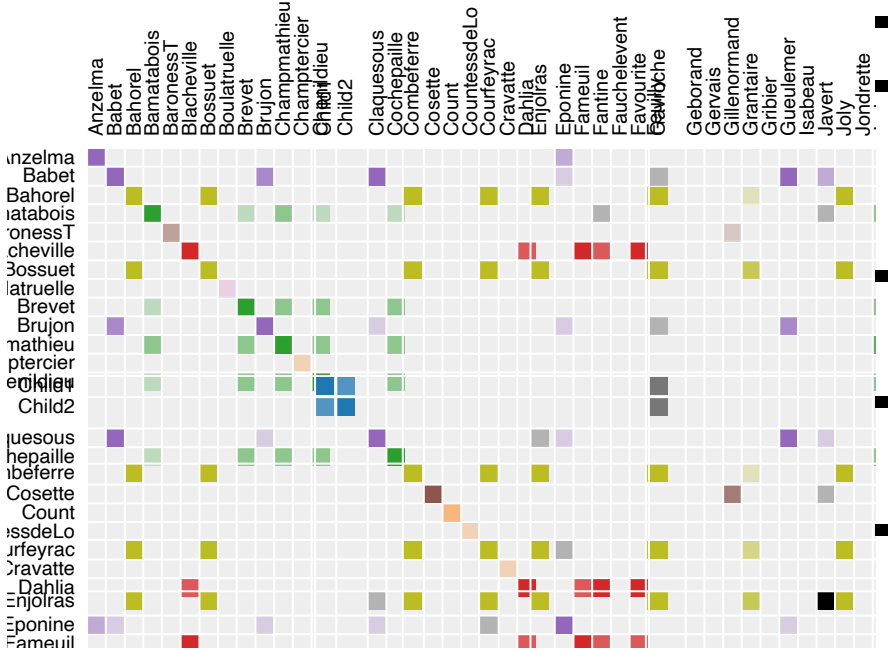
- The main diagonal is darker than the rest of the matrix.
- There are clusters of colors.
- Some of them are more intense than others.
- There is a vertical/horizontal line or a big squared

April 10, 2012 / Mike Bostock

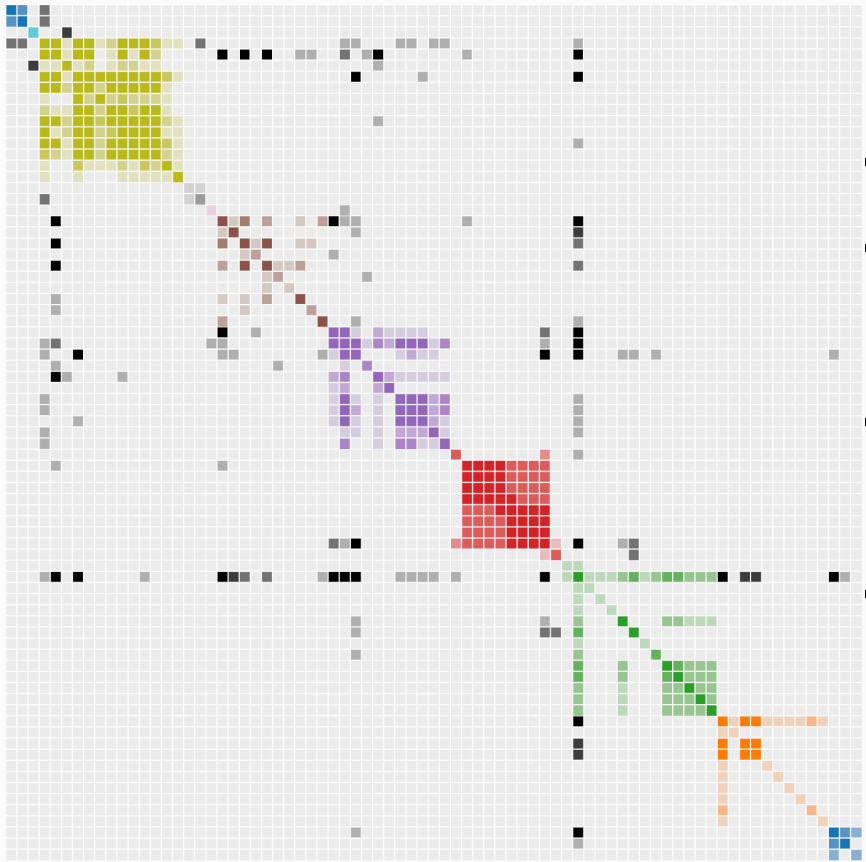
Source: Les miserables Co-ocurrence by Mike Bostock

Les Misérables (

This is a **co-occurrence matrix** showing how characters in Les Misérables appear together in the novel's chapters:



- Each row and column represents a character
- The color indicates which cluster/group the character belongs to (determined by community detection algorithms)
- The intensity/darkness shows how often two characters appear in the same chapters
- The diagonal is dark because characters always "co-occur" with themselves
- Clusters reveal groups of characters that frequently interact together in the story



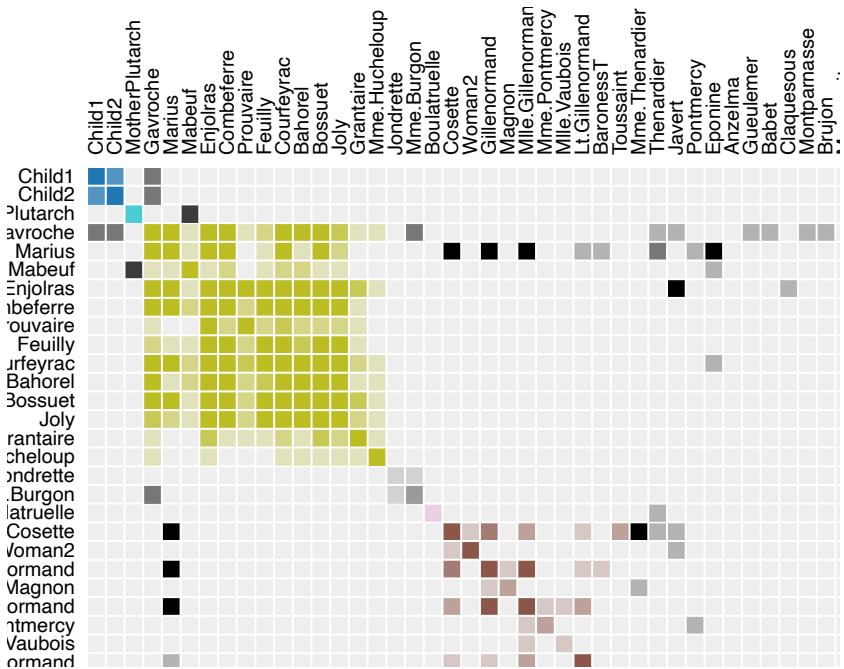
Level 2 analysis

Explain the visual patterns in the context of the data.

- The main diagonal is darker than the rest of the matrix.
  - This is because characters always "co-occur" with themselves.
- There are clusters of colors.
  - These clusters represent groups of characters that frequently appear together in the story: Valjean's associates, the Thénardiers, and the students.
- Some of them are more intense than others.
  - Some groups have stronger internal connections.
  - For example, the Thénardiers family members appear together in many scenes
- There is a vertical/horizontal line or a big squared
  - Valjean appears as a prominent line because he is the main character who interacts with many others throughout the story

April 10, 2012 / Mike Bostock

Les Misérables



As designers we need to understand **both levels of analysis**: human visual system (level 1) that pays attention to specific patterns and level 2 that interprets those patterns to get insights or actionables.

If we reorder the rows by other property (e.g. alphabetically) all those patterns disappear. That's the magic of alignment and our visual system.

Level 1, although automatic and implicit, is **needed to extract insights and business value from our analytics**.

Good data visualizers and product designers exploit the visual system to maximize the amount of information we can process.

If we reorder the characters by name using the same data, **we get no insights because Level 1 is not working** ("I cannot see anything").

Data visualization as a **product discipline**

As we've seen, good data visualization needs a user's perspective.

We cannot design a good visualization without its use case. Or in other words:

Data visualization is a product discipline.

We have **users**, with **pain points**. We build **products** (data visualizations/dashboards) to **solve their problems**.

But, historically, data visualization has been treated either as an artistic discipline or a secondary goal of data science.

Data visualization is the **last mile** of the data science pipeline.

It delivers value to the user. So if correctly executed multiplies the impact of the work of data scientists and analysts.

(IMO) You would get more impact from being a better designer than a better data scientist.

The best analysis is worthless if it can't be understood and acted upon.

People pay (more) to extract value from data not from the data itself.

Practical data visualization

How can I design the right visualization?

Key Principles for Effective Visualization

The very basics of applied human perception

- Level 1: Visual patterns (what you see when squinting)
- Level 2: Cognitive interpretation (making meaning from patterns)

Optimize for Human Perception

- Minimize eye movement
- Use alignment and proximity
- Stack when totals matter
- Generate visual patterns with data: clusters, outliers, shapes, …

Focus on User Needs

- Start with the user's questions/tasks. Consider it a product solving a problem
- Choose visualizations that make answers obvious
  - How we know is it obvious?

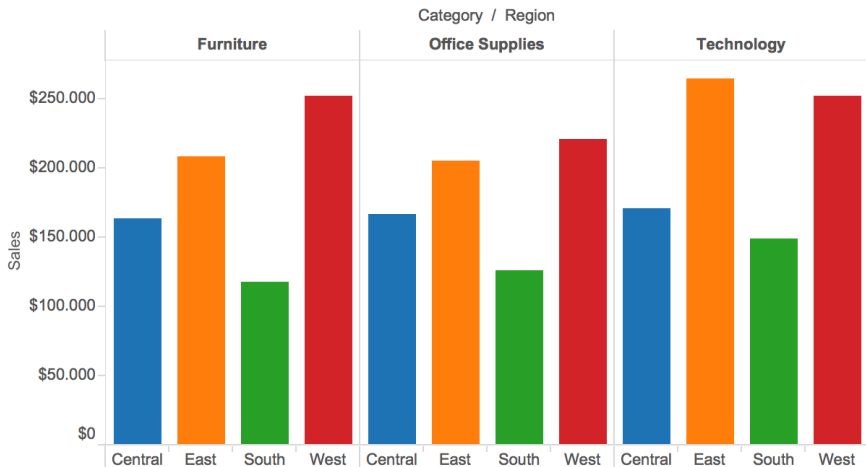
Self-observation exercise

I will show you different graphs of the same data.

Time how long it takes you to answer a question.

Watch what you do to find the answer.

Online store dataset with 4 regions and 3 product categories.



Time yourself and observe what is happening

Ready…

Which region has the highest sales?

What happened?

Ready…

Which region has the lowest sales?

Ready…

Which region has the highest furniture sales?

Ready…

Which category has the highest sales in Central (blue)?

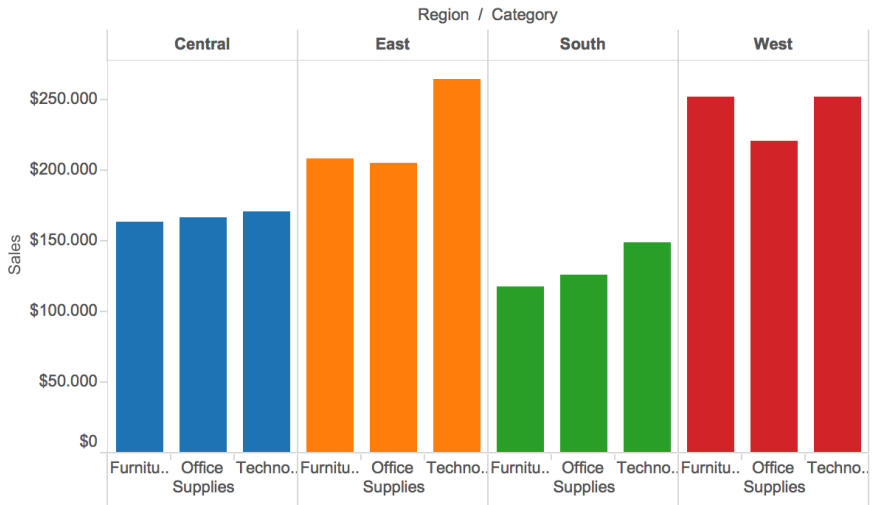
Online store dataset with 4 regions and 3 product categories.

Ready…

Which category has the highest sales in Central (blue)?

Ready…

Which region has the highest technology sales?



Ready…

Which region has the highest sales?

Ready…

Which category has the highest sales in Office Supplies (Orange)?

In general, response time is a good quality metric

Is a good proxy of how obvious is to answer the question with that visualization.

Ideally you test this with real users.

In practice, you can test it with yourself or other colleagues.

Key Principles for Effective Visualization (II)

- Minimize eye movement
- Use alignment and proximity
- Stack when totals matter
- Generate visual patterns with data: clusters, outliers, shapes, …
- Measure response time as a proxy for data visualization quality
- Be consistent with colors, shapes, axis, …
  - In our examples, orange has two meanings: region and category. Avoid that.

Data visualization for decision making

Organizations use data visualization to help make better decisions.

Example: SaaS company CEO

Let’s look at a real example: A CEO of a software subscription company wants to "make sure everything is going well."

- Understanding what the CEO needs:
  - The request is not specific - this is normal, users often can’t explain exactly what they need
  - Let’s focus on subscription revenue, which is likely the most important metric
- **Exercise: What questions should we answer to help this CEO?**

Data viz for decision making

Most examples we've seen so far use fixed data that doesn't change.

- In the real world, data changes constantly
- We need to make decisions based on these changes
- **Planning for different scenarios:**
  - Think about the different situations that might happen
  - Design visualizations that work well in many scenarios, not just one

Data viz for decision making

If we have n scenarios: $S_1, S_2, \ldots, S_n, \ldots$

Examples:

- S_1: the company is losing customers abruptly
- S_2: customers churn after 5 months
- S_3 the outflow and inflow of users is symetrical and the user base is stable
- …

We have two **very important goals**:

- Saliency: If S_n happens, we want to make sure the visualization helps us notice it
- Discrimination: If S_n happens, we want to make sure the visualization is **different** from the other scenarios

Retention by cohort

This table shows how long new members who sign up remain a member. The percentage shows how many of them who signed up are still a member in the corresponding month.

Signup	Start	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
Feb '18	100%	92%	88%	86%	86%	84%	84%	84%	84%	84%	84%	84%	
Mar '18	100%	94%	89%	87%	87%	87%	87%	87%	87%	87%	87%		
Apr '18	100%	69%	58%	57%	53%	52%	52%	49%	45%	45%			
May '18	100%	79%	67%	63%	59%	59%	57%	55%	54%				
Jun '18	100%	74%	65%	60%	60%	57%	53%	52%					
Jul '18	100%	72%	60%	58%	54%	54%	54%						
Aug '18	100%	79%	64%	52%	51%	44%							
Sep '18	100%	61%	54%	51%	51%								
Oct '18	100%	81%	75%	75%									
Nov '18	100%	92%	92%										
Dec '18	100%	88%											
Jan '19	100%												

Retention by cohort

This table shows how long new members who sign up remain a member. The percentage shows how many of them who signed up are still a member in the corresponding month.

Signup	Start	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
Feb '18	100%	92%	88%	86%	86%	84%	84%	84%	84%	84%	84%	84%	
Mar '18	100%	94%	89%	87%	87%	87%	87%	87%	87%	87%	87%		
Apr '18	100%	69%	58%	57%	53%	52%	52%	49%	45%	45%			
May '18	100%	79%	67%	63%	59%	59%	57%	55%	54%				
Jun '18	100%	74%	65%	60%	60%	57%	53%	52%					
Jul '18	100%	72%	60%	58%	54%	54%	54%						
Aug '18	100%	79%	64%	52%	51%	44%							
Sep '18	100%	61%	54%	51%	51%								
Oct '18	100%	81%	75%	75%									
Nov '18	100%	92%	92%										
Dec '18	100%	88%											
Jan '19	100%												

- Each row is a cohort of users:
  - A cohort is a group of users who started using the product in the same month
  - Each column shows how many users are still active after X months
  - Colors indicate retention rate (darker = better retention)
- This visualization helps identify patterns:
  - Vertical patterns: How retention changes for specific months of usage
  - Horizontal patterns: How different cohorts perform over time
  - Diagonal patterns: Calendar-specific effects
- Example reading:
- February 2018 cohort: 100% start active (month 0), by month 3, only 86% are still active
- We can compare this to newer/older cohorts

Retention by cohort

This table shows how long new members who sign up remain a member. The percentage shows how many of them who signed up are still a member in the corresponding month.

Signup	Start	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
Feb '18	100%	92%	88%	86%	86%	84%	84%	84%	84%	84%	84%	84%	
Mar '18	100%	94%	89%	87%	87%	87%	87%	87%	87%	87%	87%		
Apr '18	100%	69%	58%	57%	53%	52%	52%	49%	45%	45%			
May '18	100%	79%	67%	63%	59%	59%	57%	55%	54%				
Jun '18	100%	74%	65%	60%	60%	57%	53%	52%					
Jul '18	100%	72%	60%	58%	54%	54%	54%						
Aug '18	100%	79%	64%	52%	51%	44%							
Sep '18	100%	61%	54%	51%	51%								
Oct '18	100%	81%	75%	75%									
Nov '18	100%	92%	92%										
Dec '18	100%	88%											
Jan '19	100%												