

# Data Visualization

Spring Semester 20/21  
Story telling with Data

Iolanda Velho, Ph.D.  
Nova School of Business and Economics

April 13<sup>th</sup>, 2021

- Why story telling
- How to construct a story
- How to properly use data in your story

- Why story telling
- How to construct a story
- How to properly use data in your story

# Why is story telling relevant?

**There are a variety of reasons to tell stories :**

- Stories solidify abstract concepts and simplify complex messages.
- Stories are a universal language of sorts
- Stories inspire and motivate.

**They help to focus the attention of your audience – and  
attention is a scarce resource!**

# Stories transport us

When we consume uninteresting information, a certain part of our brain called the Wernicke's area is activated to translate the words into meaning.

But when we hear a story, our brains change dramatically:

- Not only the language processing parts activate,
- But so are whatever areas that would be used if you were actually in the story yourself.
  - Stories are associated with focus, motivation, attention (dopamine production), and empathy (oxytocin production)
- The brains of the storyteller and the story listener can actually synchronize, says Princeton's Uri Hasson:

**“By simply telling a story, [a person] could plant ideas, thoughts and emotions into the listeners’ brains.”**

# We are programmed to recognize patterns of information

Stories are recognizable patterns, and we use them to find meaning in the world around us

Stories are so near and dear to us that we even invent them when they're not actually there.

- In 1944, 34 Massachusetts college students were shown a short film with two triangles and a circle moving across the screen. They were then asked to describe the scene. All but one described the movements with elaborate, human narratives, including:
  - The two triangles were men fighting as a woman (the circle) tried to escape.
  - The circle was “worried.”

# Our story must follow a clear pattern so that the audience recognizes it as a story

The best stories are snapshots of a world improved by an action we want to promote. You need to set up your story to show:

- 1 - A problem to be solved, and a character that is affected by the problem:
- 2 - A way to solve the problem:
- 3 – The impact of solving the problem

# Our story must follow a clear pattern so that the audience recognizes it as a story

The best stories are snapshots of a world improved by an action we want to promote. You need to set up your story to show:

- 1) Your client wants to expand his gourmet street food truck chain to a new location in Europe.
- 2) By focusing in markets with fewer medium high /high end restaurants, he could tap into that market at a lower cost
- 3) He would make gazillions of Euros

# Our story must follow a clear pattern so that the audience recognizes it as a story

The best stories are snapshots of a world improved by an action we want to promote. You need to set up your story to show:

- 1) Your client wants to expand his gourmet street food truck chain to a new location in Europe.
- 2) By focusing in markets with fewer medium high /high end restaurants, he could tap into that market at a lower cost
- 3) He would make gazillions of Euros

Difficult

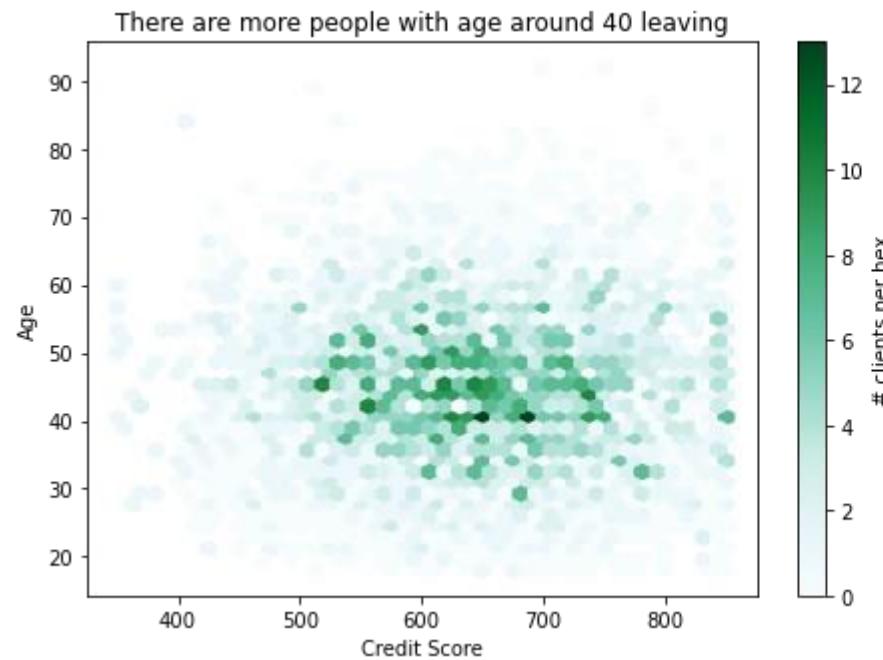
- Why story telling
- **How to construct a story**
- How to properly use data in your story

# To help introduce new concepts, create contrast between the expected narrative vs our great insight

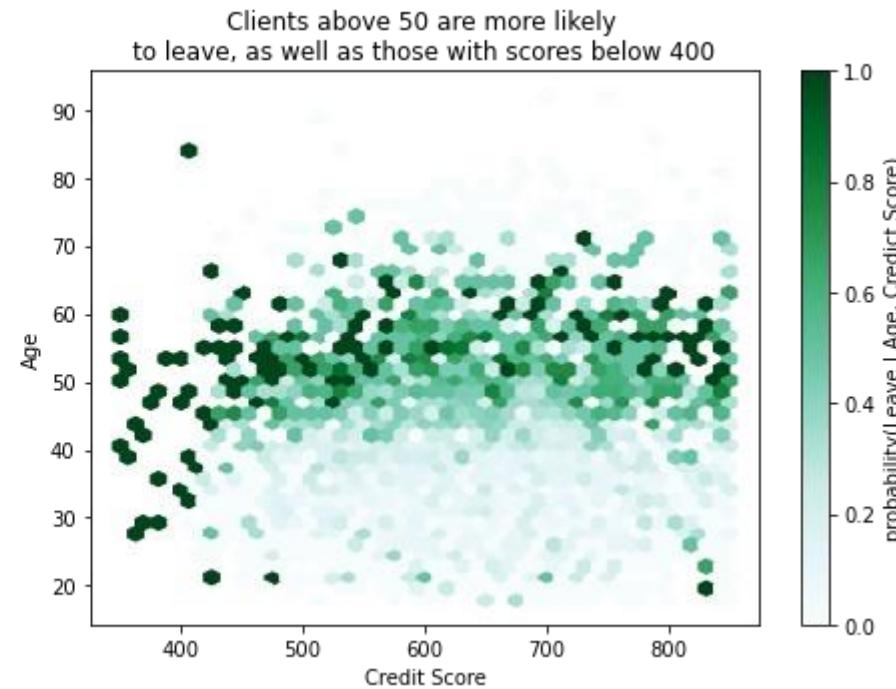
- A) You start by showing what people expect to find,
  - B) show the problem in the result,
  - C) introduce your solution,
  - D) show the results with your solution
- 
- This is the basic structure of most papers.
  - It helps explaining difficult approaches: first you show what people already know – and then add your novelty

# To help introduce new concepts, create contrast between the expected narrative vs our great insight

A well known analysis and expected result



A twist that is easier to explain after introducing basic concepts



# Use your stories to guide the project development

**1- Formulate a plausible hypothesis** – markets with a smaller percentage of expensive restaurants have better opportunities for my gourmet food truck

**2 – Define the key analysis that support/disprove this hypotheses**

- Identify examples of these markets
- Identify the rating of smaller business with food types similar to yours
- If your hypothesis is correct – you should have better rating in these markets

**3 - Do those analysis and only those**

- double check your reasoning
- get one or two insights that can lead to one or two new analysis or hypothesis.  
Eg. the fact that ratings are biased for those with fewer reviews!

**Stop! Finish writing your story/prepare your document.**

- Even if you disproved your hypothesis, that is a good insight and if the hypothesis is well formulated, you can still find a solution to the problem

# When presenting your work, tell your story in big letters

How you tell your story depends on the medium. But most often you can use presentations or similar: **You have a big title, and some white space**

## **Use the title to tell your story**

- If you gave the presentation to someone, just by reading all titles they should understand your message
- Each title should be informative and affirmative: e.g., avoid passives

## **Use the white space to support the claim in your title**

- With graphics, diagrams ... even text ... each slide should have only the content that supports the title

# When presenting your work, tell your story in big letters

Quantitative information is precise, but makes it difficult to find patterns or trends

| Region        | Jan   | Feb   | Mar   | Apr   | May  |
|---------------|-------|-------|-------|-------|------|
| Domestic      | 1,983 | 2,343 | 2,593 | 2,283 | 2,57 |
| International | 574   | 636   | 673   | 593   | 64   |
| Total         | 2,557 | 2,979 | 3,266 | 2,876 | 3,21 |

- This table does two things extremely well: efficient means to look up values for a part
- But if we're looking for patterns, trends, or of the story contained in these numbers, consider just two at a time, this table fails.



2009 Sales (thousands of U.S. \$)

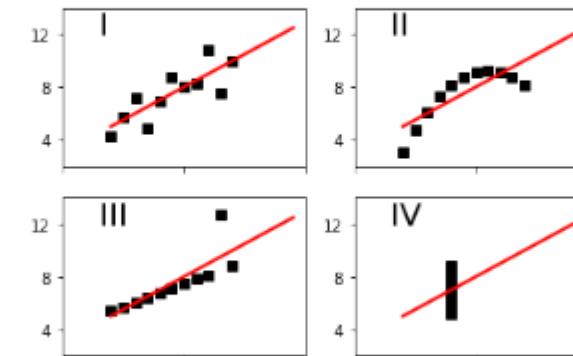
Qualitative representations allow for a more global perspective, but make low level comparisons difficult



- Each character in the Game of Thrones can be represented as a vertex
- A link between characters means that they interacted, spoke of one another, or that another character spoke of them together (a link between two characters whenever their names appeared within 15 words of one another)
- Qualitative information can also be transformed to represent the complex semantics of the data

Game of Thrones Characters Network

Furthermore, qualitative and global information are fundamental in the selection the relevant quantitative metrics



Anscombe's quartet: same descriptive statistic, very different distributions

E.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (February 1973), 17-21.

- Why story telling
- How to construct a story
- How to properly use data in your story

# Follow the principles of visualization design (from 1<sup>st</sup> class) when representing your data

- Picture Superiority Effect: human is able to retain more information seen through pictures than through words
- Choosing effective visual encodings requires knowledge of visual perception, because an effective visual should tap into the brain's "pre-attentive visual processing" (e.g. color, size, orientation, etc.)
- Tell the truth and nothing but the truth (don't lie, and don't lie by omission)
- Use encodings that people decode better (where better = faster and/or more accurate)

# Follow the principles of visualization design (from 1<sup>st</sup> class) when representing your data

1. Show the data
  - Data are the most important part of the visualization and should be presented in the clearest way possible
  - But that does not mean that all of the data must be shown
2. Reduce the clutter (the use of unnecessary or distracting visual elements)
  - unnecessary tick marks, labels, or text
  - unnecessary icons or pictures
  - ornamental shading and gradients
3. Integrate the text and graph
  - visualizations should be constructed to complement the text and at the same time to contain enough information to stand alone

# Each type of chart has its own peculiarities

Bar charts

Pie charts

Line charts

# Each type of chart has its own peculiarities

Bar charts

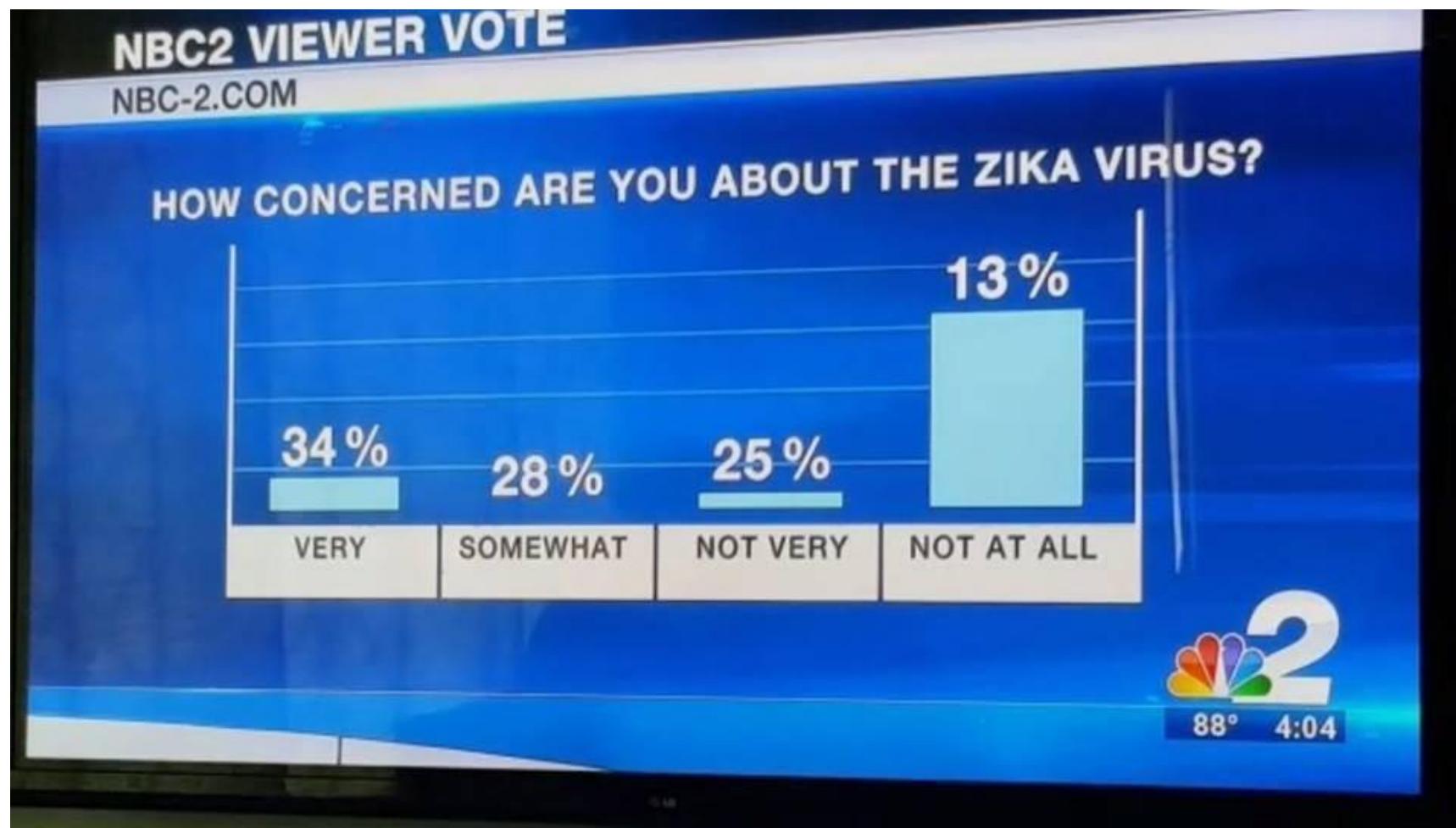
Pie charts

Line charts

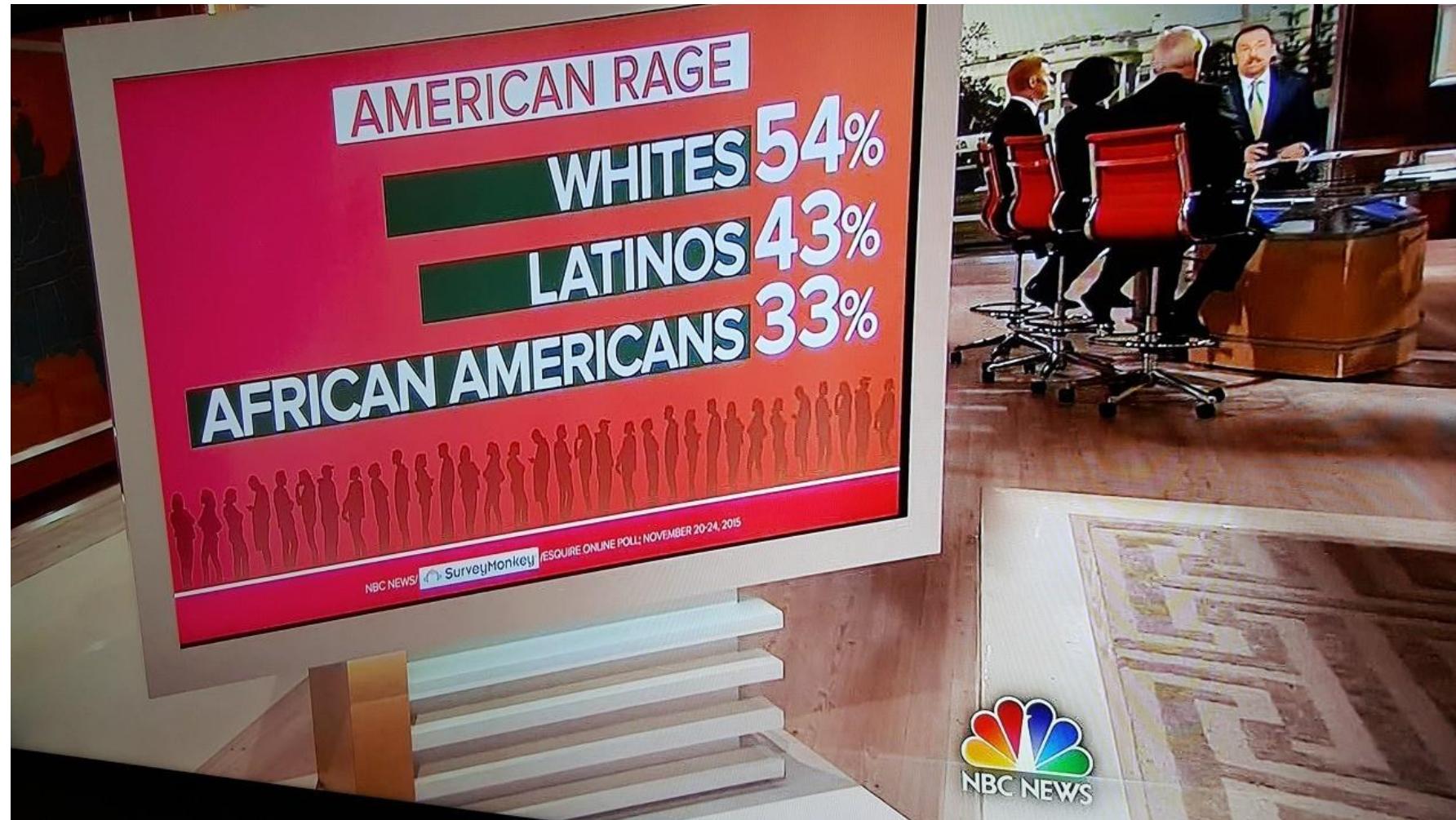
## Key takeaways for bar plots

- By its design, a bar plot emphasizes the **absolute magnitude of values** associated with each category
- That means, the length of the bar should proportionally represent the magnitude of values
- Bar chart axes should include zero
- Do NOT use stacked bar when it should be grouped!
- Do NOT compare the bars between different data
- Grouped bar chart when you want to compare the values in sub-categories

What is wrong in this plot?



What is wrong in this plot?

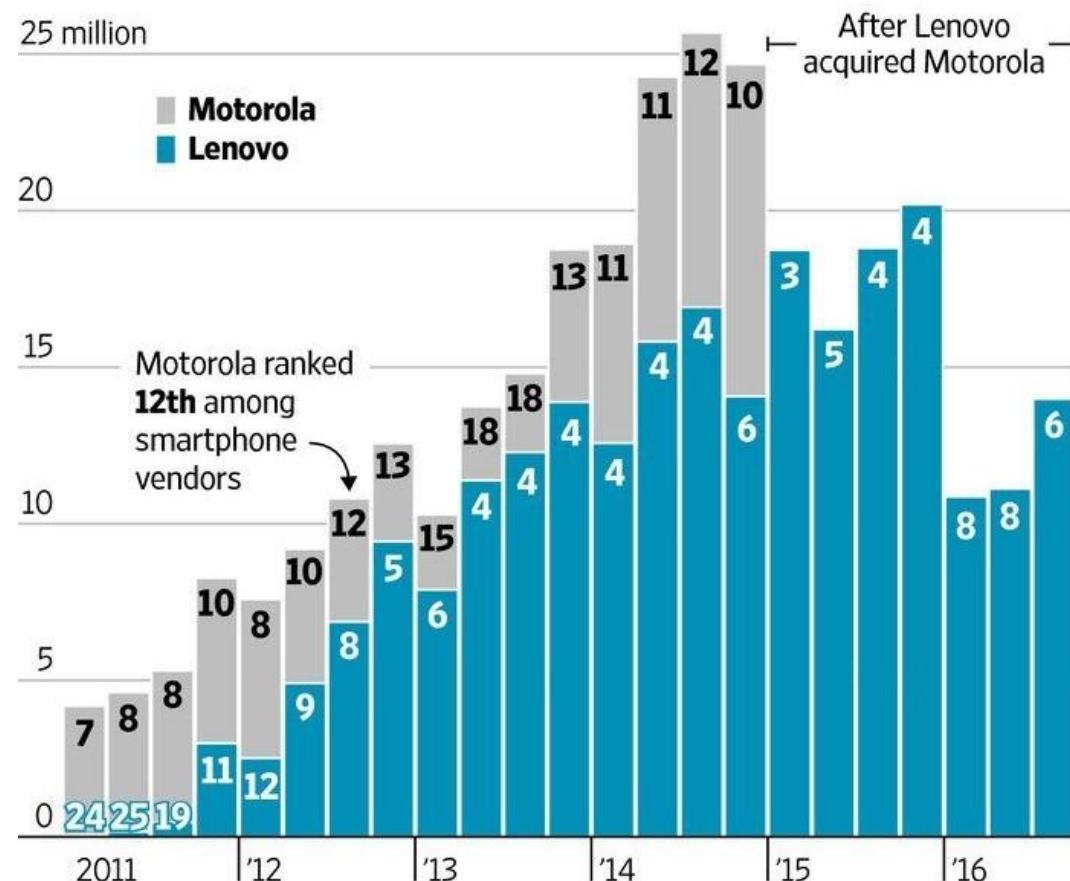


# What is wrong in this plot?

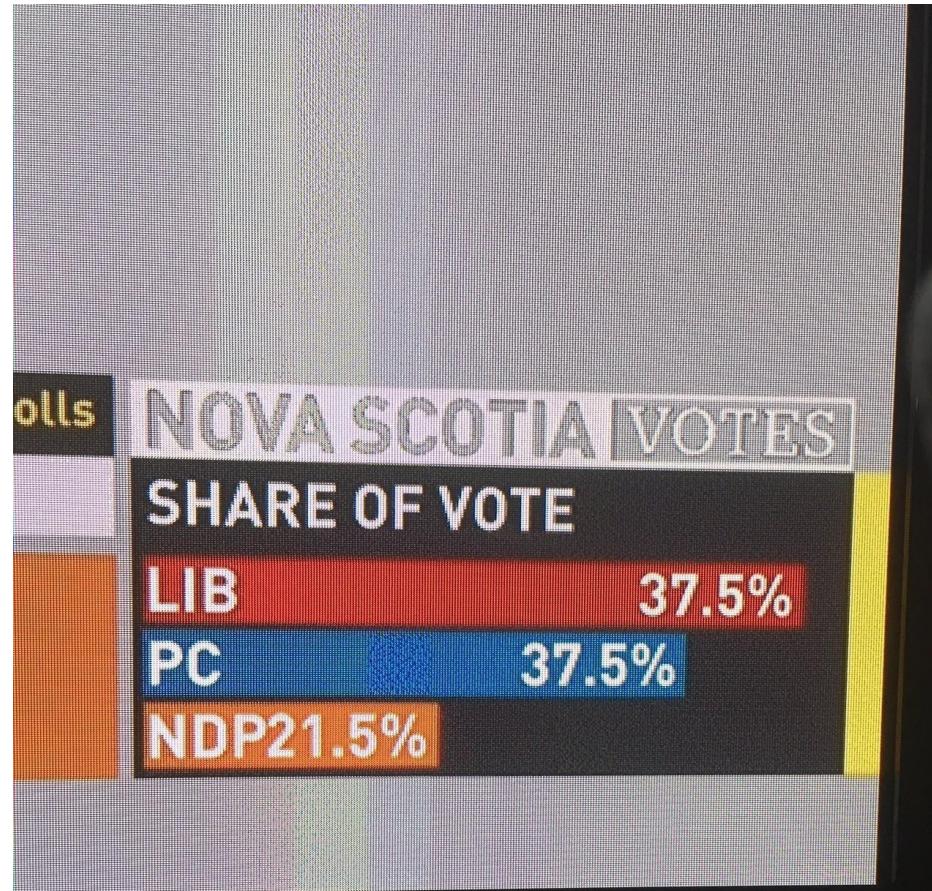
## Smartphone Hang-ups

China's Lenovo Group has struggled to integrate the smartphone business of Motorola, which it acquired in October 2014, part of a surge in Chinese acquisitions of foreign companies.

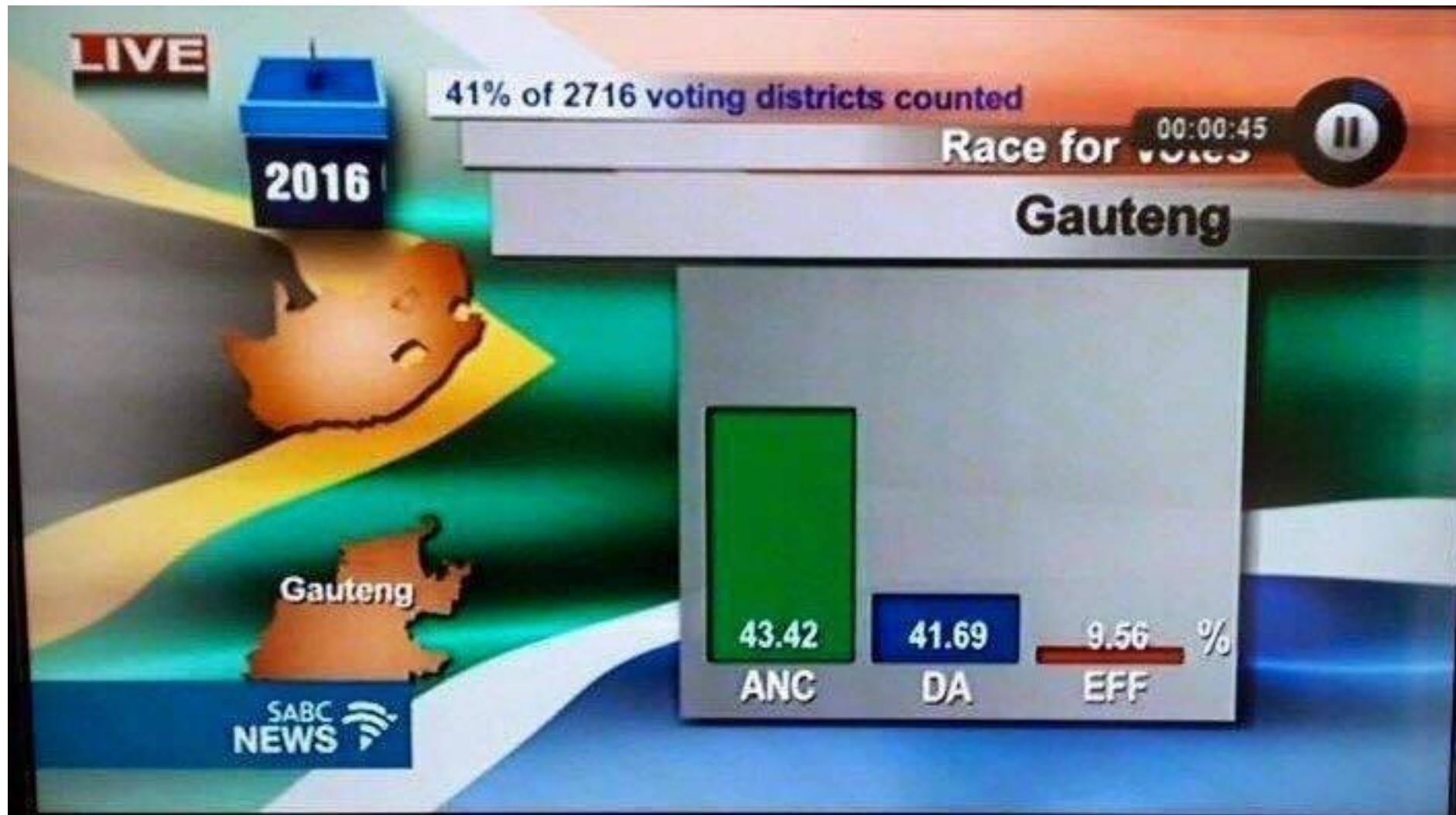
### World-wide smartphone shipments and vendor ranking



What is wrong in this plot?



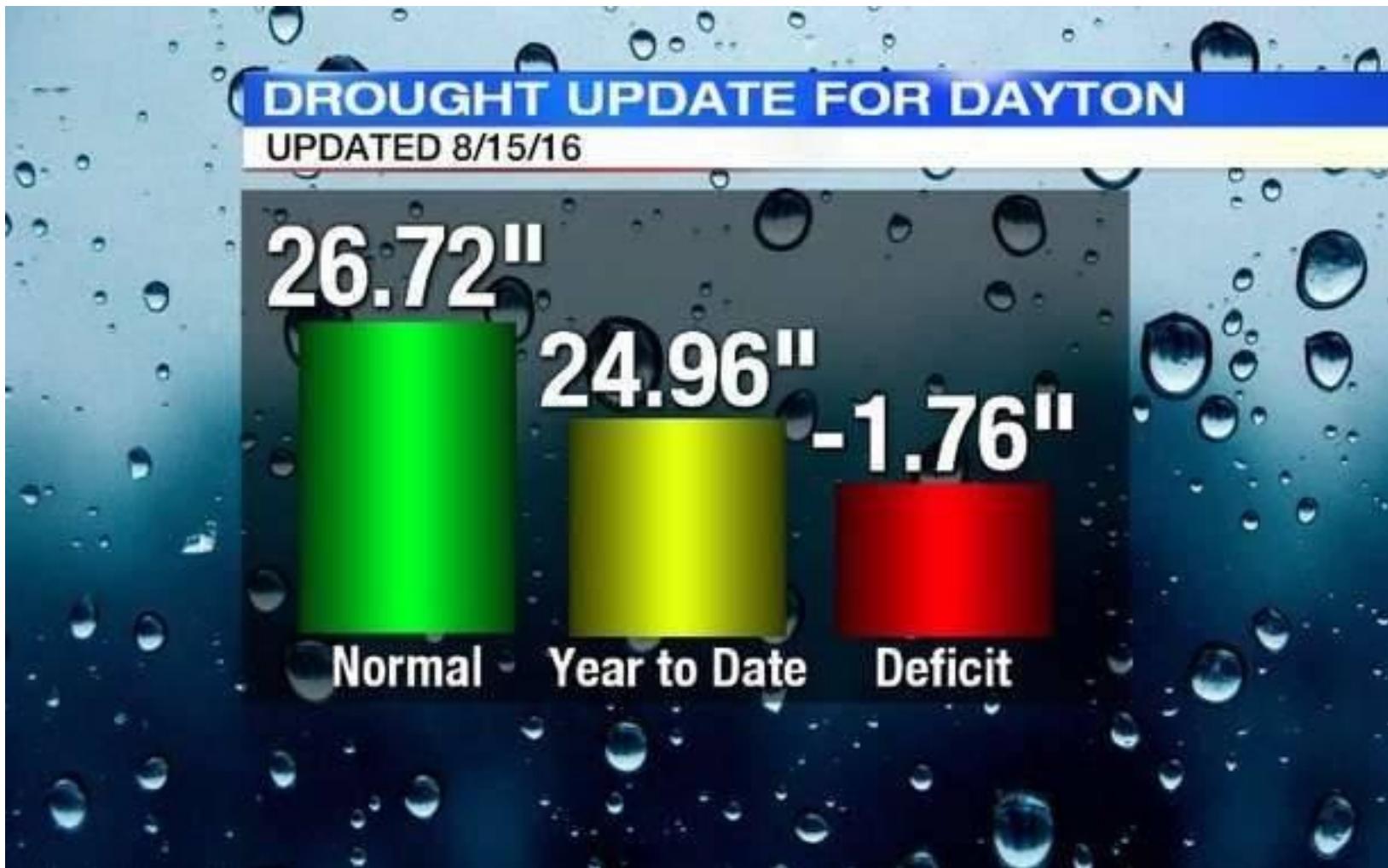
What is wrong in this plot?



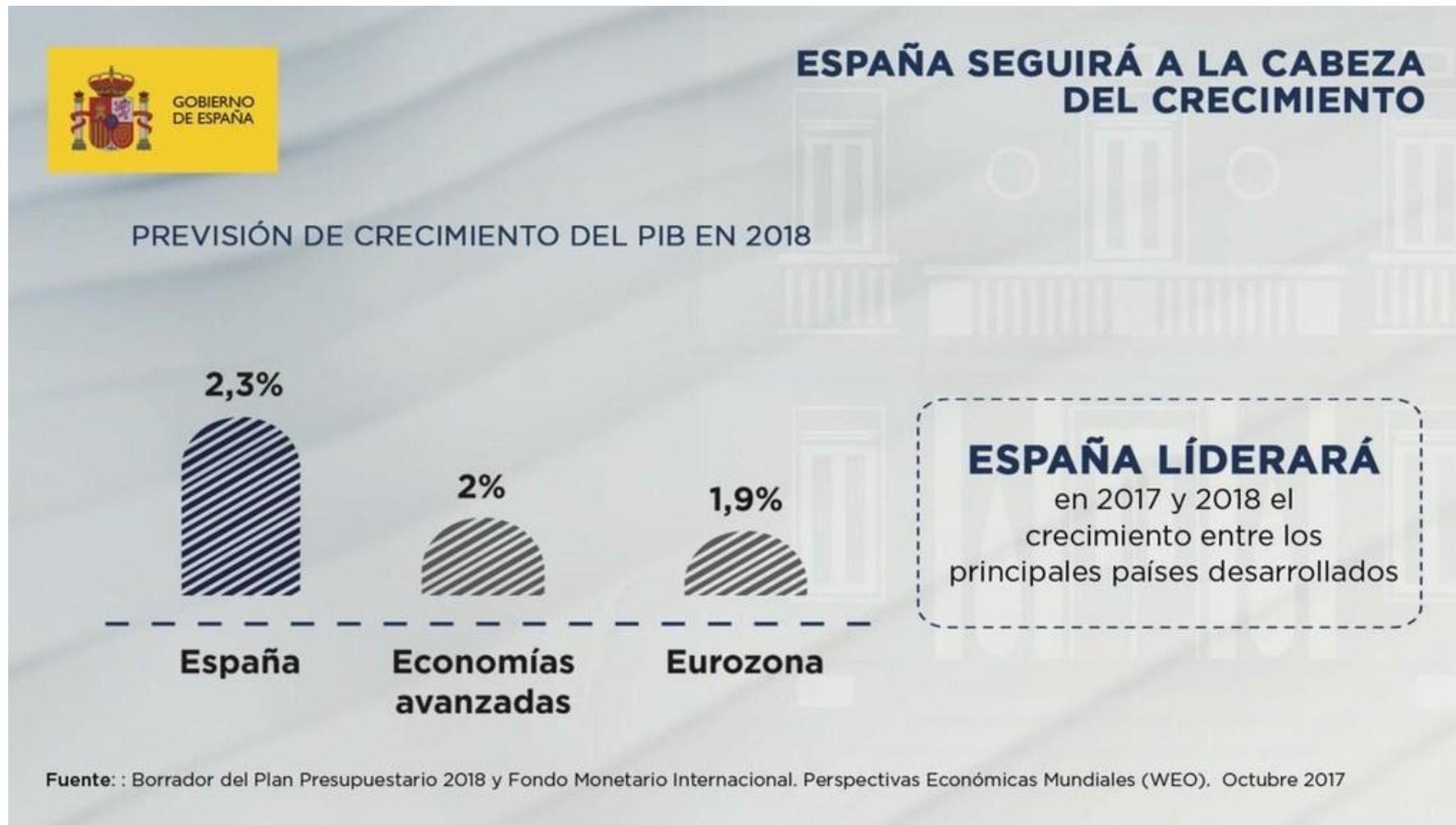
What is wrong in this plot?



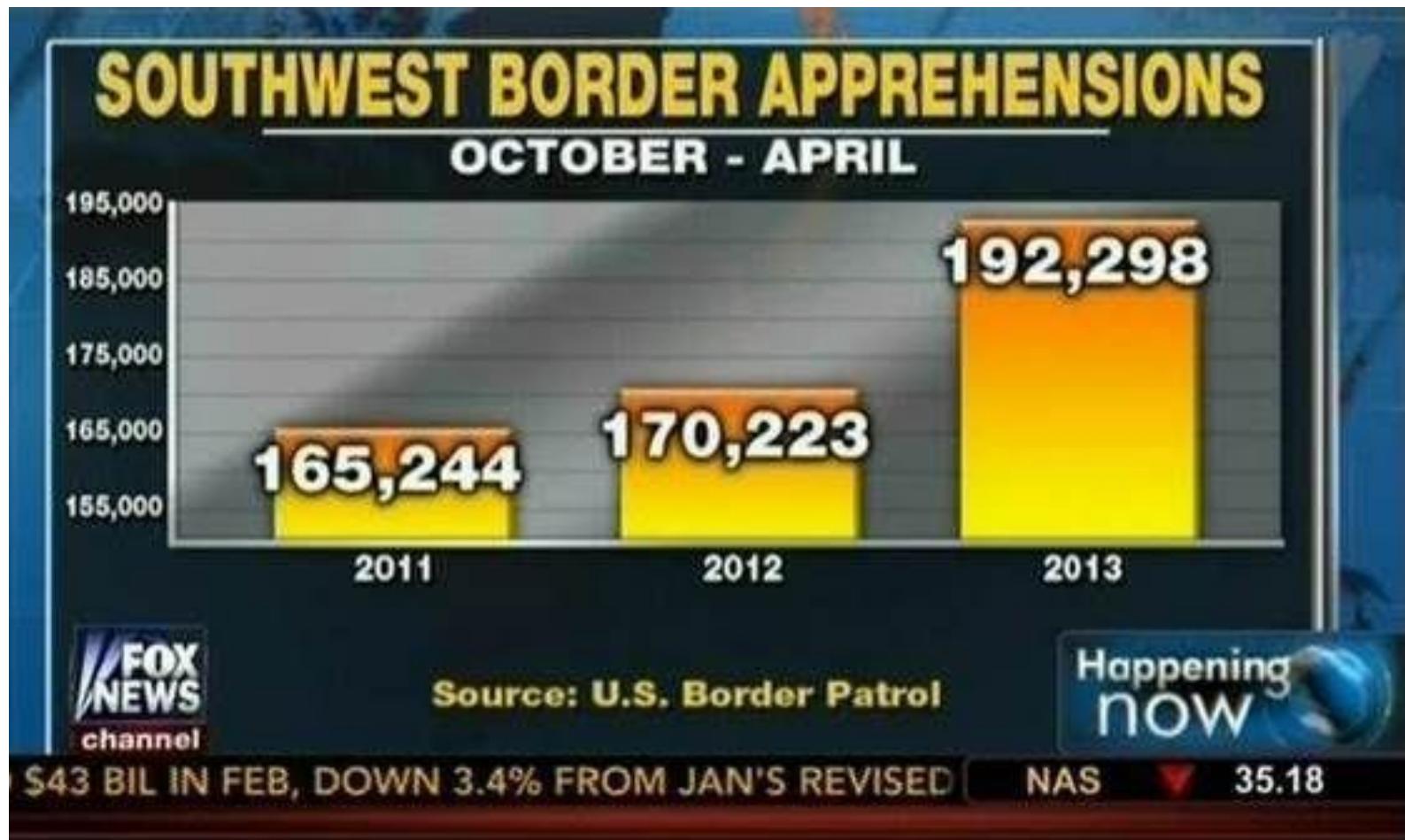
What is wrong in this plot?



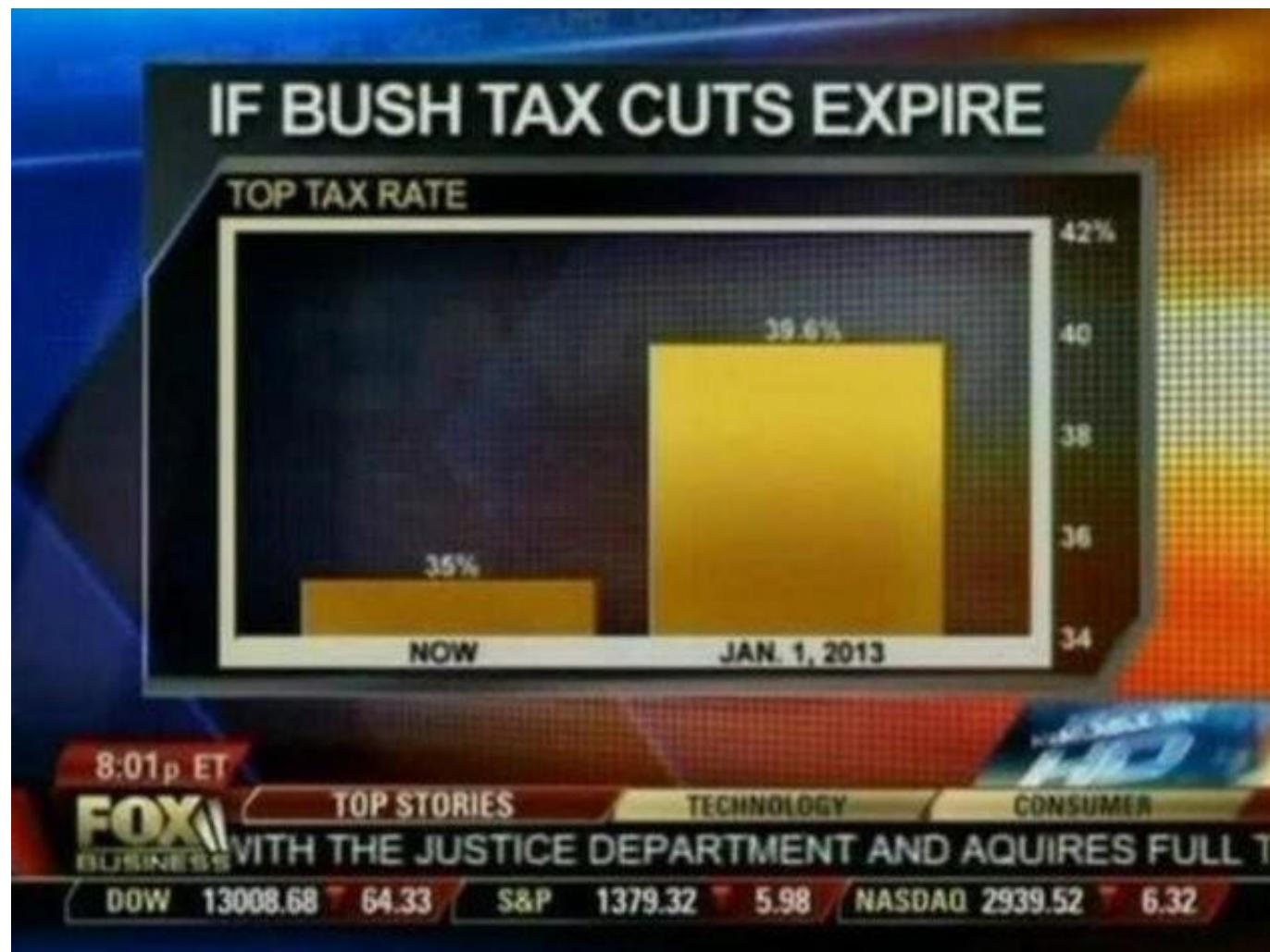
# What is wrong in this plot?



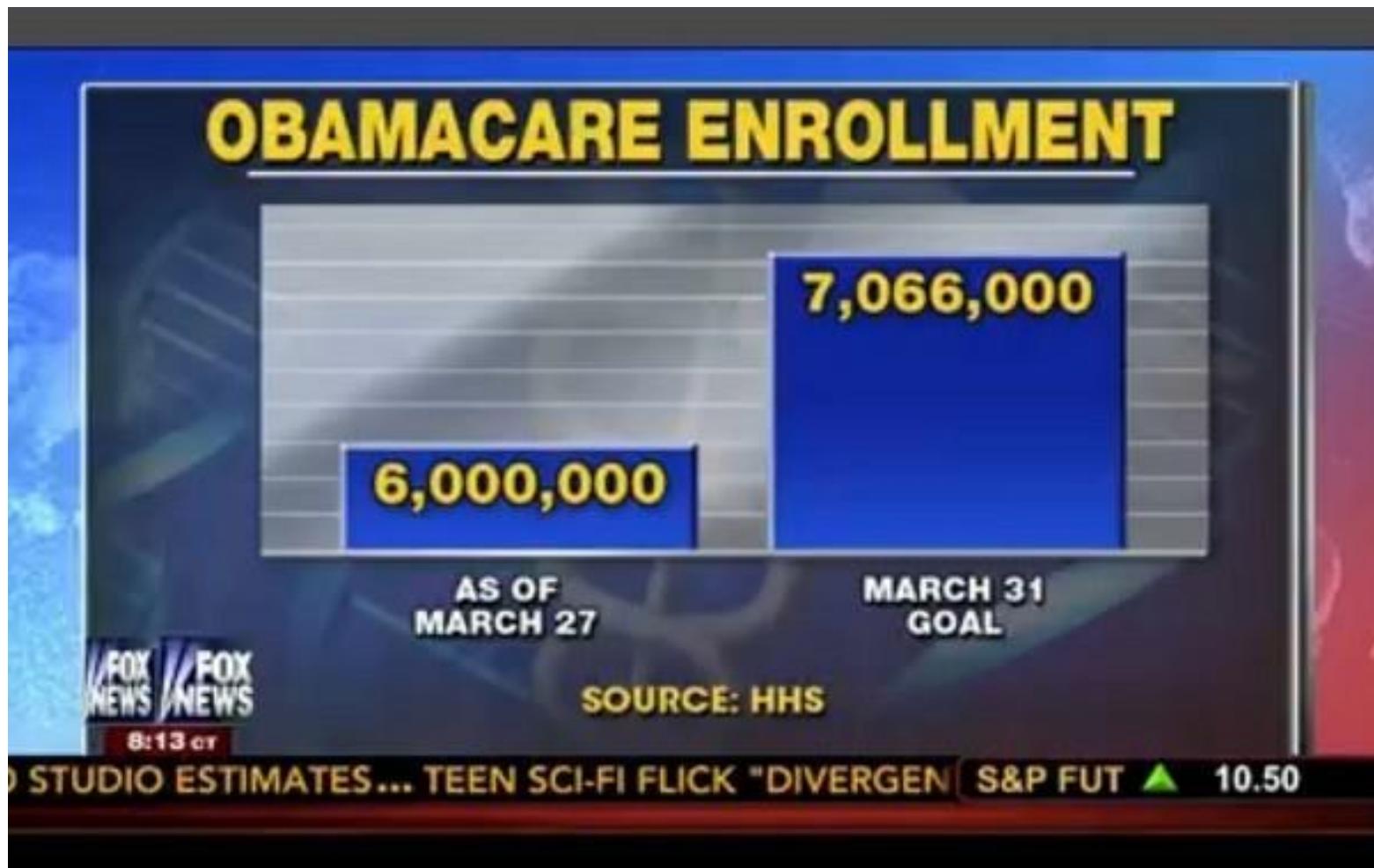
What is wrong in this plot?



What is wrong in this plot?



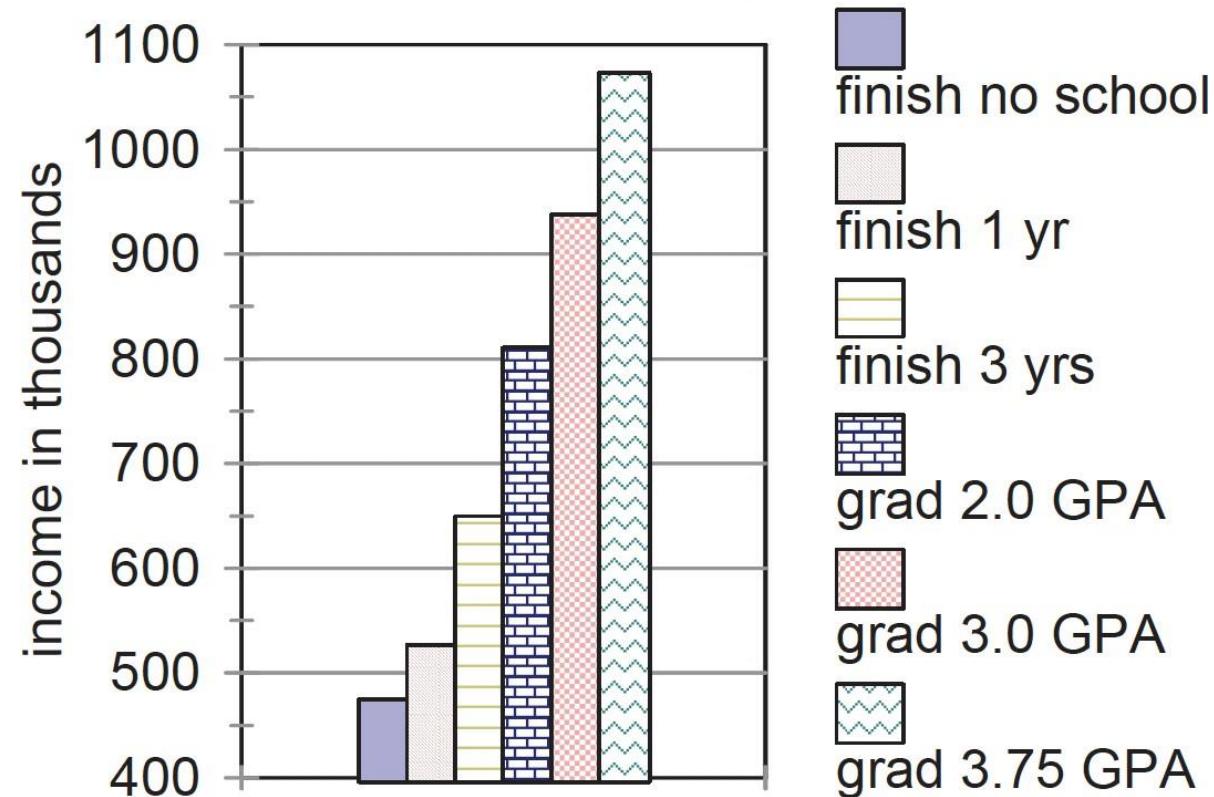
What is wrong in this plot?



# A bar plot example

**Figure 2 Discounted Expected Lifetime Earnings,  $VN(t')$**

- y-axis does not start at 0
- clutters for texture, colors
- legend instead of x-axis



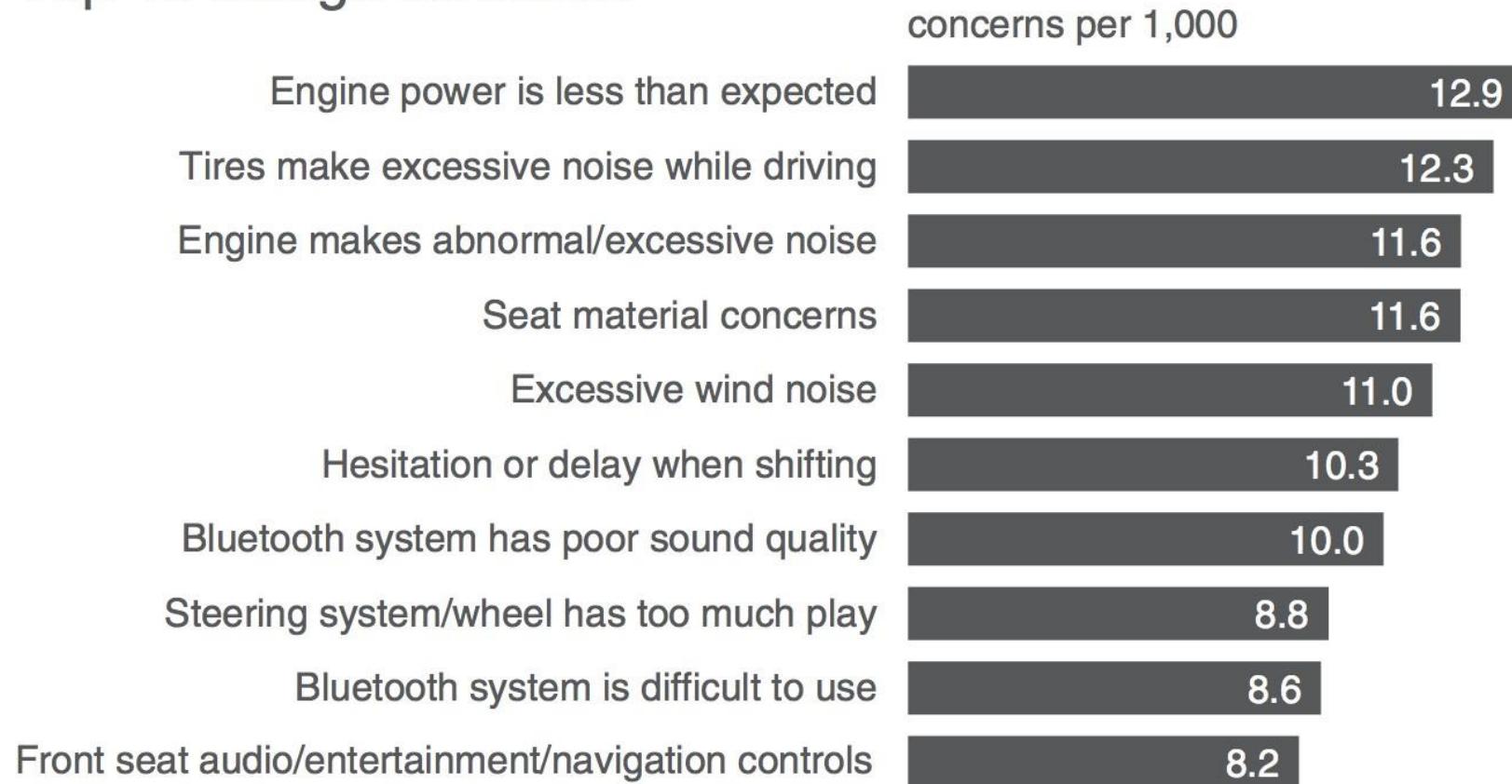
## Another bar plot example

- the same kinds of data are plotted used different types of encoding
- gradient color shading in the bar is darker at the bottom than at the top where the data are truly encoded
- too many gridlines



# Preattentive attributes in graphs

## Top 10 design concerns



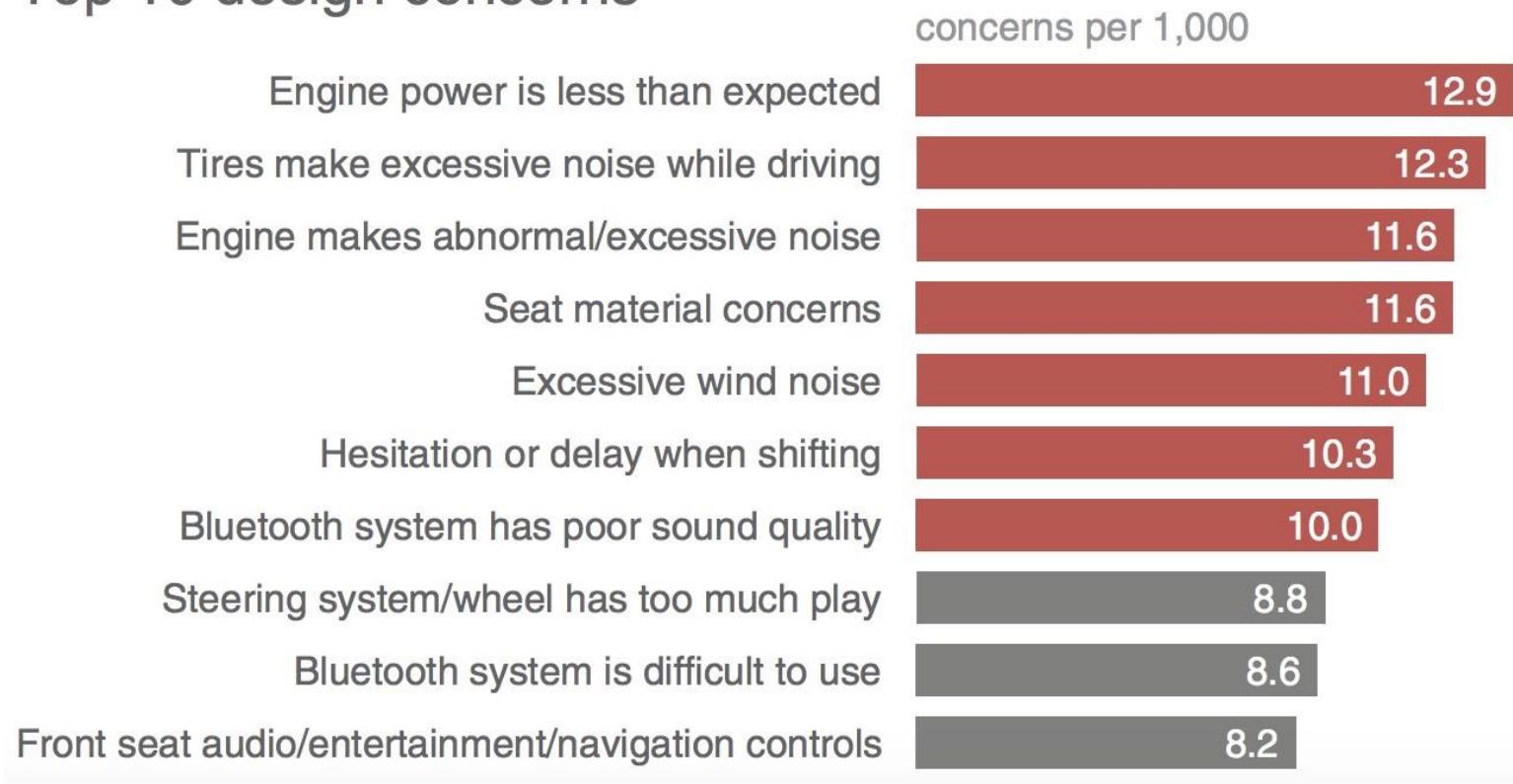
With no preattentive attributes, you need to process all of the information.

# Preattentive attributes in graphs

**7 of the top 10 design concerns have 10 or more concerns per 1,000.**

Discussion: is this an acceptable default rate?

## Top 10 design concerns

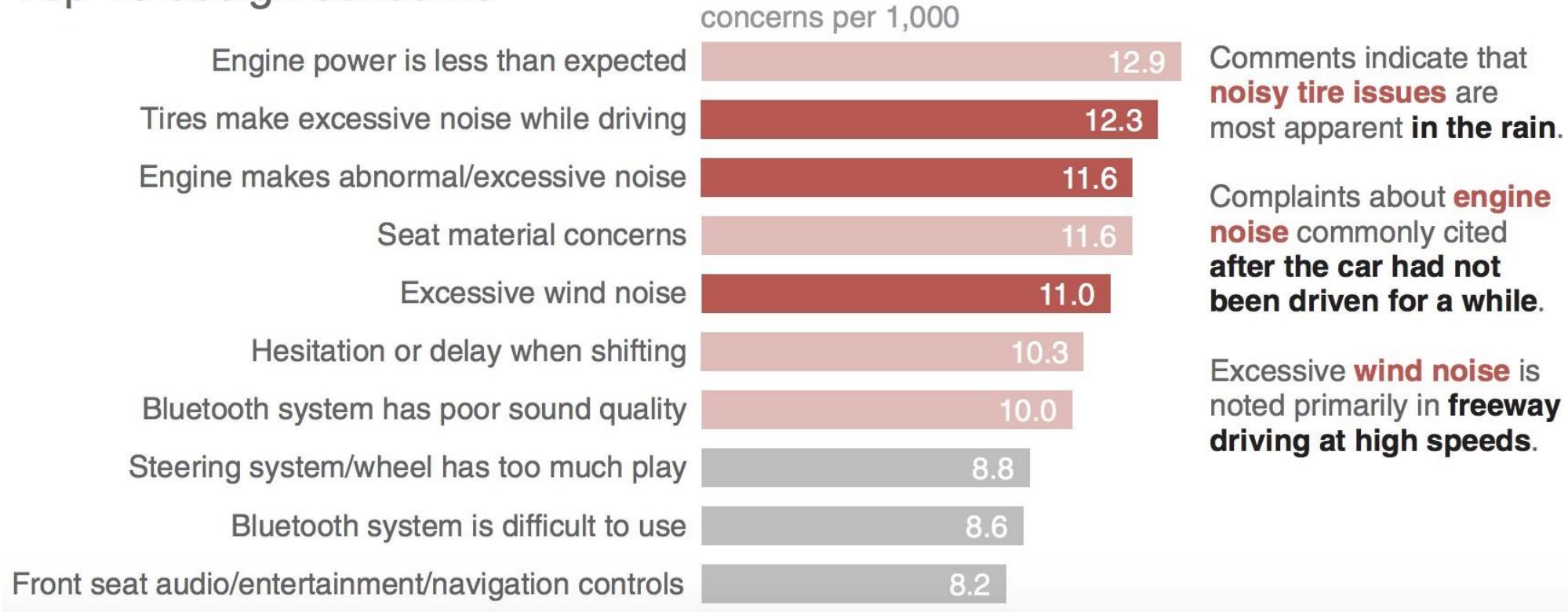


This is most likely to be created during the exploration phase, but you can leverage visual cues to highlight information at the explanation phase

# Preattentive attributes in graphs

Of the top design concerns, three are noise-related.

## Top 10 design concerns



Comments indicate that **noisy tire issues** are most apparent **in the rain**.

Complaints about **engine noise** commonly cited **after the car had not been driven for a while**.

Excessive **wind noise** is noted primarily in **freeway driving at high speeds**.

With modified color to create a visual hierarchy of information to lead audience from the macro to the micro part of the story

# Each type of chart has its own peculiarities

Bar charts

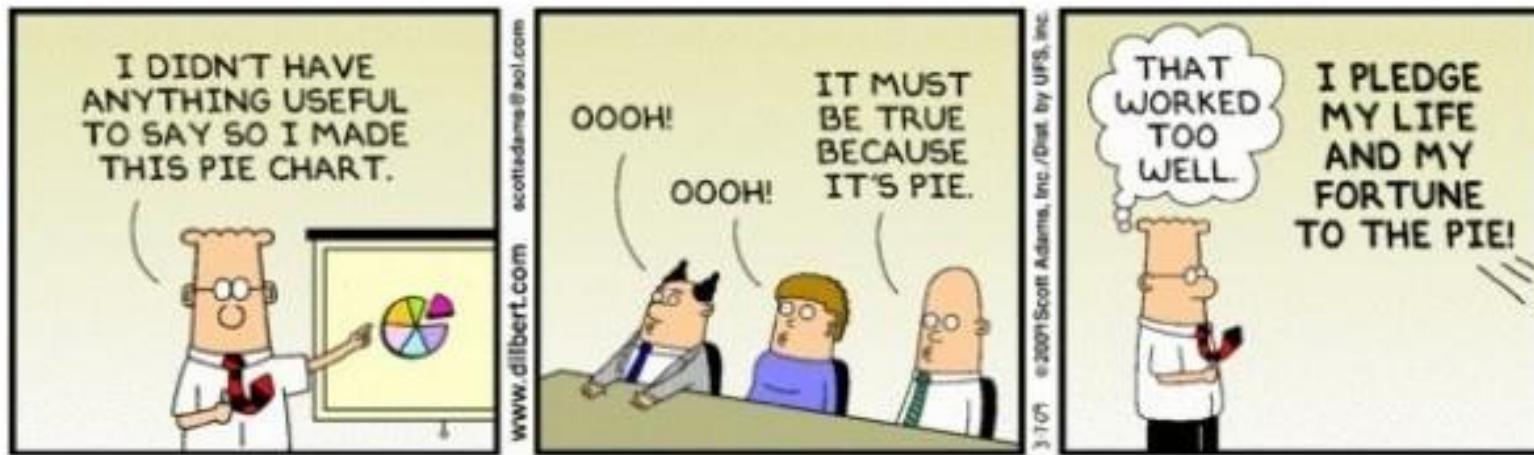
Pie charts

Line charts

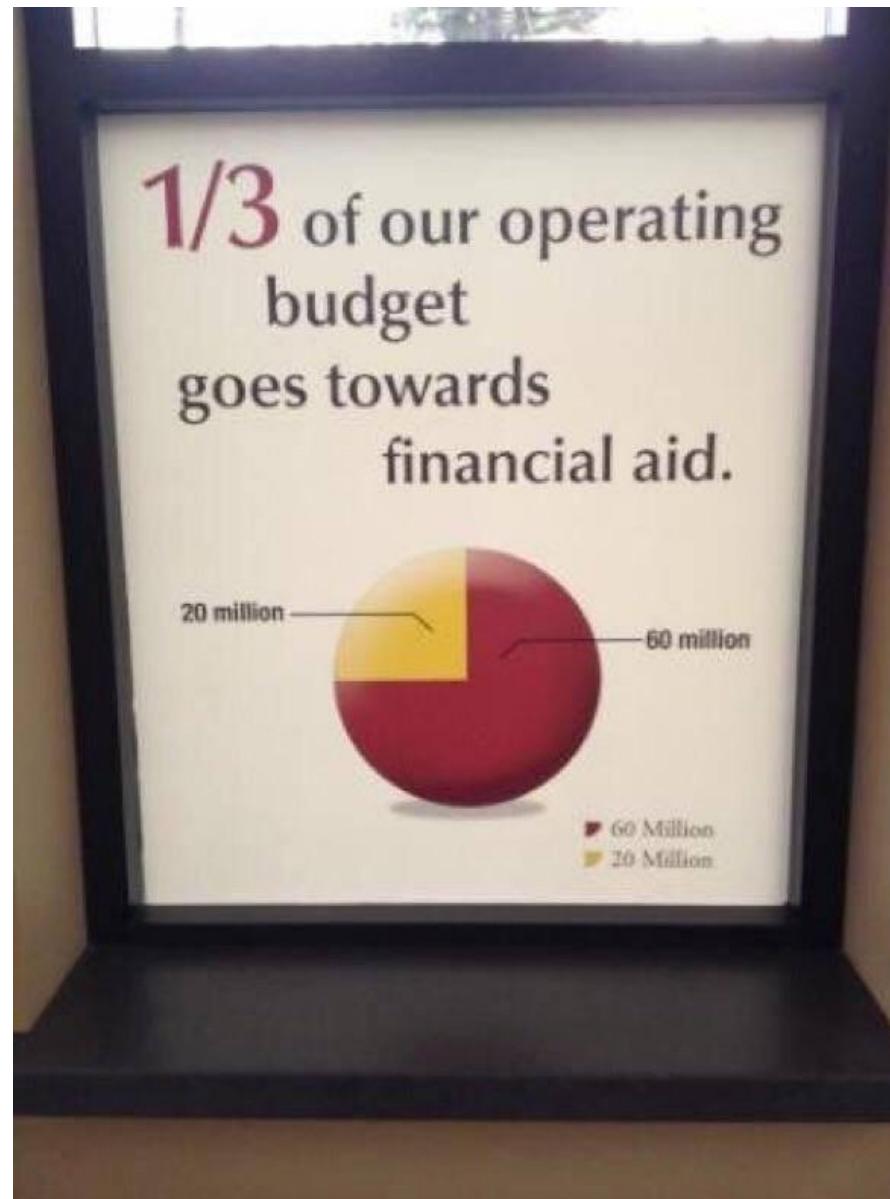
## Key takeaways for pie charts

- Humans are pretty bad at reading angles
- Do NOT compare pie charts that represent different data
- Do NOT miscalculate the pie size (sum it to 100%)
- Do NOT use a lot of small pies (better no more than 5)

# Pie chart, should I still use it?

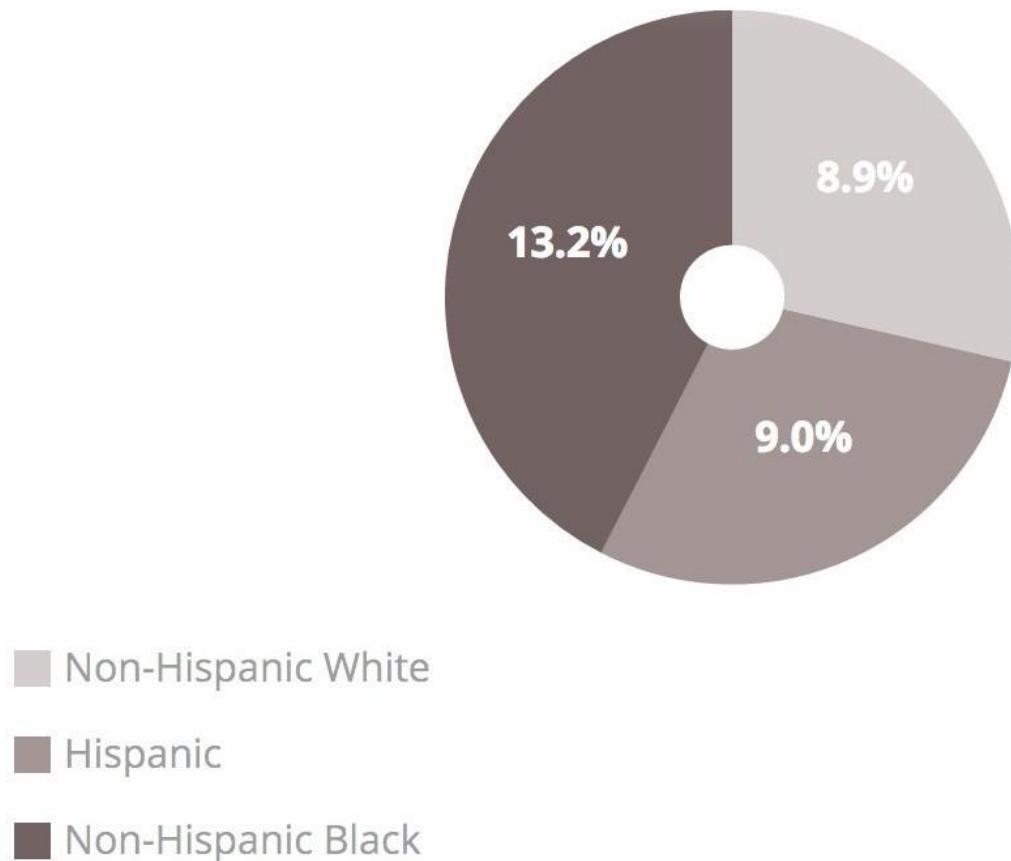


What is wrong in this plot?



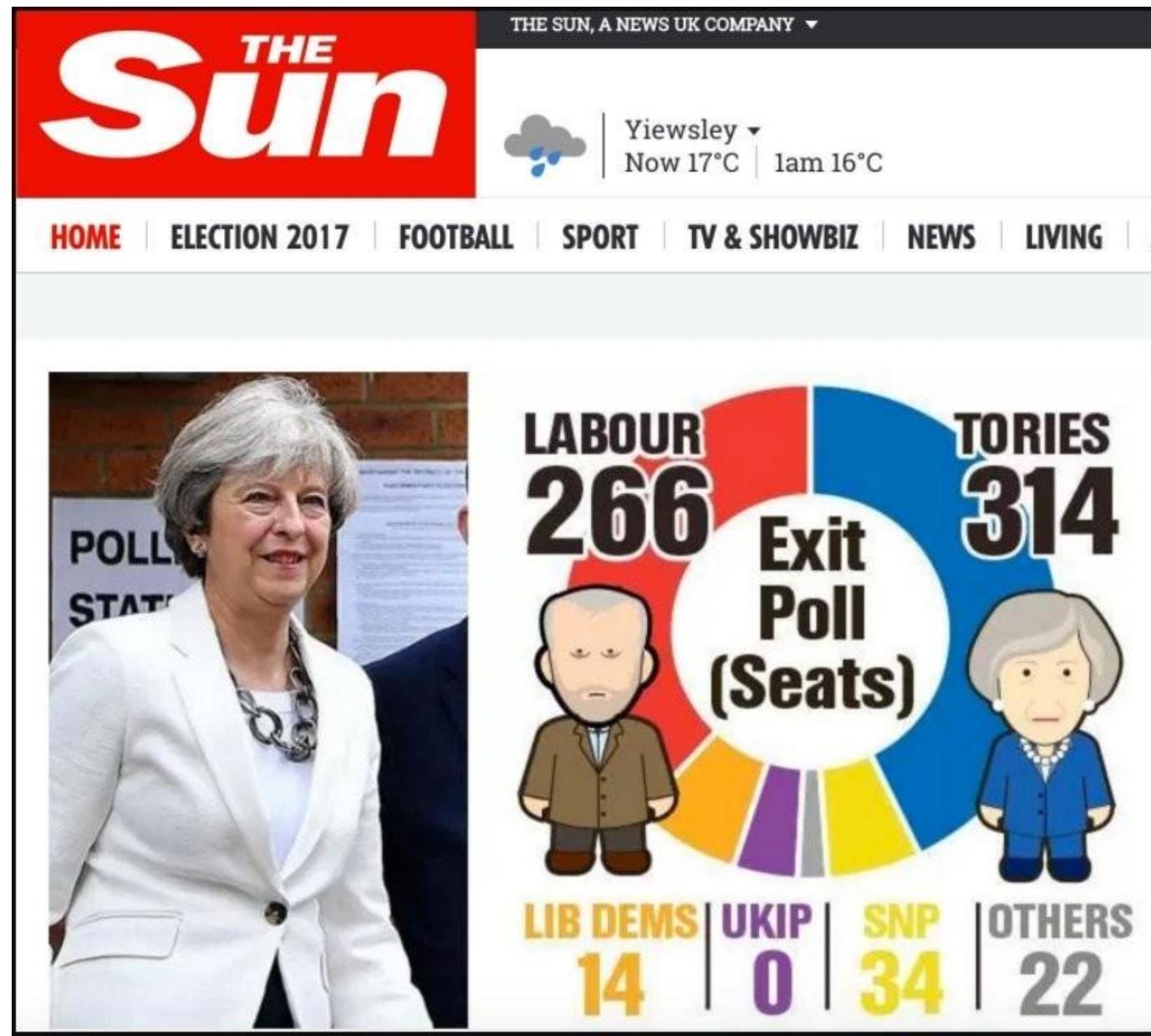
What is wrong in this plot?

### PRETERM BIRTH BY RACE & ETHNICITY



*Preterm defined as less than 37 weeks gestation.*  
Source: CDC, 2015

# What is wrong in this plot?

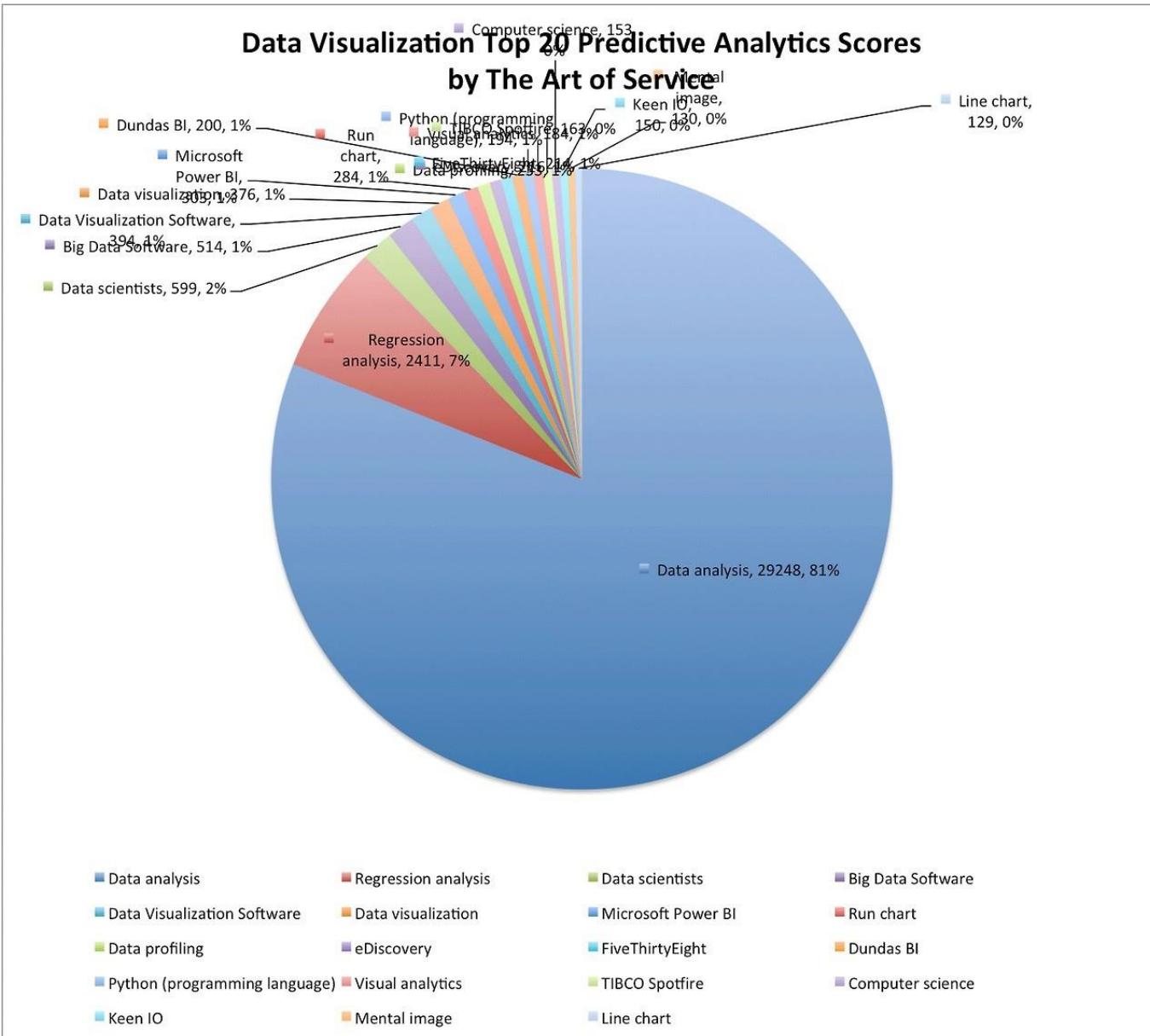


What is wrong in this plot?



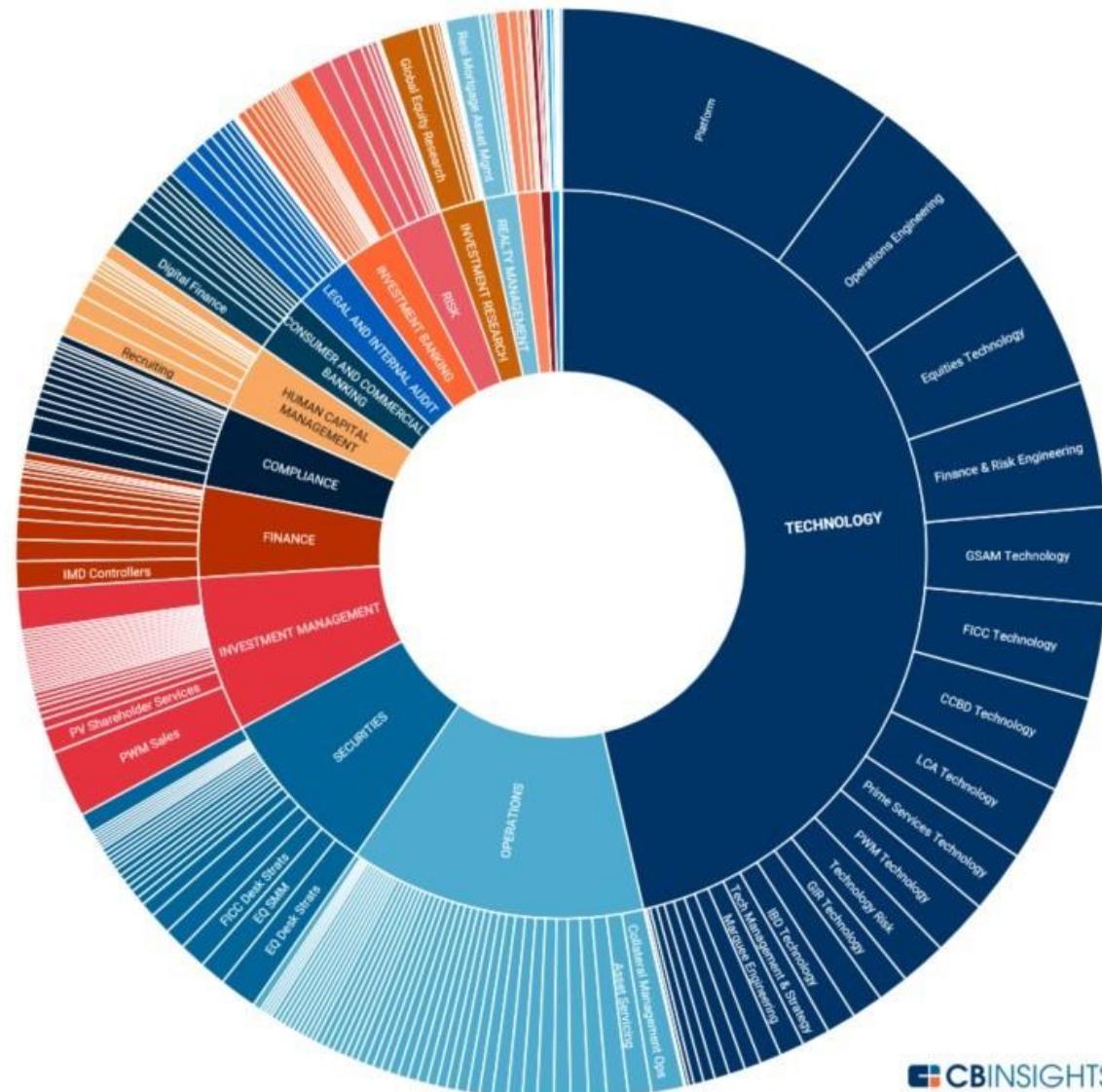
**300-500 million**  
malaria cases  
worldwide every  
year vs **90%** in **Africa**

# What is wrong in this plot?



# Can you understand this?

 **Goldman Sachs open job listings**  
Distribution of active job listings as of 9.14/2017

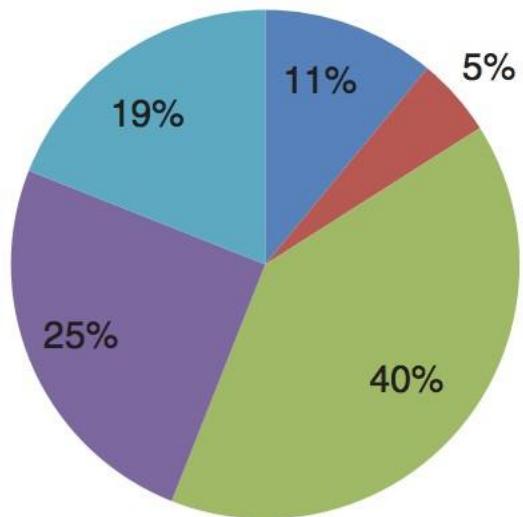


# Alternative to pies

## Survey results: summer learning program on science

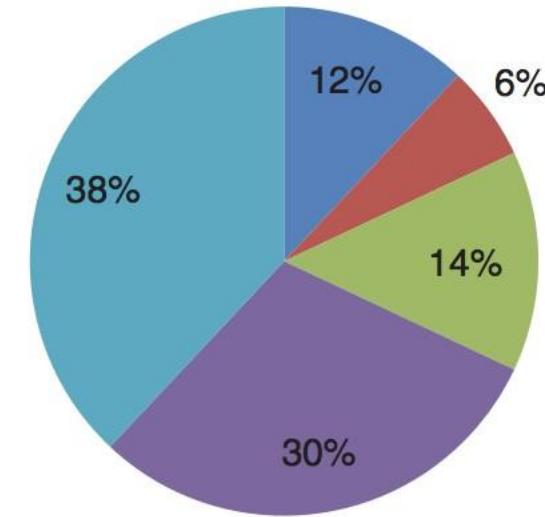
### PRE: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



### POST: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



It is rather difficult to compare segments across pies

What shall we do to improve the understanding of the audience?

## Alternative to pies

Pilot program was a success

After the pilot program,

**68%**

**of kids expressed interest towards science,**  
compared to 44% going into the program.

Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

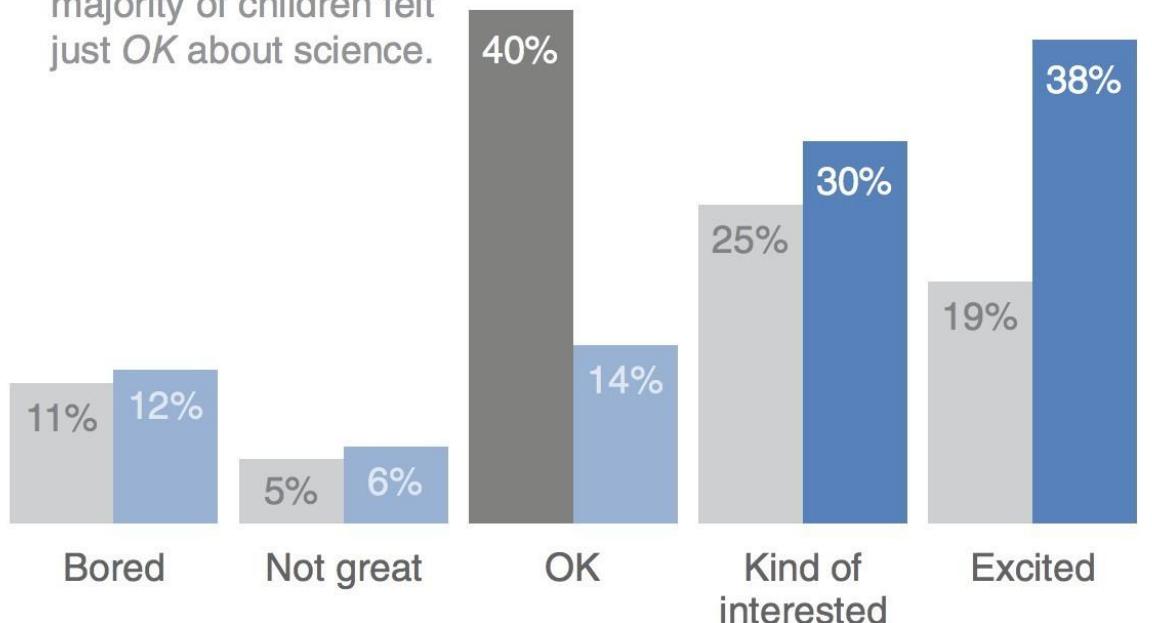
If the improvement in positive sentiment is the main message we want to impart to our audience, we can consider making that the only thing we communicate

# Alternative to pies

## Pilot program was a success

How do you feel about science?

**BEFORE** program, the majority of children felt just *OK* about science.



**AFTER** program, more children were *Kind of interested* & *Excited* about science.

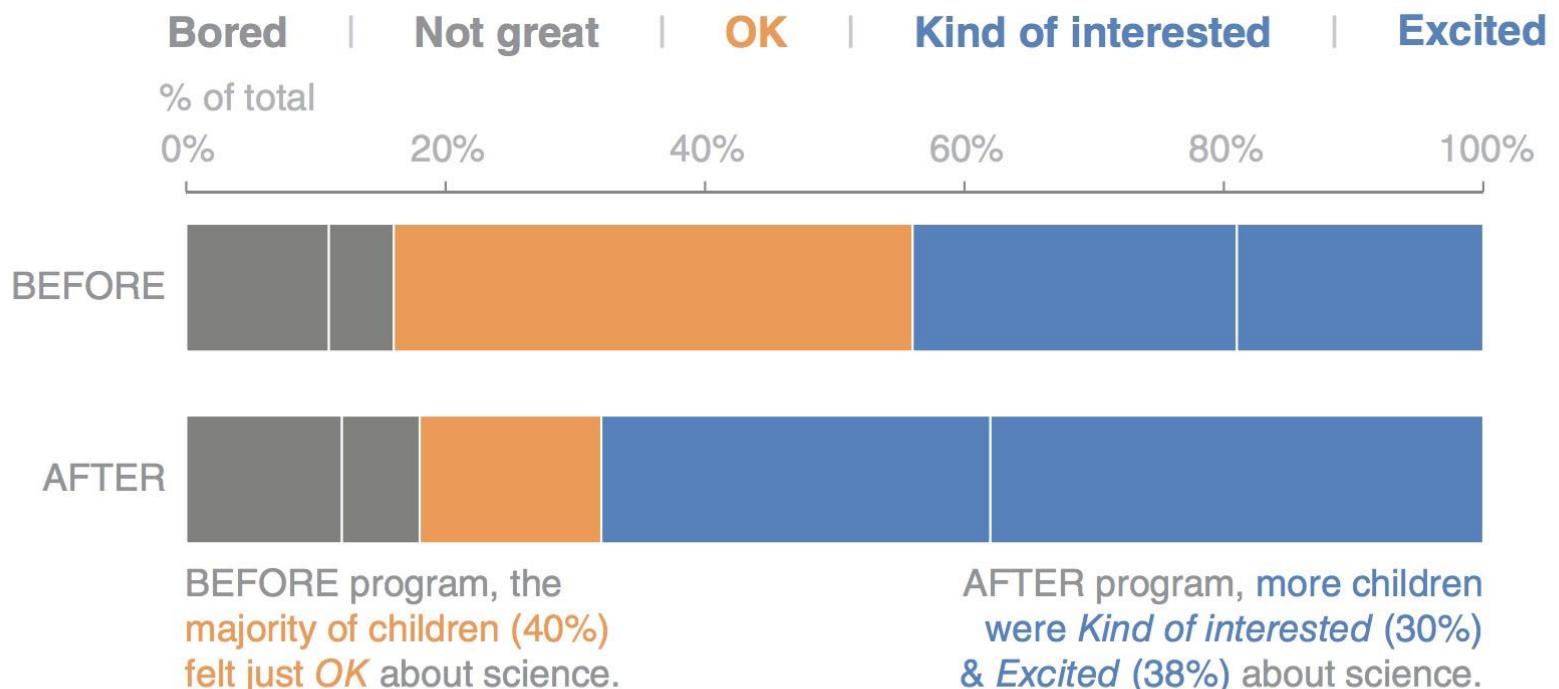
Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

The simple bar plot allows for easy comparison between two categories  
Colors and texts may further help highlight information for the audience

# Alternative to pies

## Pilot program was a success

How do you feel about science?



Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

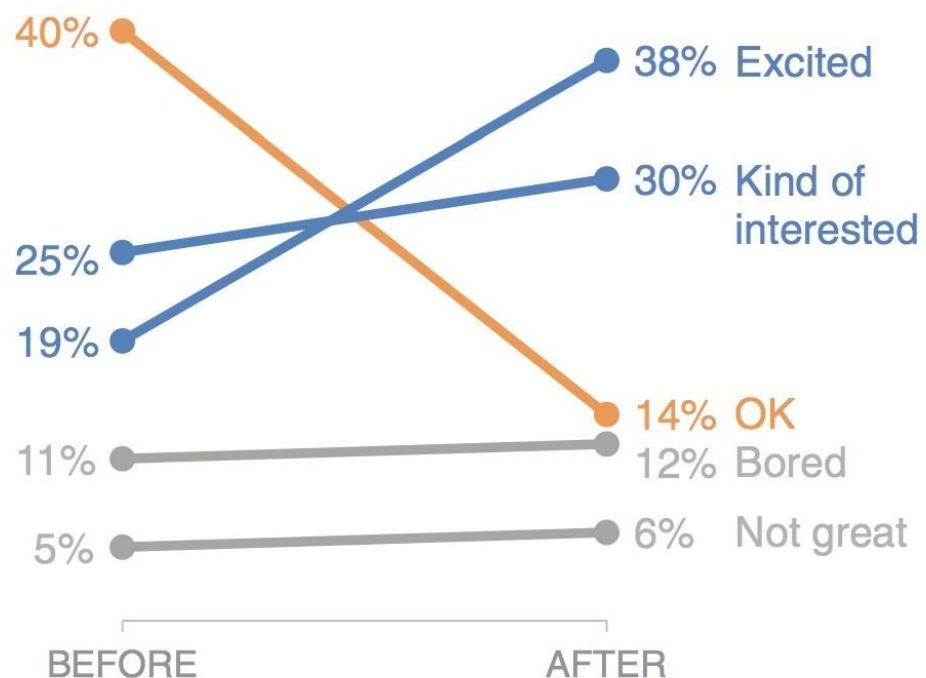
When the part-to-whole concept is important, stacked bar plot is useful

Audience can compare the negative segments at the left and positive segments at the right

# Alternative to pies

## Pilot program was a success

How do you feel about science?



BEFORE program, the majority of children felt just **OK** about science.

AFTER program, more children were *Kind of interested* & *Excited* about science.

Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Slopegraph shows the visual percentage change and visual ordering of categories from the most to least

# Each type of chart has its own peculiarities

Bar charts

Pie charts

Line charts

## Key takeaways for line charts

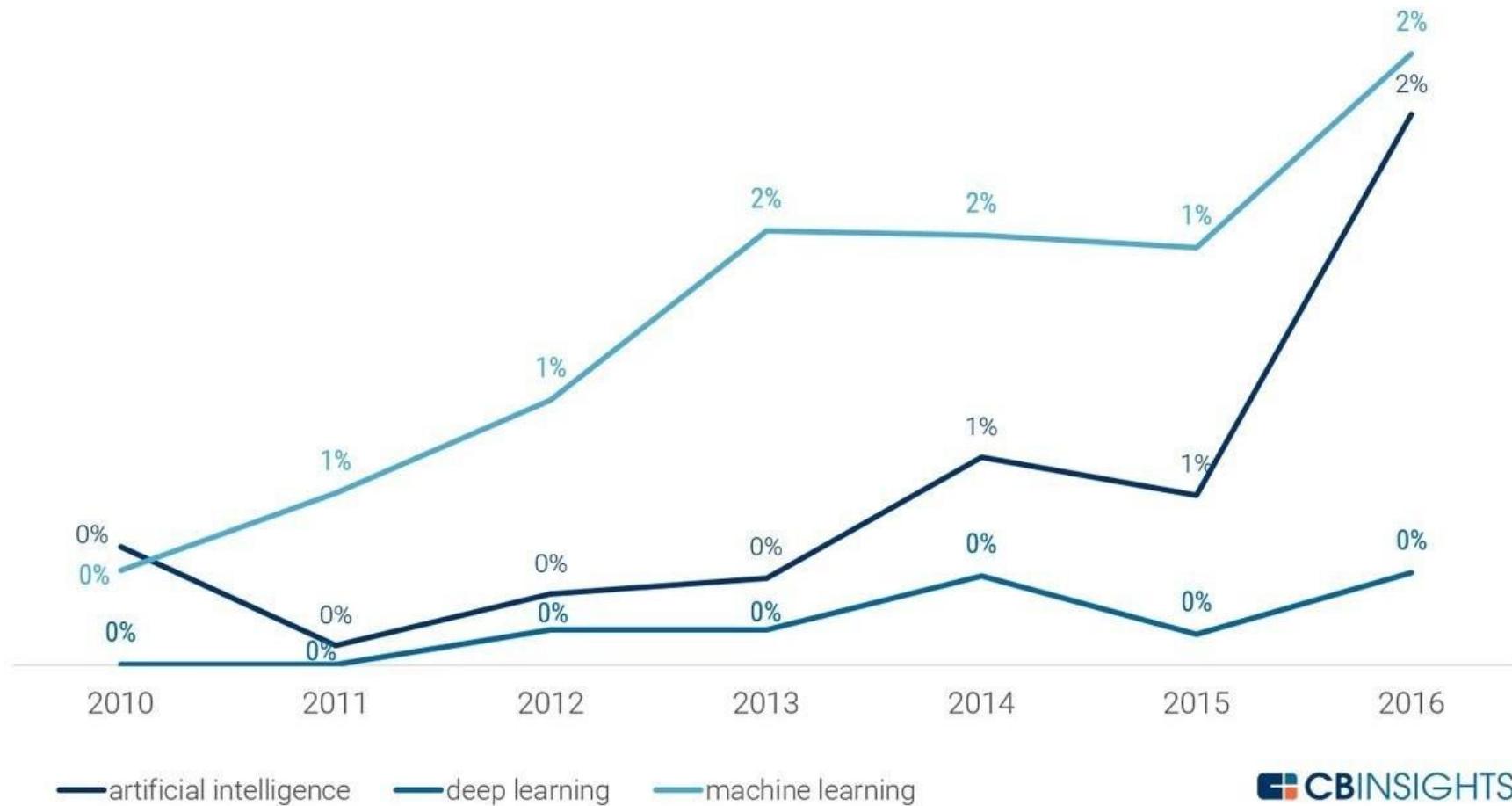
- By its design, line graph emphasizes the change in the dependent variable (usually the y value) as the independent variable (usually the x value) changes
- Do NOT use line chart if the x value does not change or belong to different categories
- Do NOT use line chart to represent y-values at different scale (e.g. you may consider two y-axes, but with caution)

# What is wrong in this plot?

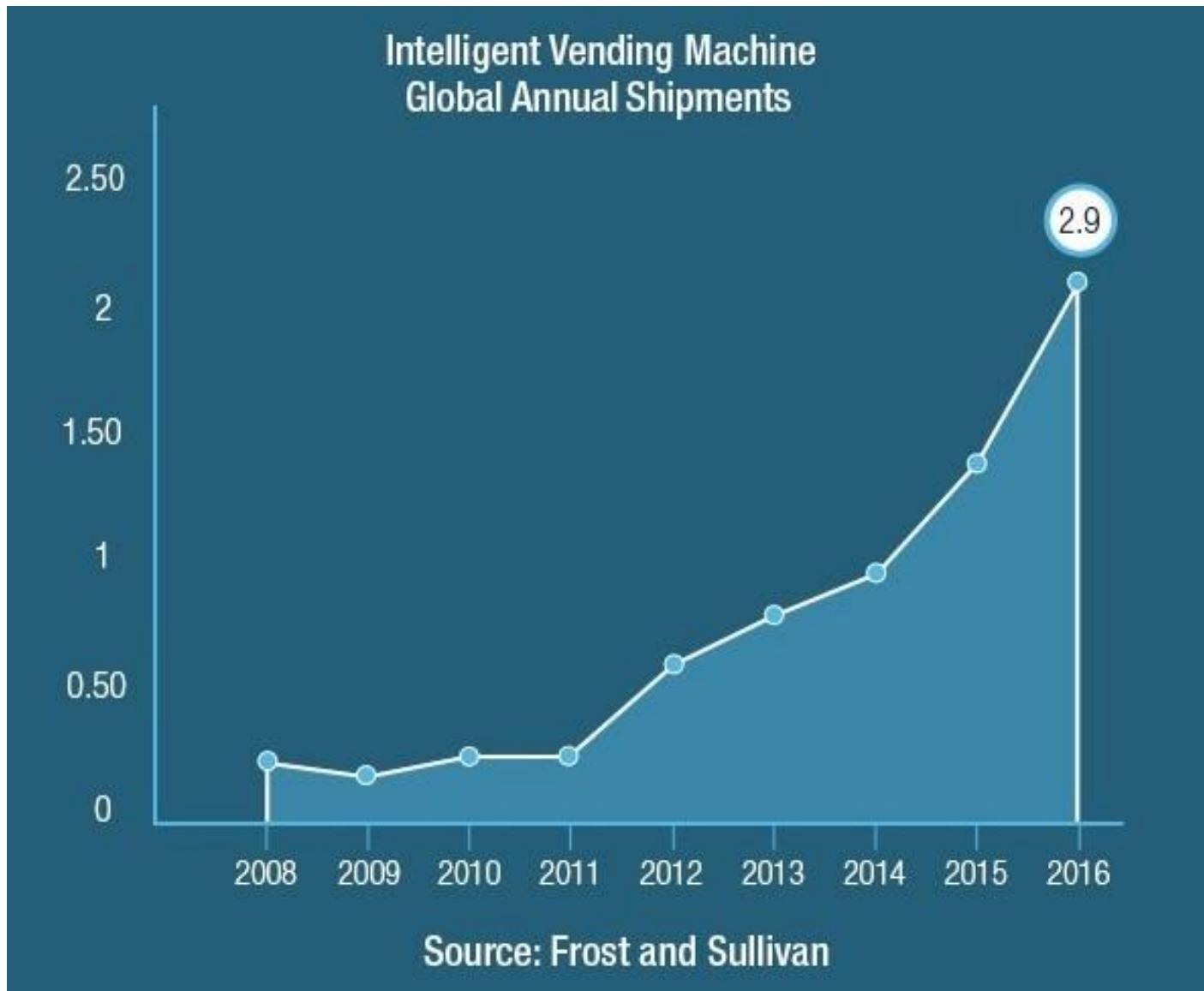


## LEARNING HOW TO LEARN

Percentage of VC-backed companies with certain words in their company description over time



What is wrong in this plot?



What is wrong in this plot?

# Save \$10,000s, Go Solar

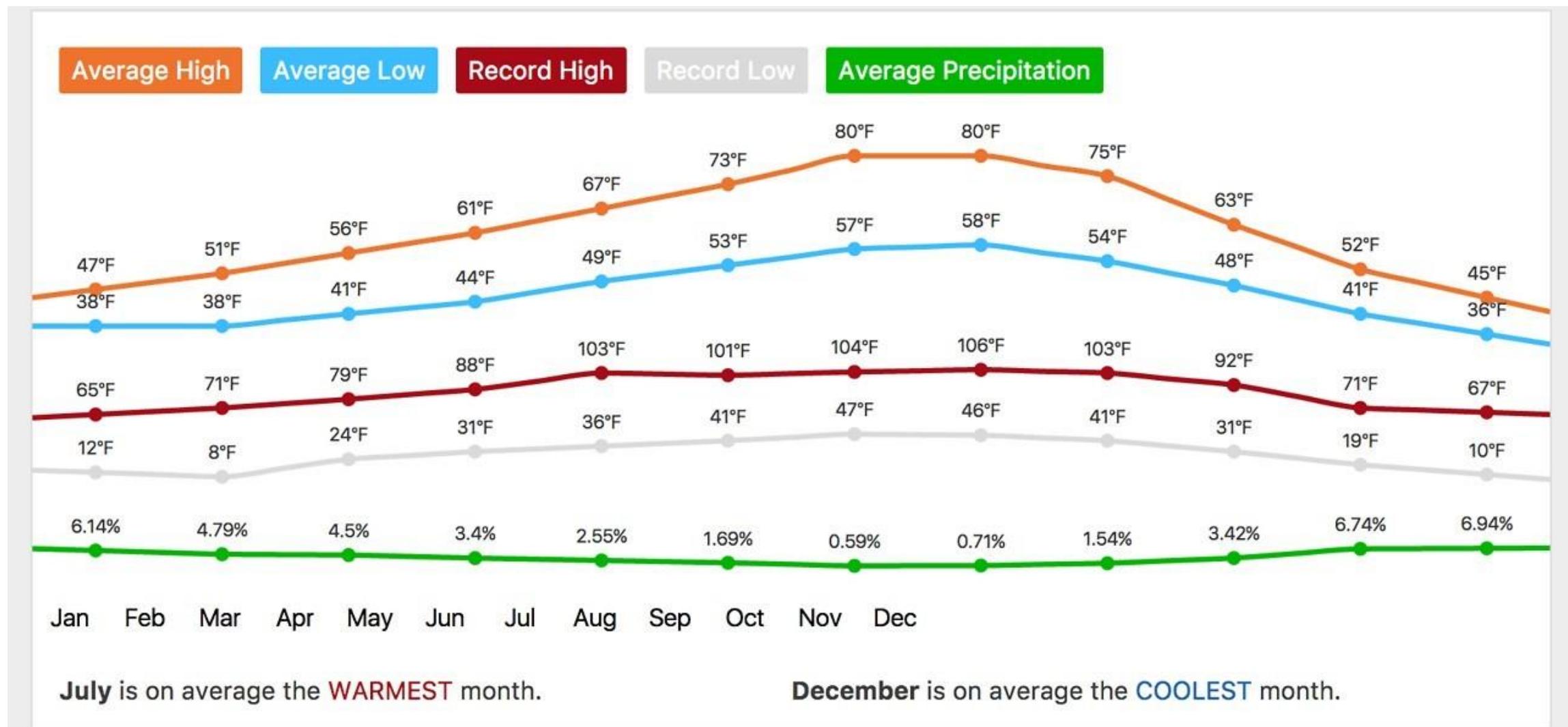
“ ”

**"I'd put my money on the sun and solar energy."**

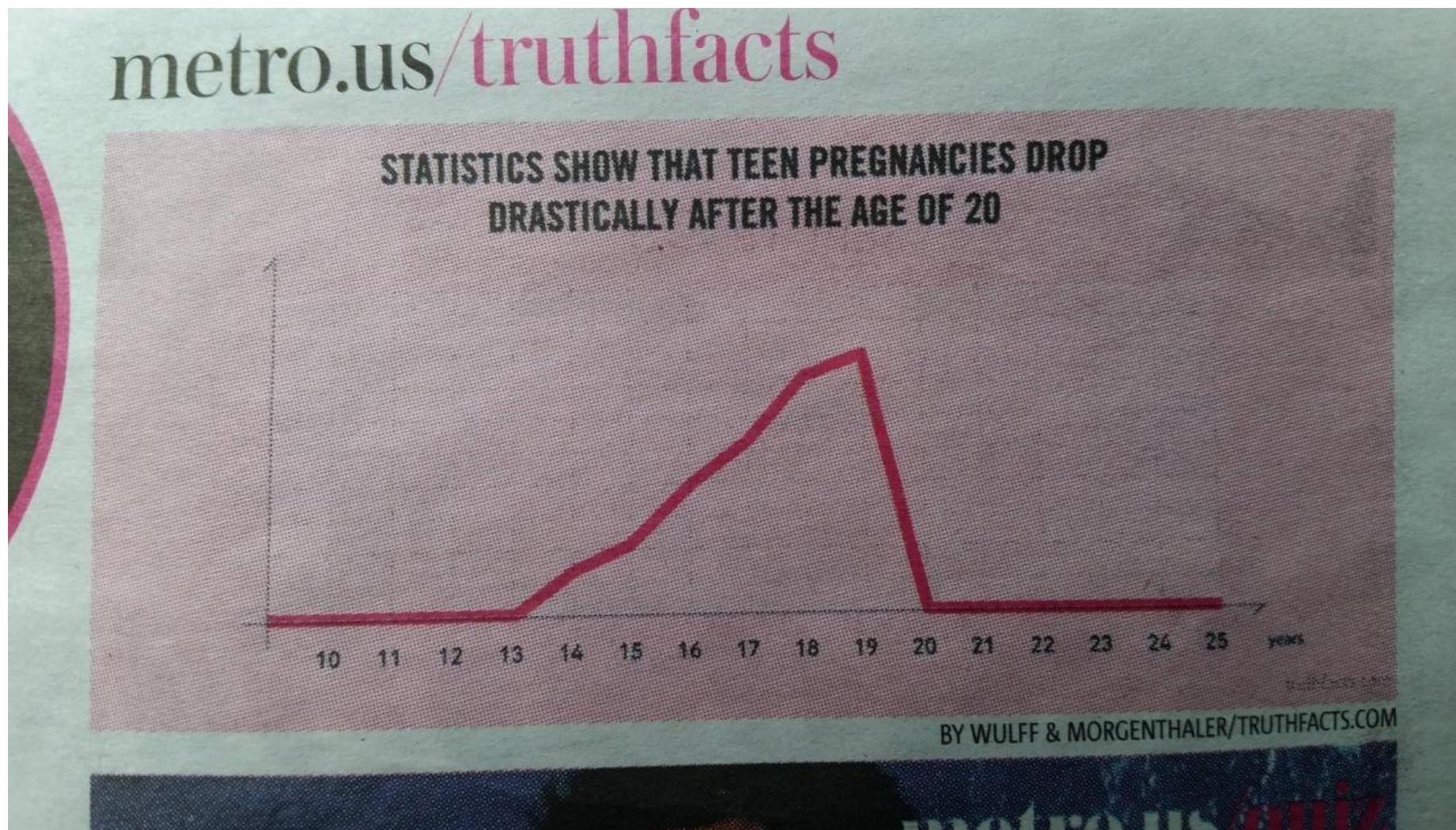
Thomas Edison



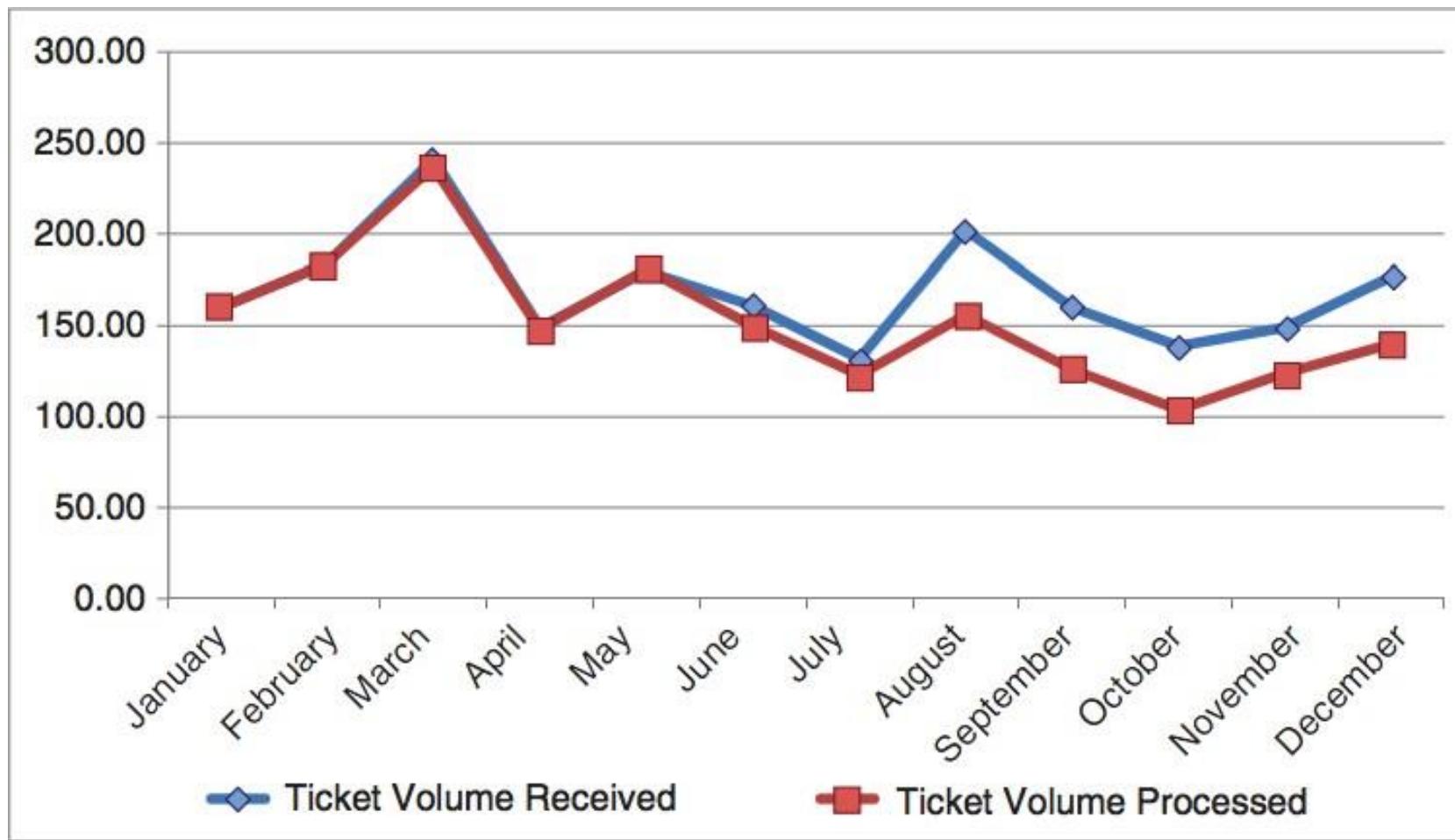
# What is wrong in this plot?



What is wrong in this plot?



## Typical starting point...



There is some evidence your team's productivity is suffering from being short-staffed

## Decluttering (1): remove the chart border

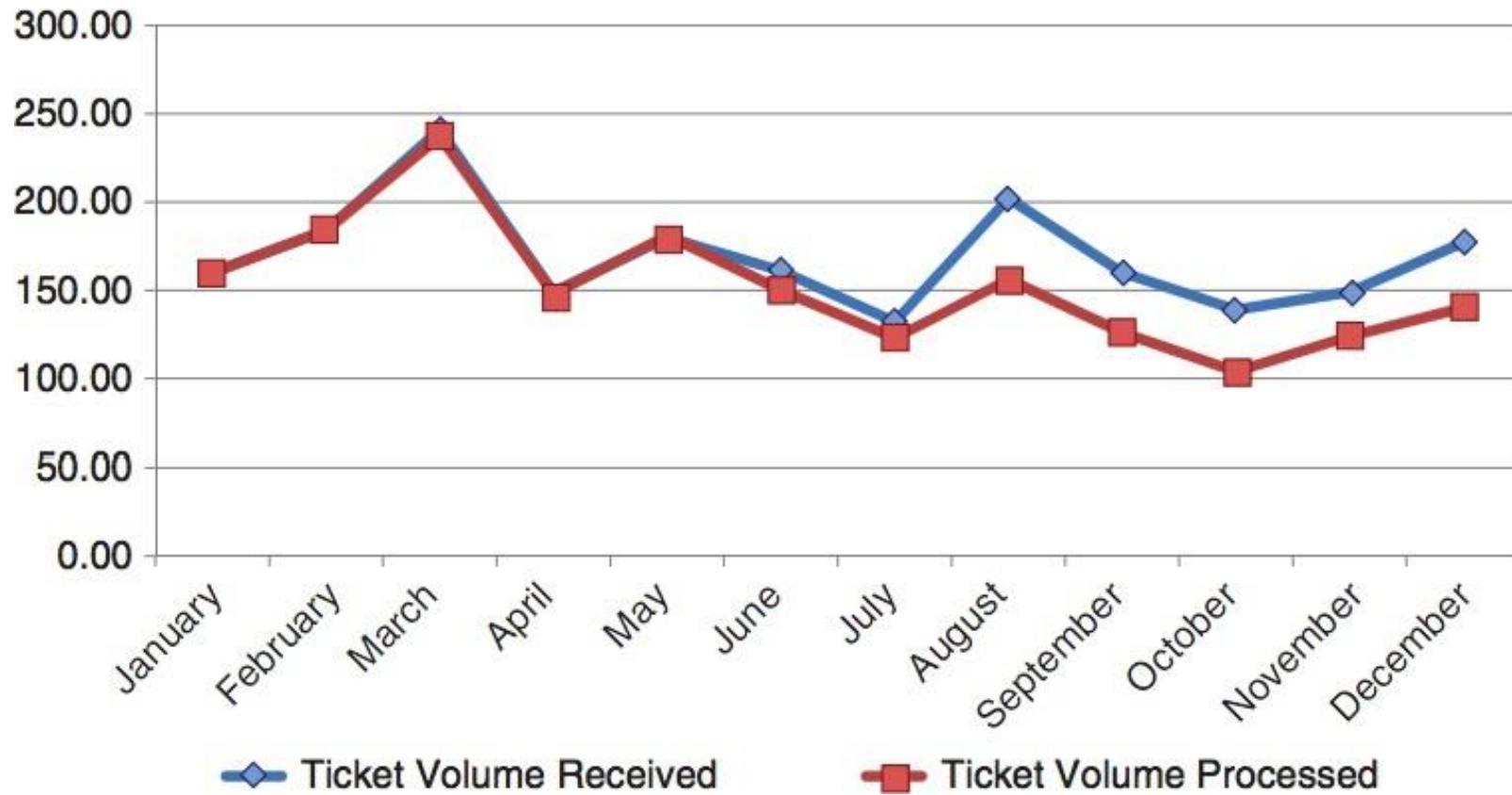
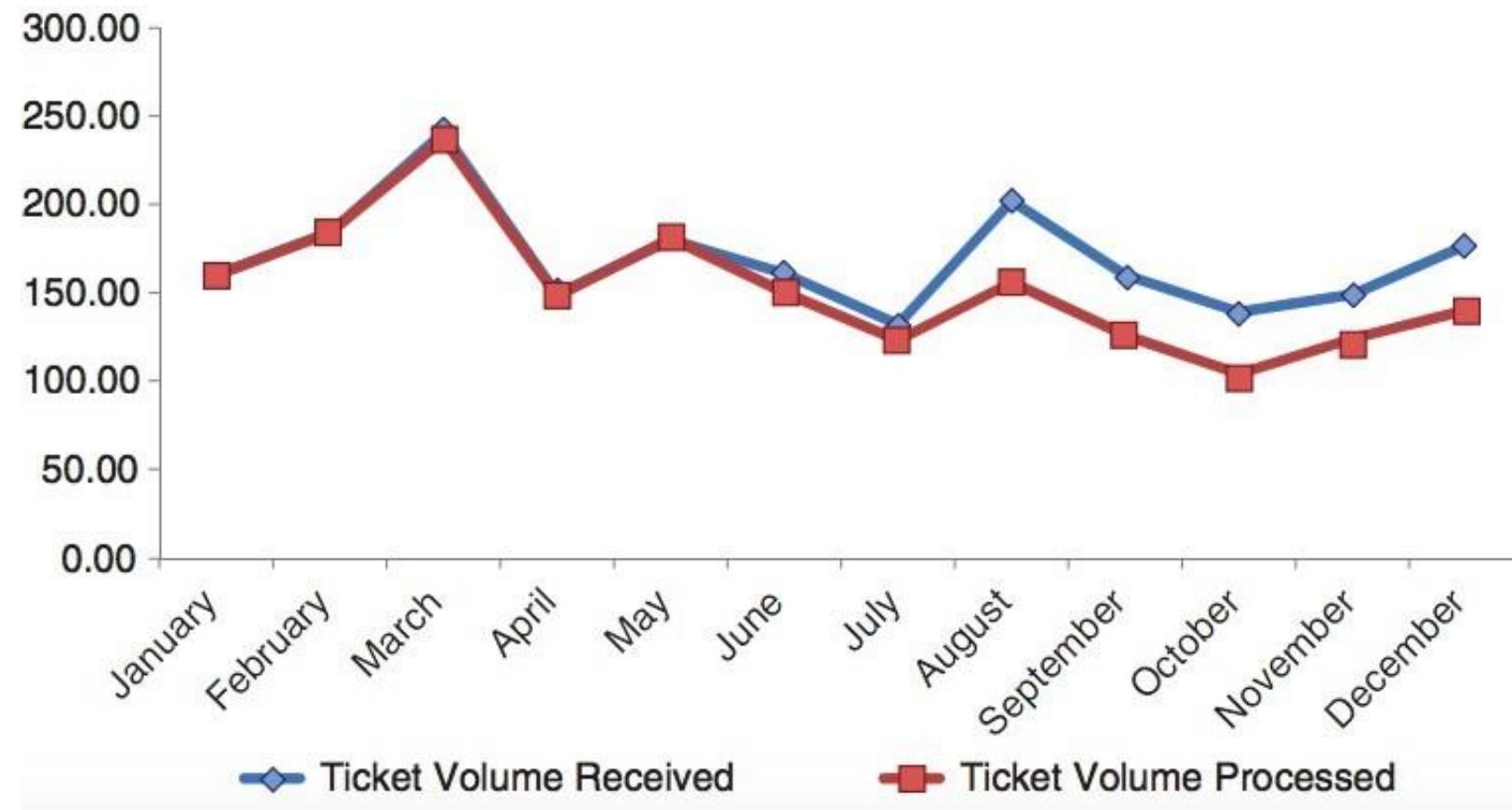


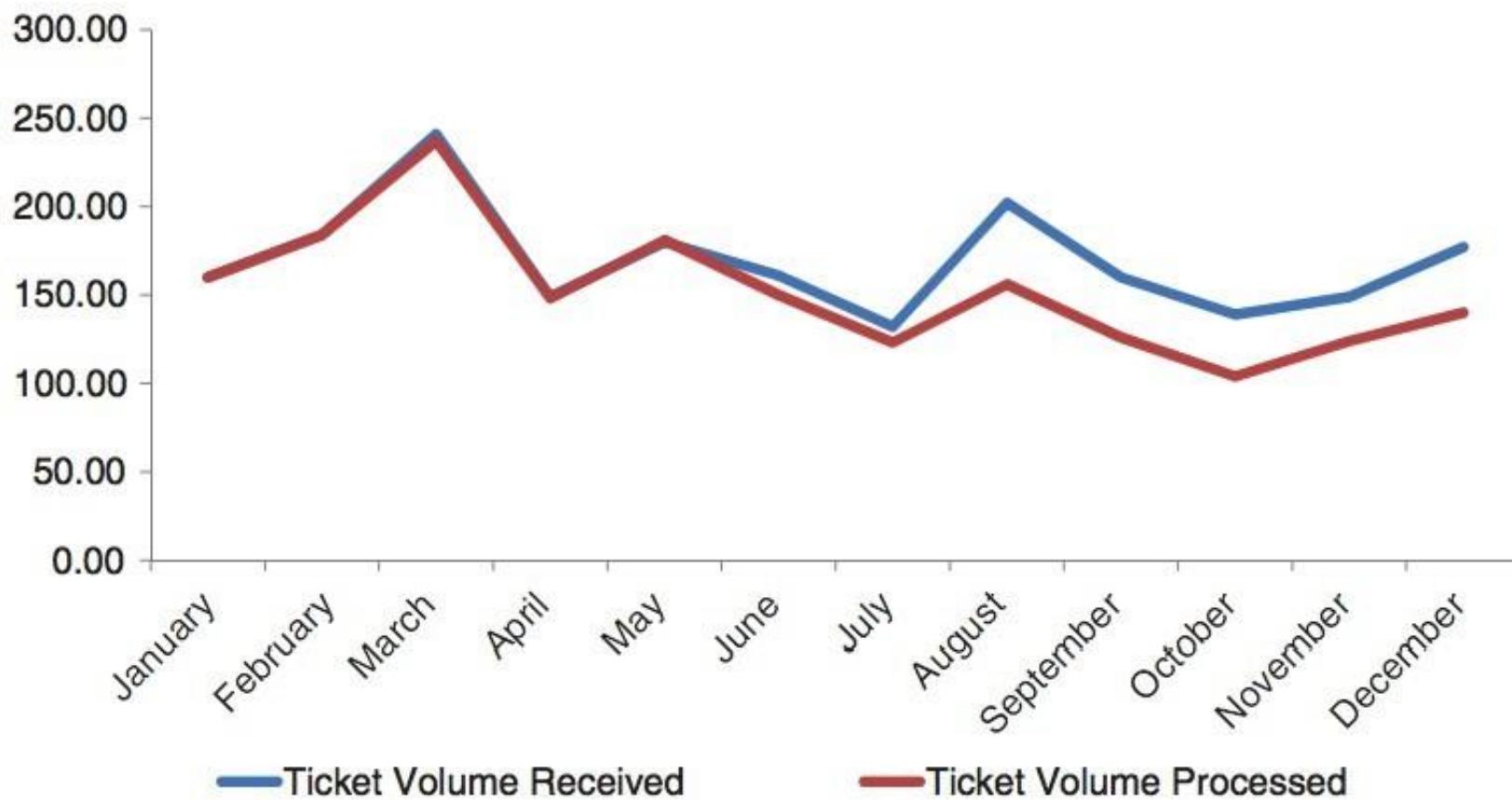
Chart borders are usually unnecessary, as covered in the Gestalt principle of closure

## Decluttering (2): remove gridlines



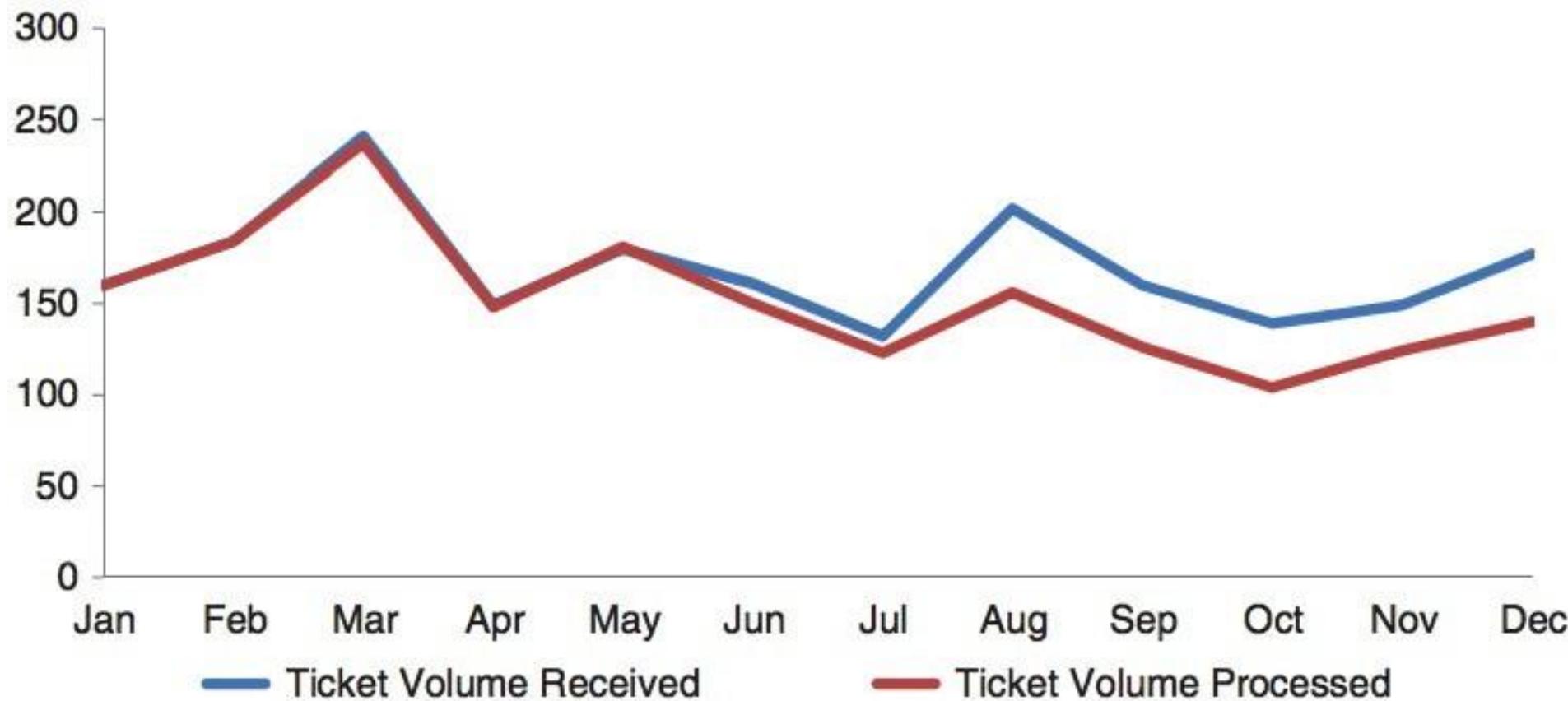
Do not let gridlines compete visually with your data and ensure the data will stand out

## Decluttering (3): remove data markers



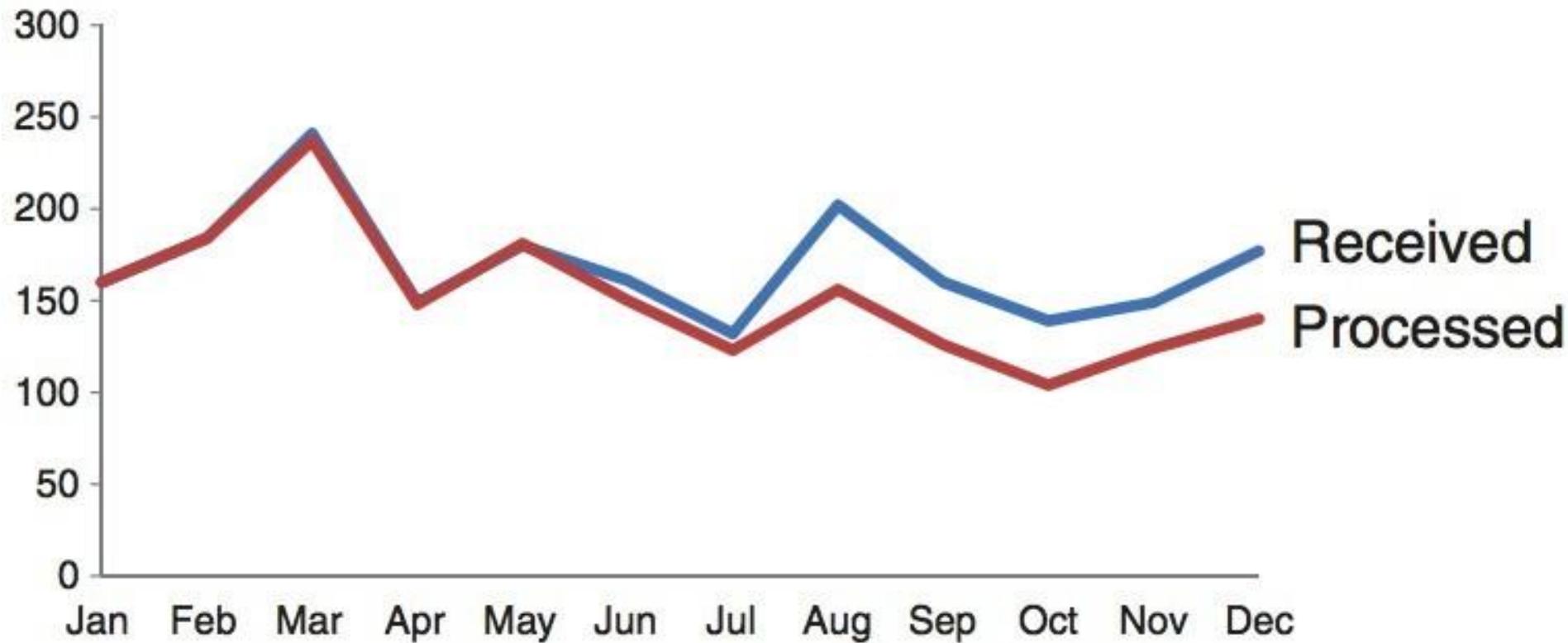
Use only markers on purpose, or leave out cognitive load to process data already depicted visually with the lines

## Decluttering (4): clean up axis labels



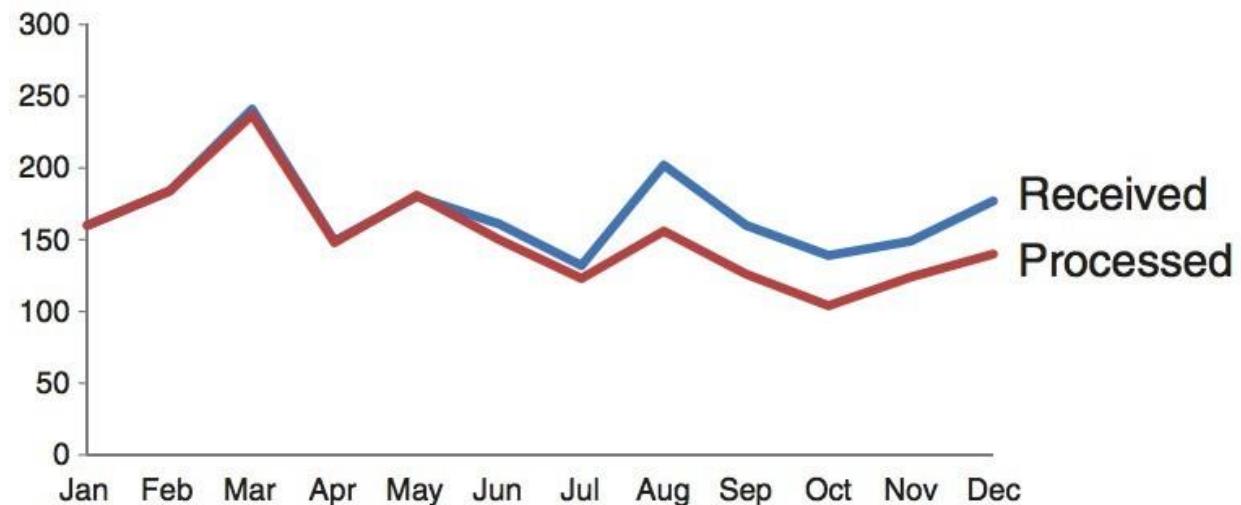
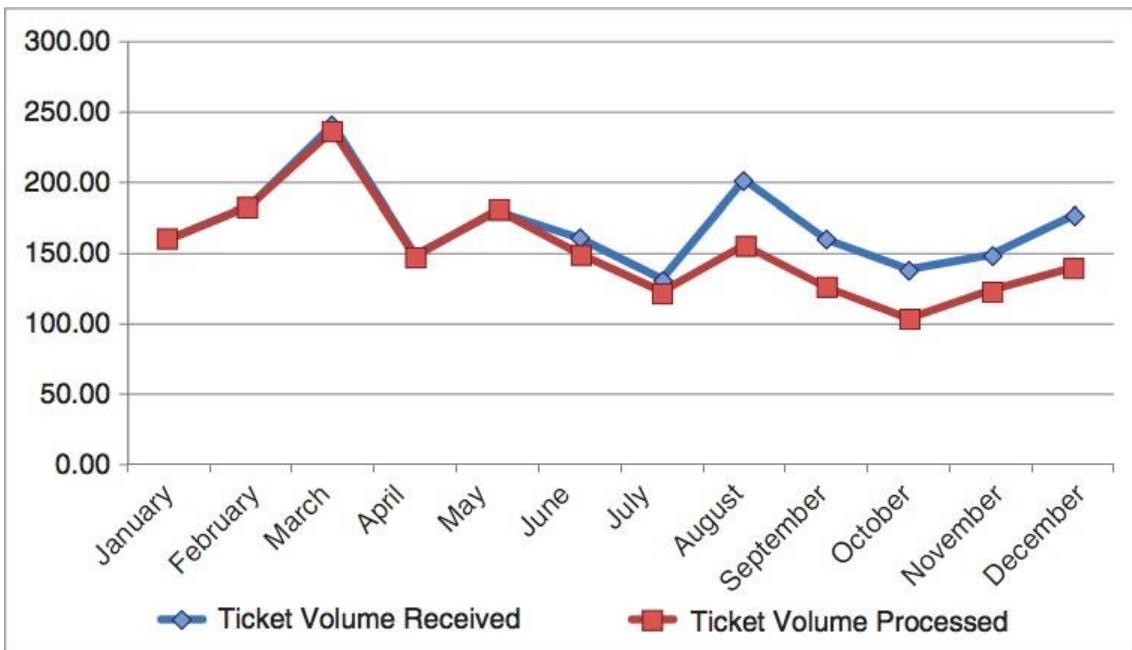
Decimal values on y-axis labels carry no informative value, and yet just make the numbers look more complicated than they are

## Decluttering (5): label data directly

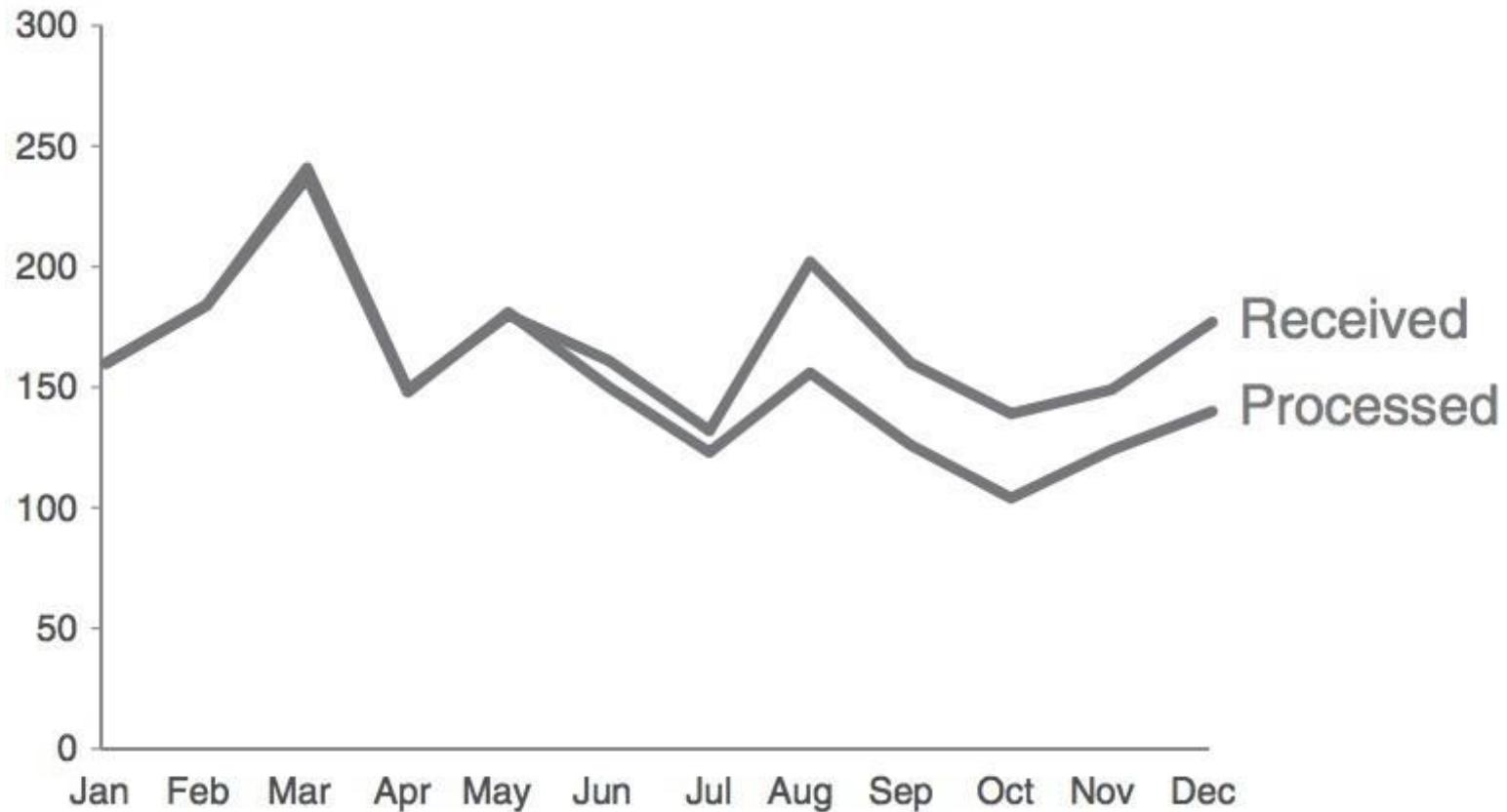


Leverage the Gestalt principle of proximity and put the data labels right next to the data

# Before/After Decluttering

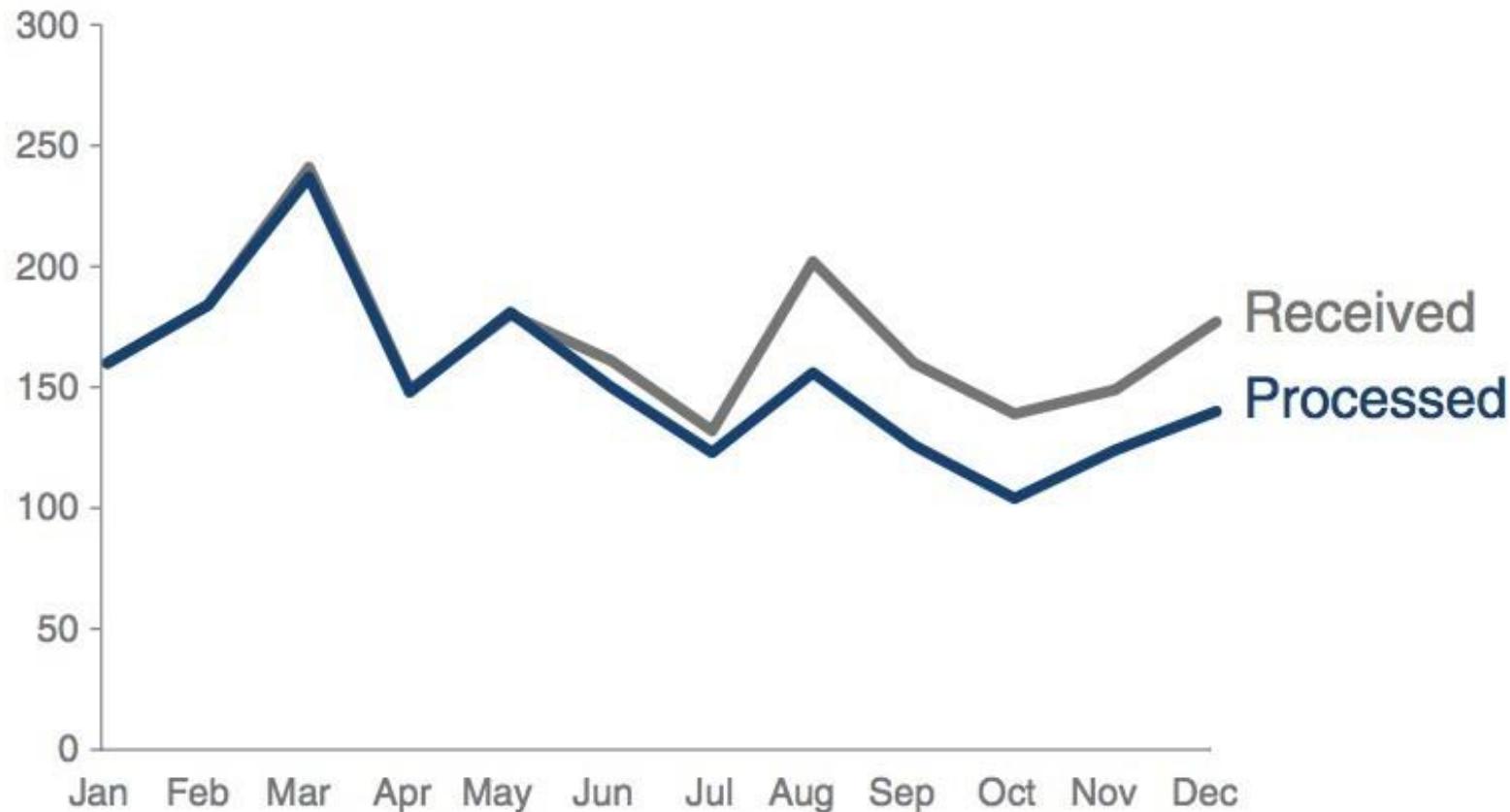


## Focus audience's attention using preattentive features



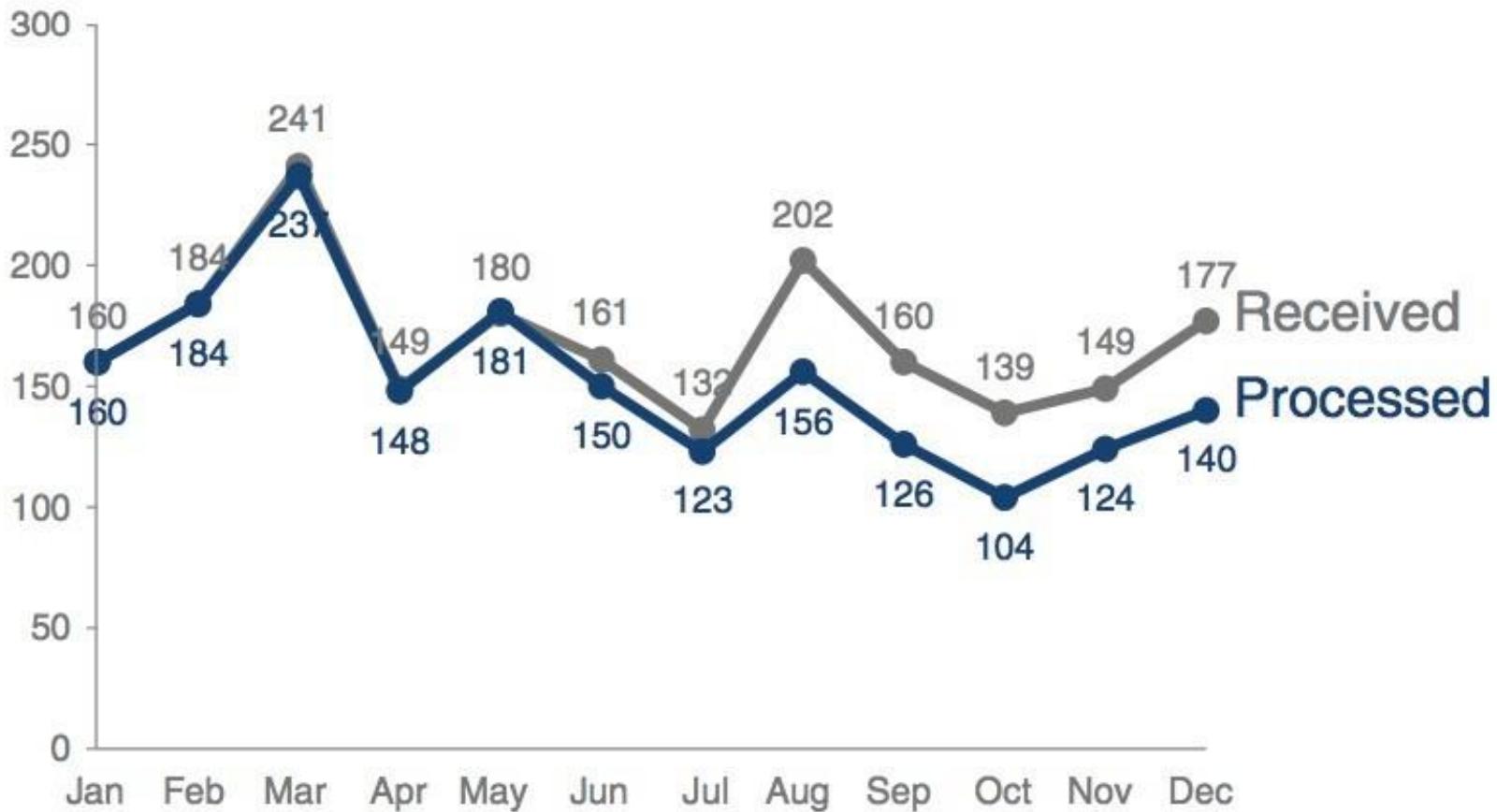
First, push everything to the background to decide what should bring to the forefront or highlight

## Focus audience's attention using preattentive features



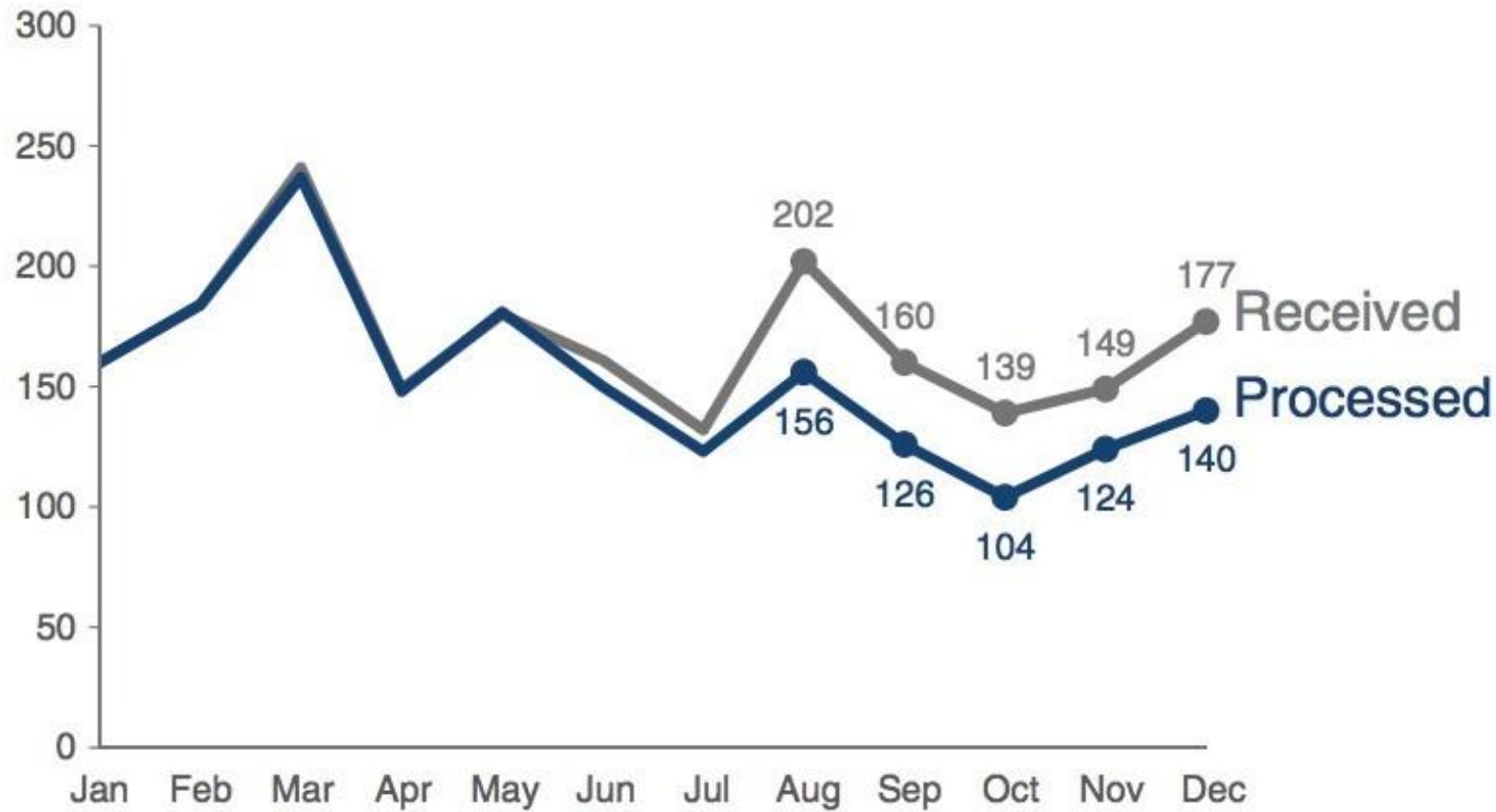
Make the data stand out to draw attention on the fact that needs to be emphasized

## Focus audience's attention using preattentive features



The added marks of data points and numeric labels are one useful preattentive attribute

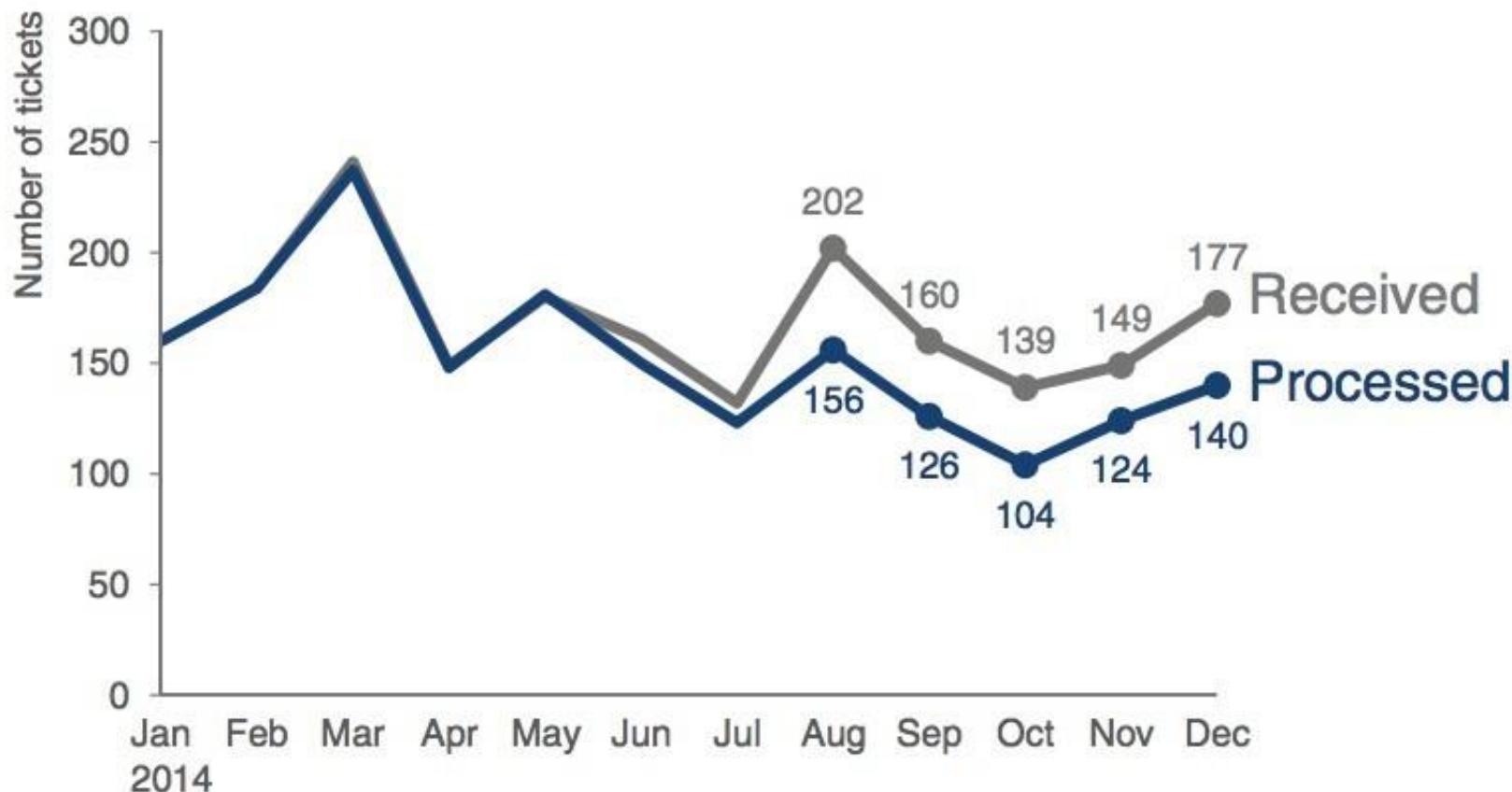
## Focus audience's attention using preattentive features



Add marks strategically to draw audience's attention more quickly to the right side of the graph

# Think like a designer

## Ticket volume over time



Data source: XYZ Dashboard, as of 12/31/2014

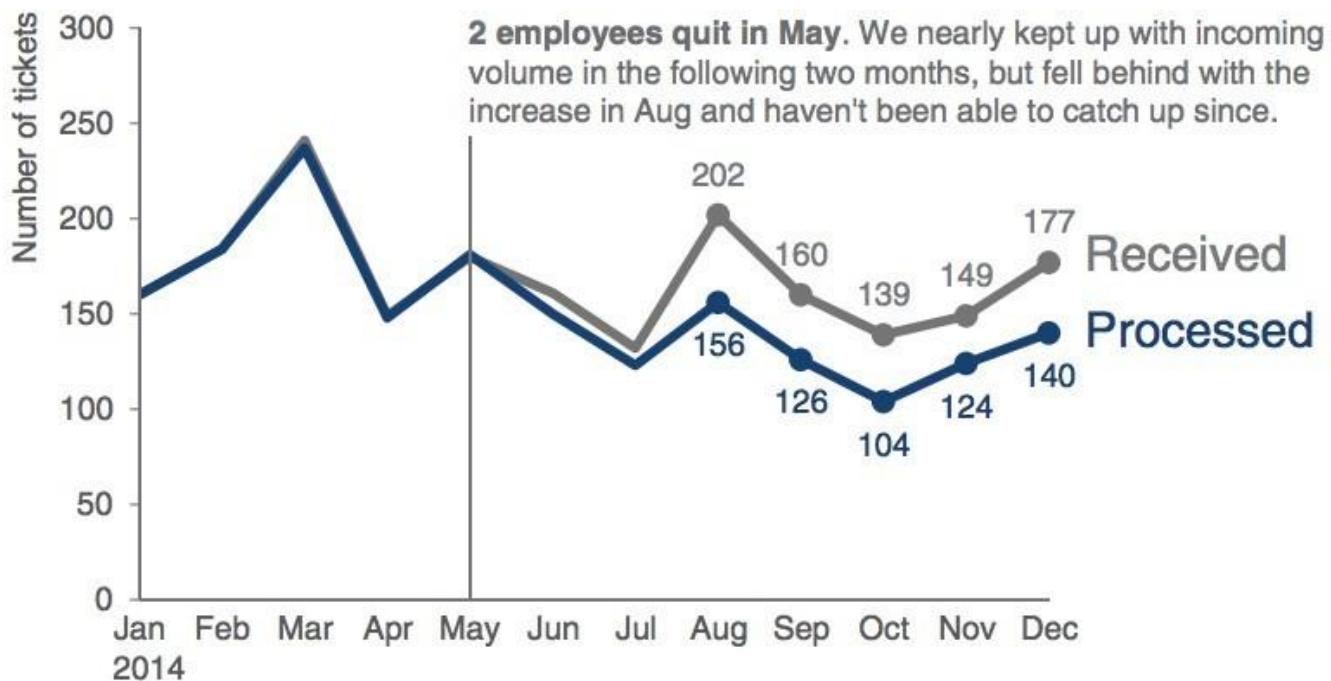
Slopegraph shows the visual percentage change and visual ordering of categories from the most to least

# Think like a designer

## Please approve the hire of 2 FTEs

to backfill those who quit in the past year

### Ticket volume over time



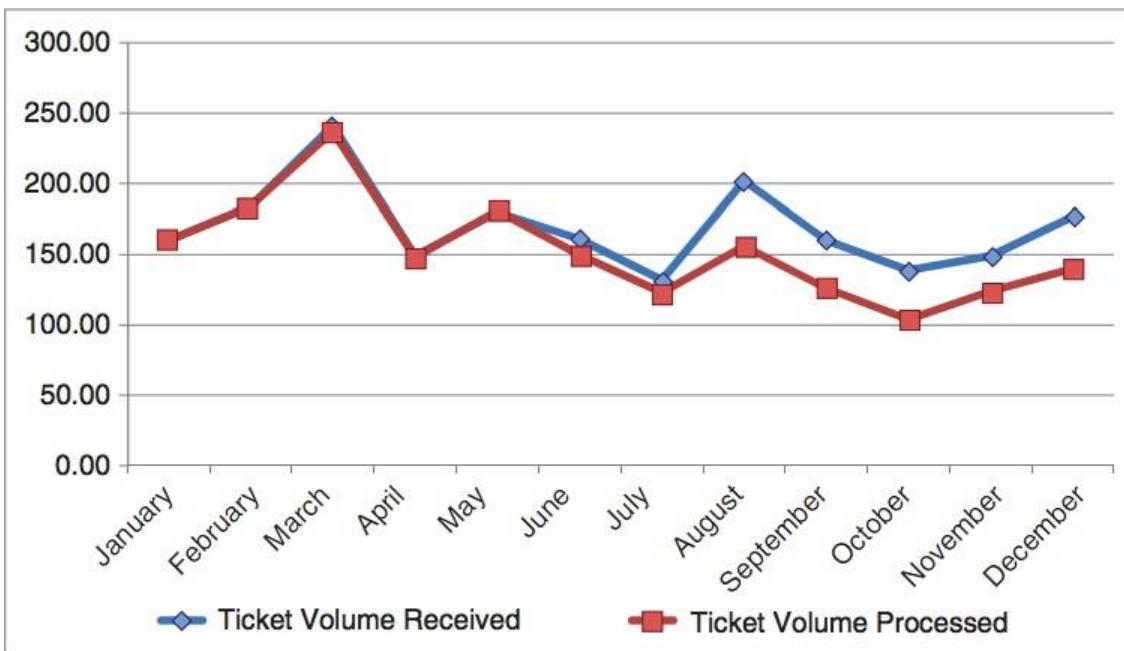
Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

Slopegraph shows the visual percentage change and visual ordering of categories from the most to least

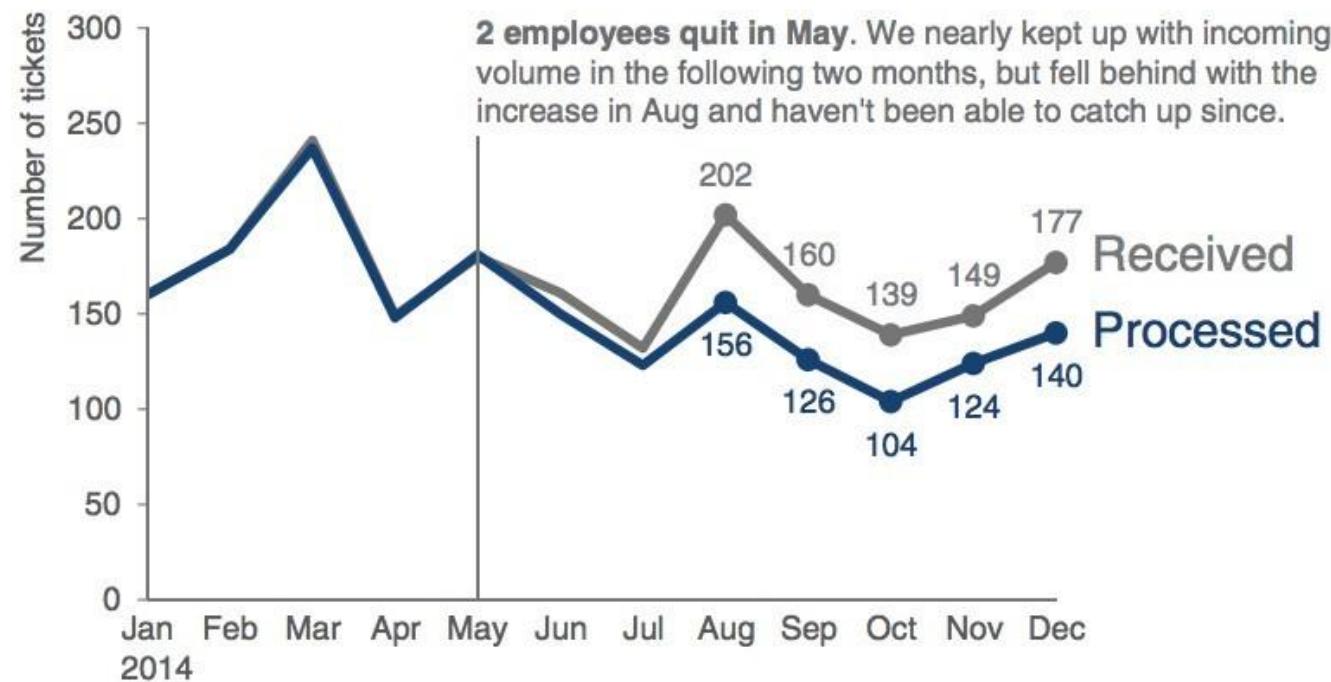
# Think like a designer

## Please approve the hire of 2 FTEs

to backfill those who quit in the past year



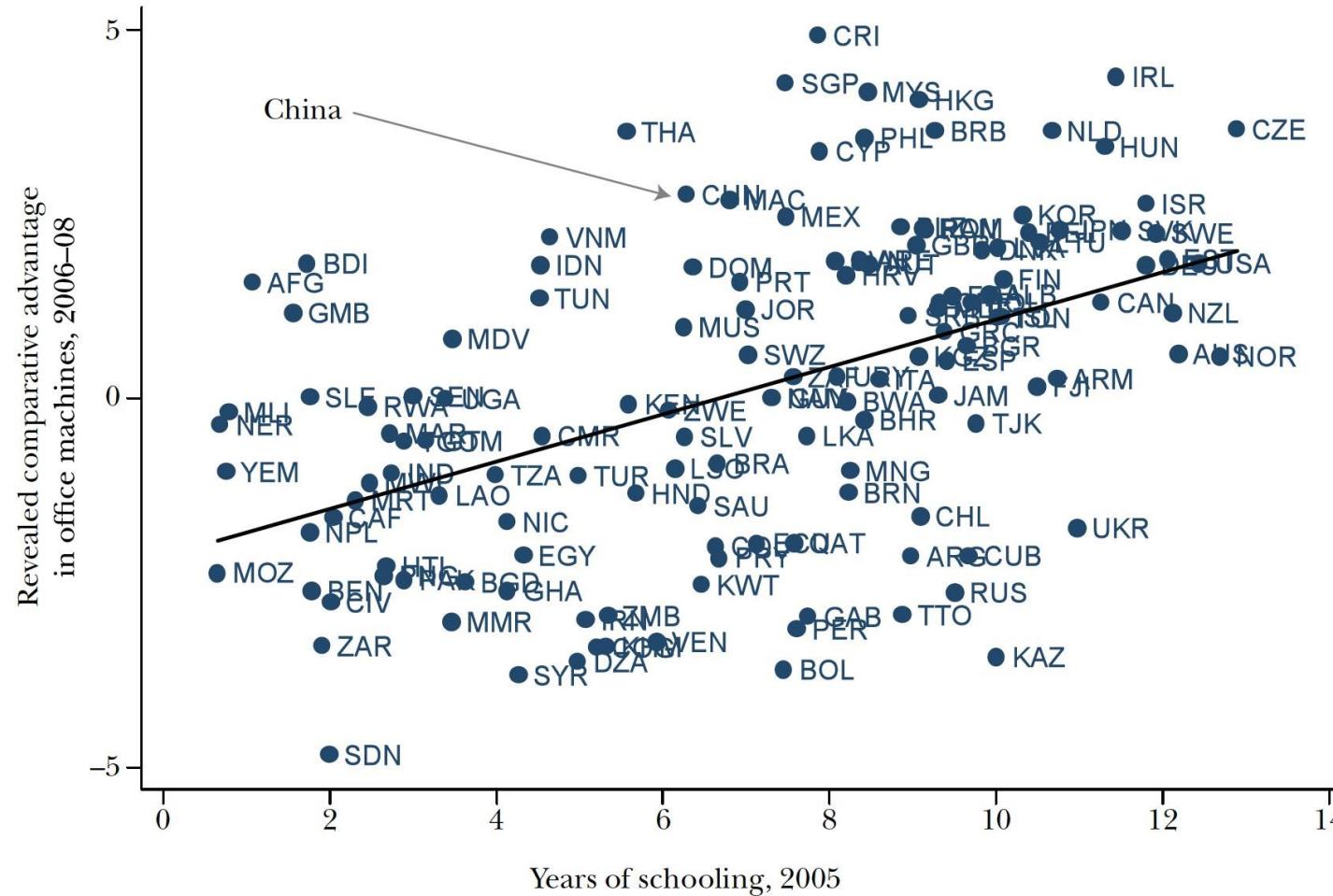
### Ticket volume over time



Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

# A scatter plot example

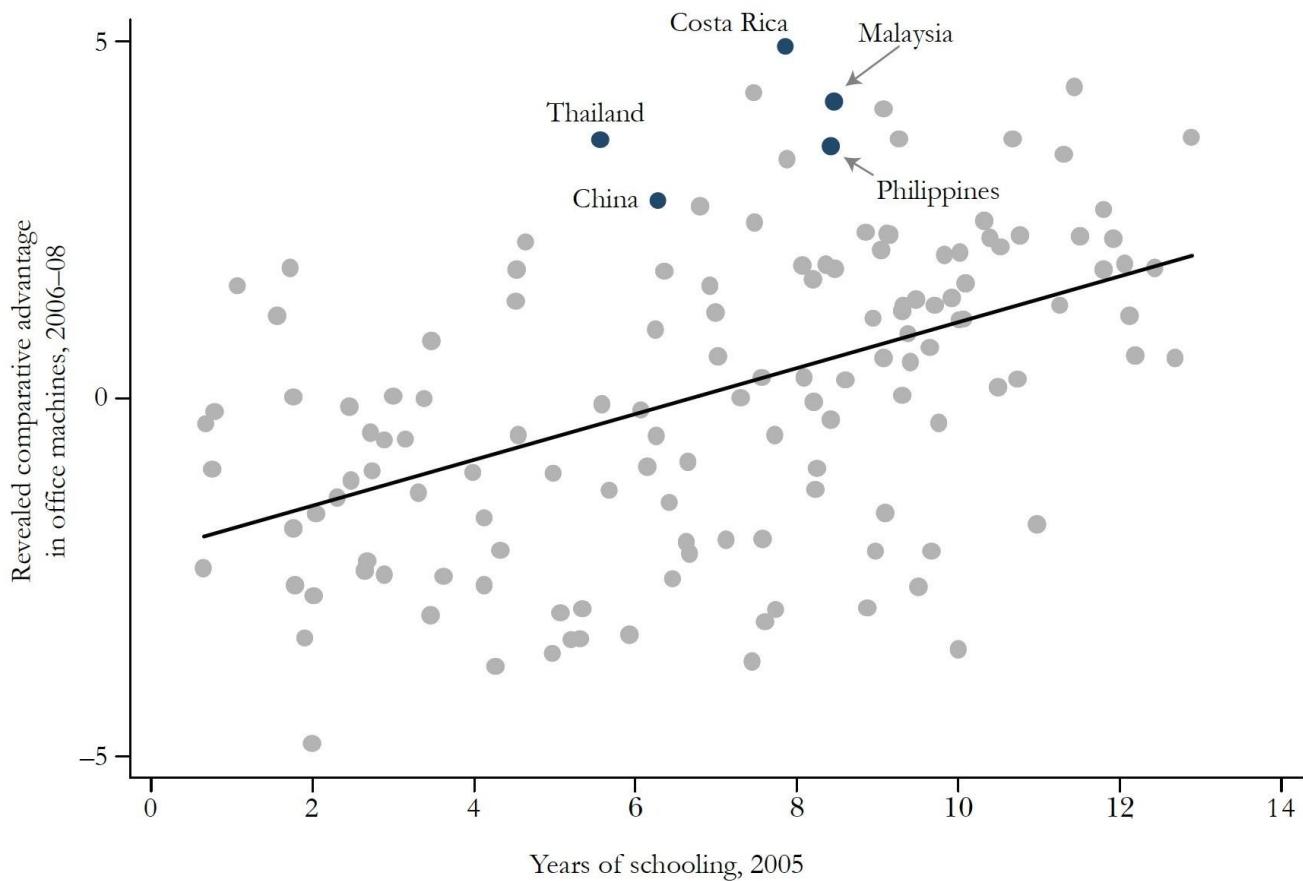
Education and Exports of Office Machines



- The figure plots countries' revealed comparative advantage in office machines . . . averaged over 2006 to 2008, against the average years of schooling of the adult population in 2005 . . .
- China is above the regression line, indicating that its specialization in the sector is greater than one would expect given its level of education, but it is hardly an extreme outlier.
- Other middle-income countries —including Costa Rica, the Philippines, Malaysia , and Thailand —have larger positive residuals

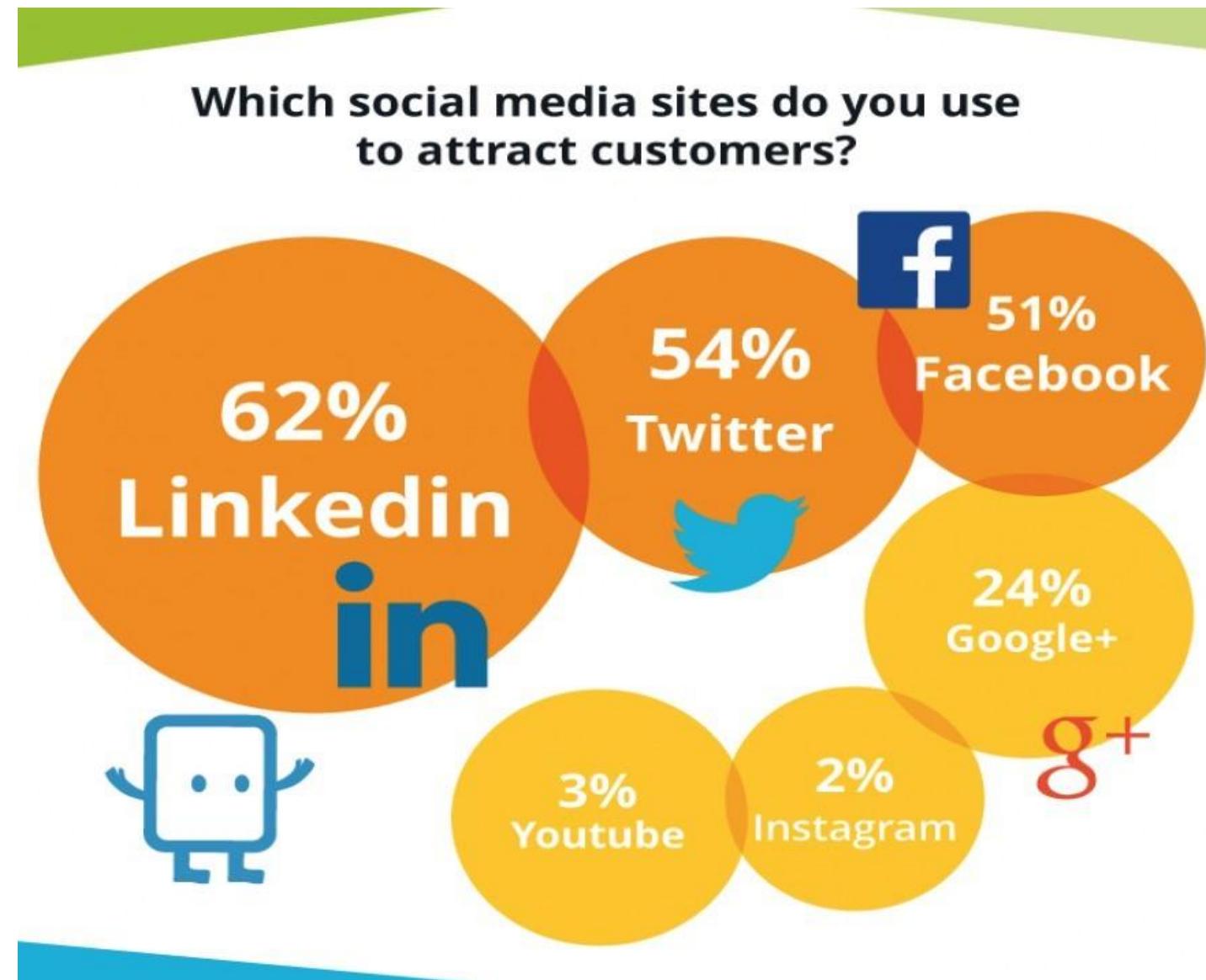
# A scatter plot example

Education and Exports of Office Machines



- eliminate all labels other than those for the five countries under discussion
- make five data points darker, thus deemphasizing the other points but still showing the important information

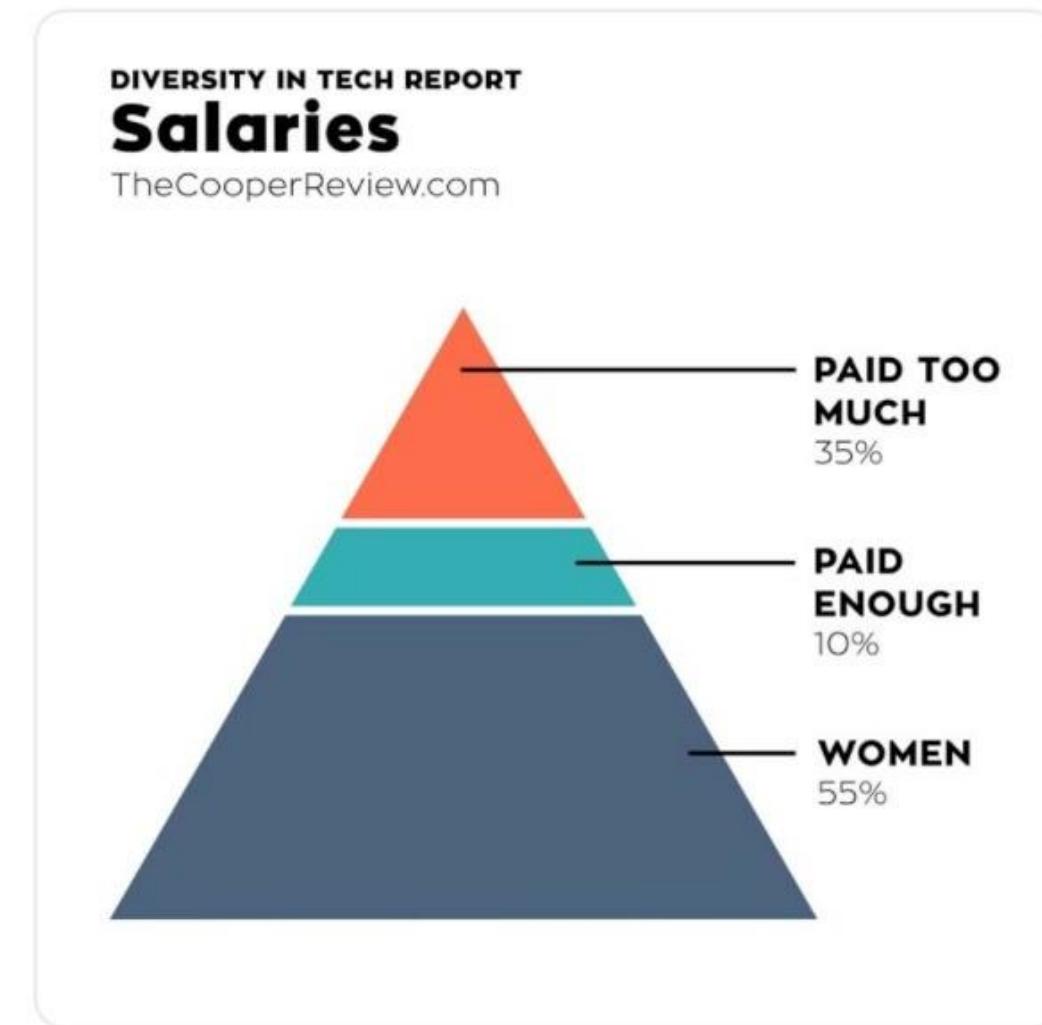
What is wrong in this plot?



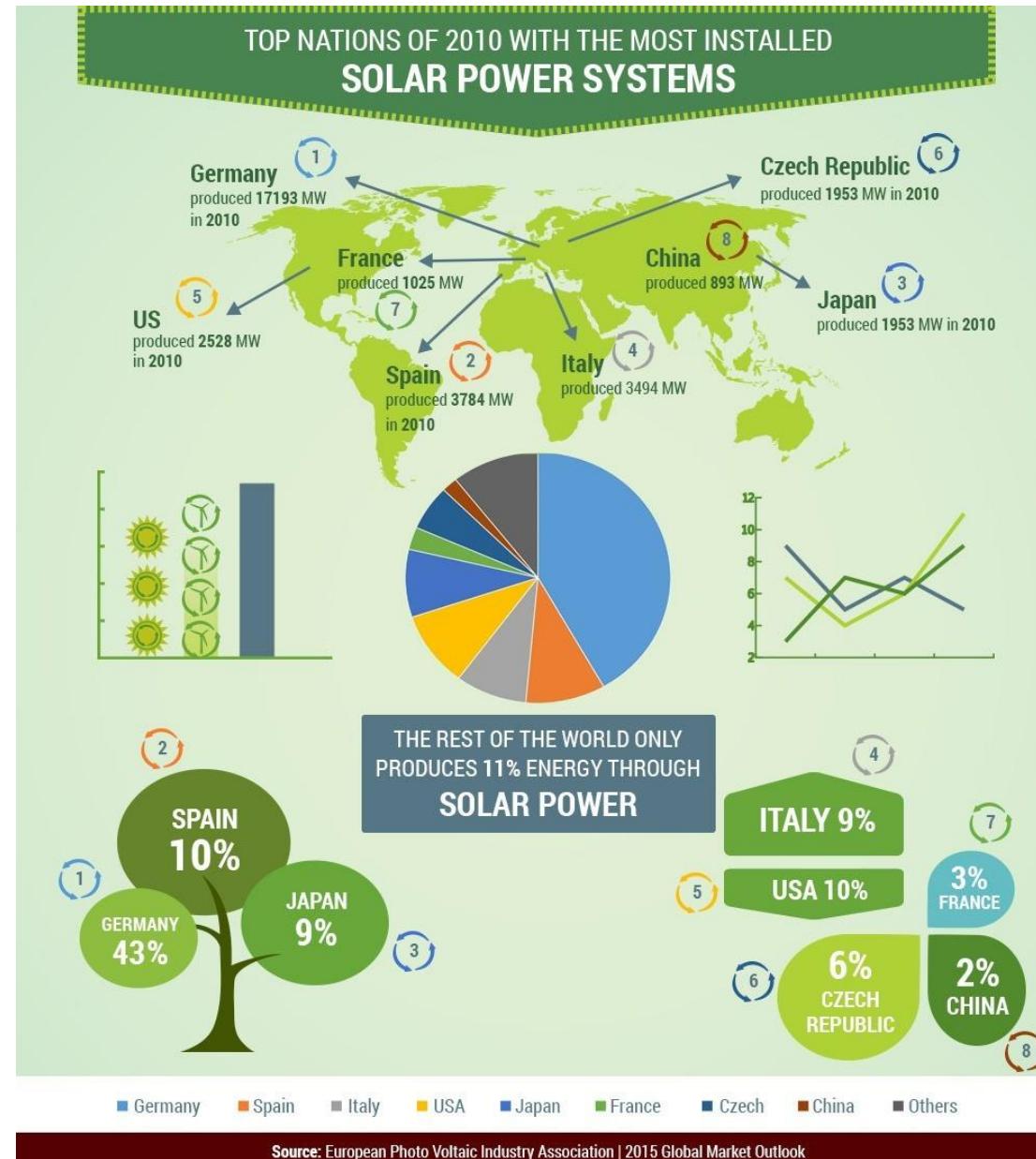
# What is the wrong in this plot?

## 7. Salaries

Salaries are also diverse.



# For those who are obsessed with Infographics...

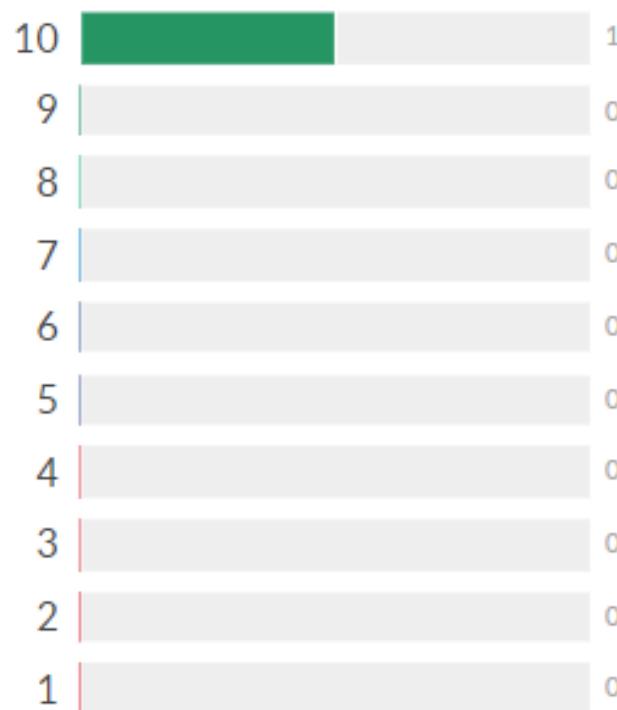


Do you need a visualization in this case?



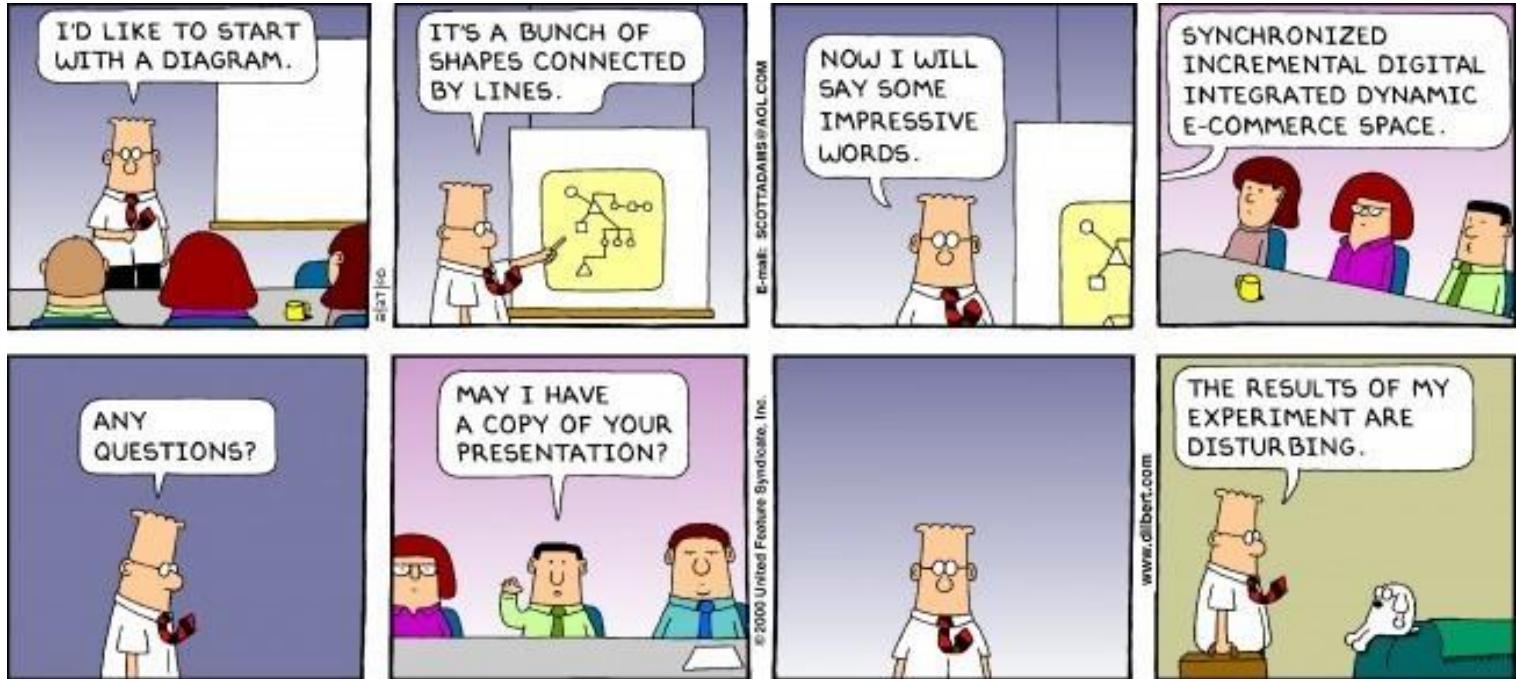
Do you need a visualization in this case?

#### RATINGS BREAKDOWN



#### RATINGS DETAIL

|                |        |
|----------------|--------|
| Avg. Rating    | 10.000 |
| No. of Ratings | 1      |
| Std. Deviation | 0.00   |



# Questions?