DATA SOCIETY®

”There is no such thing as information overload. There is only bad design."

- Edward Tufte, statistician, professor, & author

Who we are

Data Society's mission is to integrate Big Data and machine learning best practices across entire teams and empower professionals to identify new insights

- We provide:
  - High-quality data science training programs
  - Customized executive workshops
  - Custom software solutions and consulting services

- Since 2014, we've worked with thousands of professionals to make their data work for them

About the course

- Instructor introduction

- Schedule:
  - January 26-29 (4 days)
  - 11am – 2pm EST
  - 1 short break each class
  - Q&A at the end of each session

About the course

- Course objectives:
  - Create a narrative that accurately supports the data, provides context, and reveals actionable insights
  - Understand the design principles involved in creating effective and accurate visualizations
  - Recognize misleading or inaccurate charts and graphs provided by others

- Course materials:
  - Slides
  - Participant guide
    - Needed during class
    - Contains class activities, data visualization checklists, and more!

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Best practices for virtual classes

- Find a quiet place, free of as many distractions as possible. Headphones are recommended.
- Remove or silence alerts from cell phones, e-mail pop-ups, etc.
- Participate in activities and ask questions!
- Ask questions through Q and A tab
- Switch between "All panelists" (to communicate directly with instructor and support) and "Panelists and participants" (to interact with all participants) tabs in the chat window
- Give your honest feedback so we can troubleshoot problems and improve the course.

Activity: class survey

- Please complete the class survey found at: https://www.surveymonkey.com/r/MCJNZQ6

- A clickable link can be found on page 4 of your participant guide under **class survey – part I** .

Polling question

Where are you joining us from today?

- Home
- Office
- Somewhere else

Polling question

What is the primary reason you are taking this course?

- I tell stories with data as part of my job.
- I want to be a better consumer of data.
- I want to obtain marketable skills.
- Other

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

What is data visualization?

- Data visualization is any attempt to make data more easily digestible by rendering it in a visual context.

- Common data visualizations include tables, charts, graphs, and dashboards.

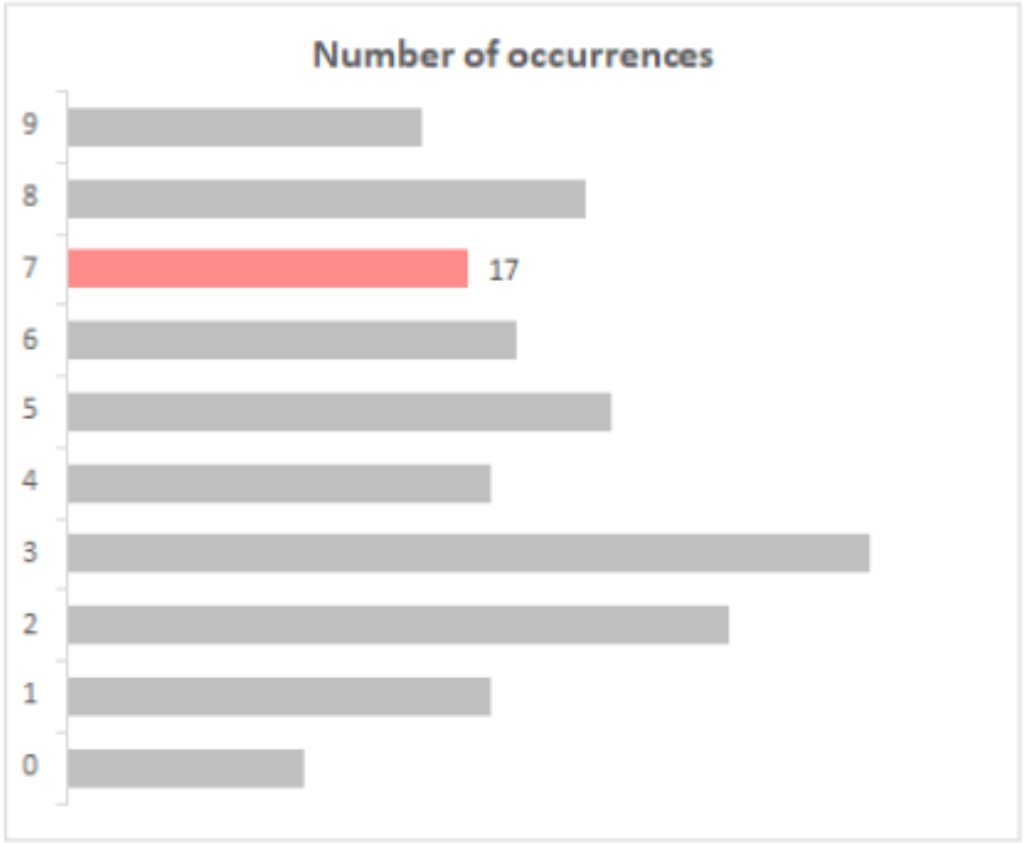
Data viz example

Count how many times the number 7 appears in the grid below.

5	2	8	3	6	1	9	3	6	2	5	3	7	4	3	8	3
8	5	8	9	6	2	1	4	4	3	9	3	6	5	2	4	9
1	0	2	7	5	2	8	3	6	1	6	2	9	3	8	3	8
5	8	4	7	2	0	3	7	3	5	4	7	1	8	2	0	1
2	5	3	6	4	3	9	1	0	8	9	5	7	3	4	5	3
2	7	5	2	8	3	6	1	6	2	9	3	8	3	8	5	8
4	7	2	0	3	7	3	5	4	7	1	8	2	0	1	9	6
2	1	4	4	3	9	3	6	5	2	4	9	1	0	2	7	5
2	8	3	6	1	6	2	9	3	8	3	8	5	8	4	7	2
0	3	7	3	5	4	7	1	8	2	0	1	2	5	3	6	4
3	9	1	8	9	5	0	7	3	4	5	3	2	7	5	2	8
3	6	1	6	2	4	6	2	7	5	9	1	5	2	6	3	6

Data viz example

Is it easier to tell now?



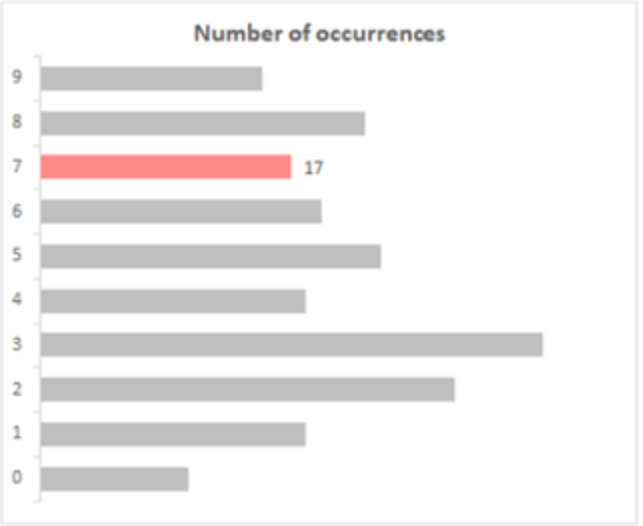
5	2	8	3	6	1	9	3	6	2	5	3	7	4	3	8	3
8	5	8	9	6	2	1	4	4	3	9	3	6	5	2	4	9
1	0	2	7	5	2	8	3	6	1	6	2	9	3	8	3	8
5	8	4	7	2	0	3	7	3	5	4	7	1	8	2	0	1
2	5	3	6	4	3	9	1	0	8	9	5	7	3	4	5	3
2	7	5	2	8	3	6	1	6	2	9	3	8	3	8	5	8
4	7	2	0	3	7	3	5	4	7	1	8	2	0	1	9	6
2	1	4	4	3	9	3	6	5	2	4	9	1	0	2	7	5
2	8	3	6	1	6	2	9	3	8	3	8	5	8	4	7	2
0	3	7	3	5	4	7	1	8	2	0	1	2	5	3	6	4
3	9	1	8	9	5	0	7	3	4	5	3	2	7	5	2	8
3	6	1	6	2	4	6	2	7	5	9	1	5	2	6	3	6

Exploratory data analysis

- Exploratory data analysis (EDA) is the process of reviewing new data to discover patterns, to spot anomalies, to test hypotheses, and to check assumptions.

- Just like in our counting example, EDA is made easier by using summary statistics and data visualizations.

- Often, summary statistics alone aren't enough.



EDA example

MONTH	SALES REP 1	SALES REP 2	SALES REP 3
1-Jan	101	34	171
2-Feb	199	111	354
3-Mar	300	633	460
4-Apr	405	313	370
5-May	495	409	145

What insights can you draw from this data?

EDA example

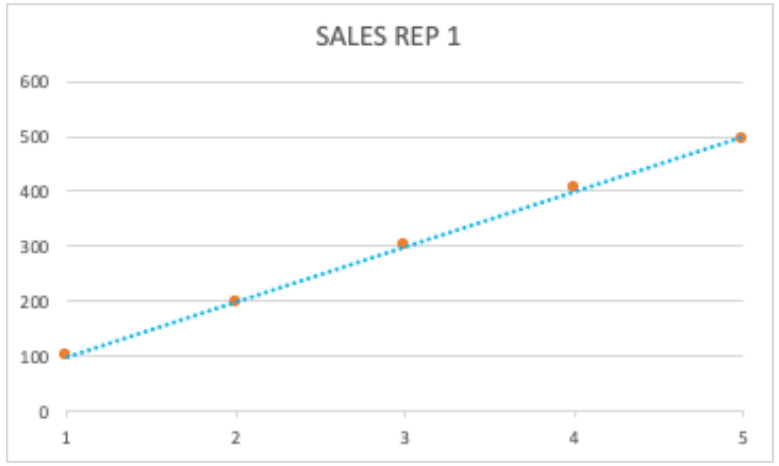
MONTH	SALES REP 1	SALES REP 2	SALES REP 3
1-Jan	101	34	171
2-Feb	199	111	354
3-Mar	300	633	460
4-Apr	405	313	370
5-May	495	409	145
SUM	1,500	1,500	1,500
AVERAGE	300	300	300

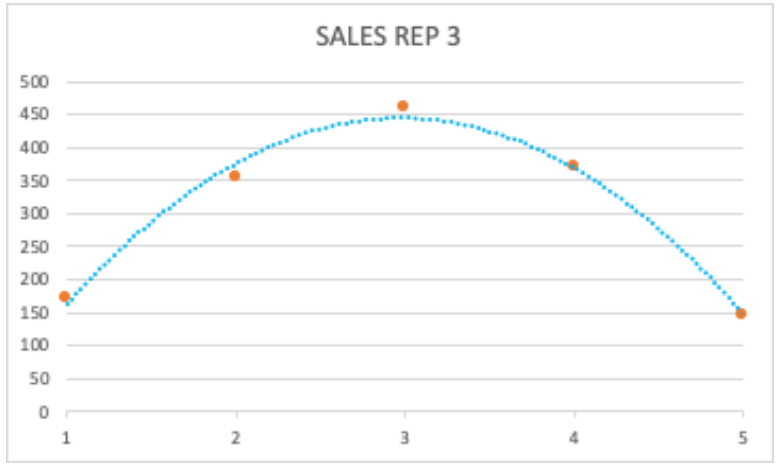
What insights can you draw from this data?

EDA example

MONTH	SALES REP 1	SALES REP 2	SALES REP 3
1-Jan	101	34	171
2-Feb	199	111	354
3-Mar	300	633	460
4-Apr	405	313	370
5-May	495	409	145
SUM	1,500	1,500	1,500
AVERAGE	300	300	300

What insights can you draw from this data?

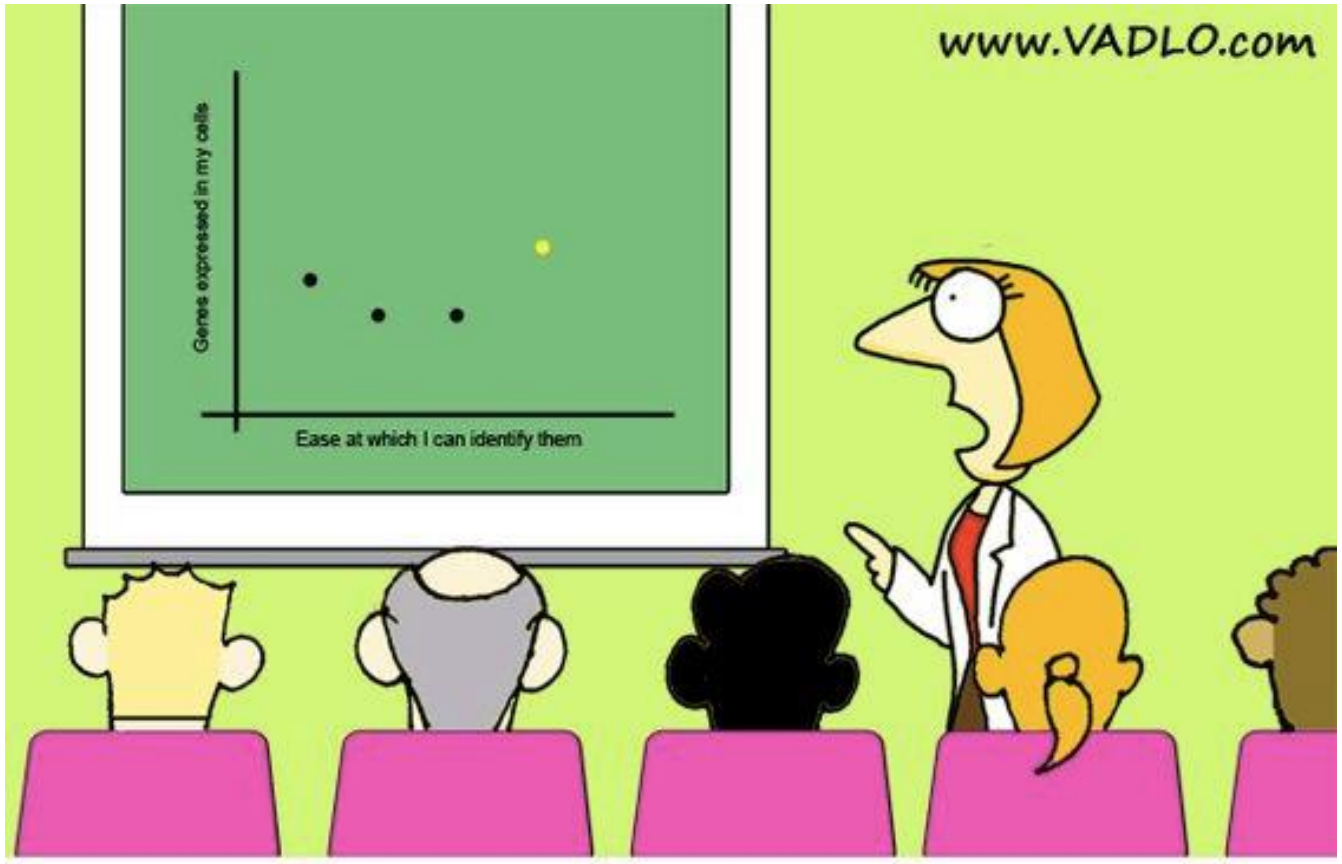




Explanatory data

- We also use data visualization to transform raw data into something compelling for an external audience.

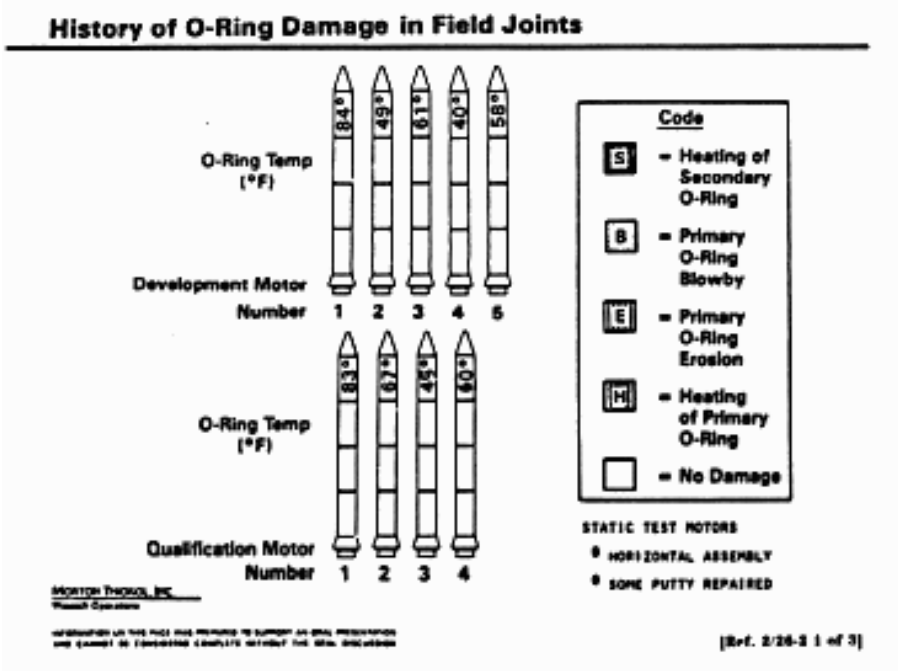
- Using visualizations incorrectly can cause you to lose your audience, lose the value in your data, and ultimately lead to poor decision making.

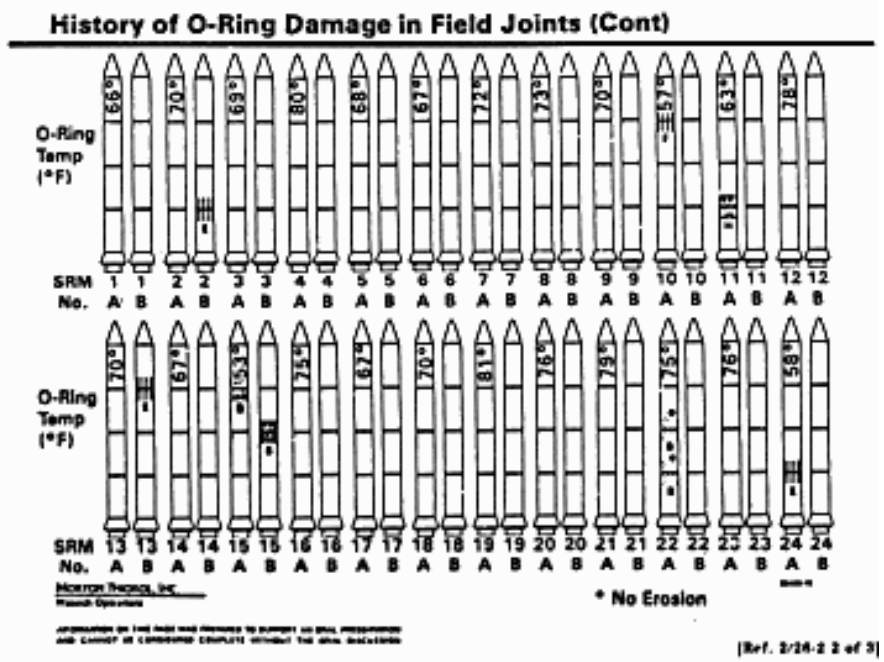


"Same graph as last year,
but now I have an additional dot."

Example: The Challenger

- On January 27, 1986, concerned engineers presented data and the following charts to try to illustrate the damage cold temperatures would have on the O-rings of the Challenger space shuttle.





Source: Presidential Commission on the Space Shuttle Challenger Accident, vol. 5 (Washington, DC: US Government Printing Office, 1986.) pp. 895-896.

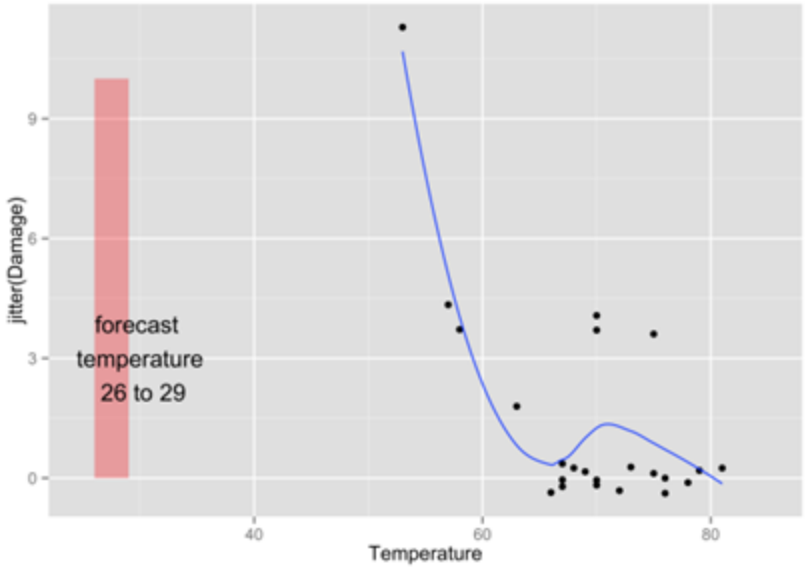
Example: The Challenger

- January 28, 1986, the Challenger space shuttle exploded within seconds of takeoff.

- Data visualization legend Edward Tufte argues that the shuttle's engineers failed to communicate dangers because their data wasn't presented in an easily digestible form.

The chart below shows O-ring damage on the y-axis and temperature on the x-axis.

Is it easier to see the issue?



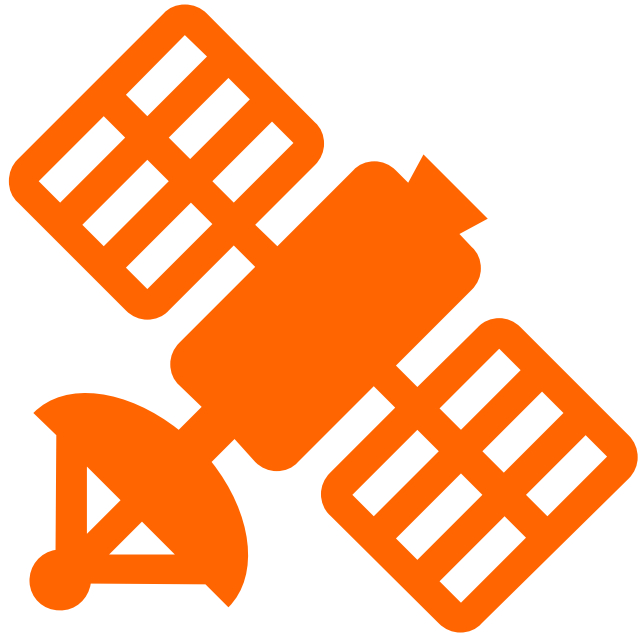
https://vizdatar.wordpress.com/2015/05/06/space-shuttle-challenger-explosion-2/

Example: Space junk

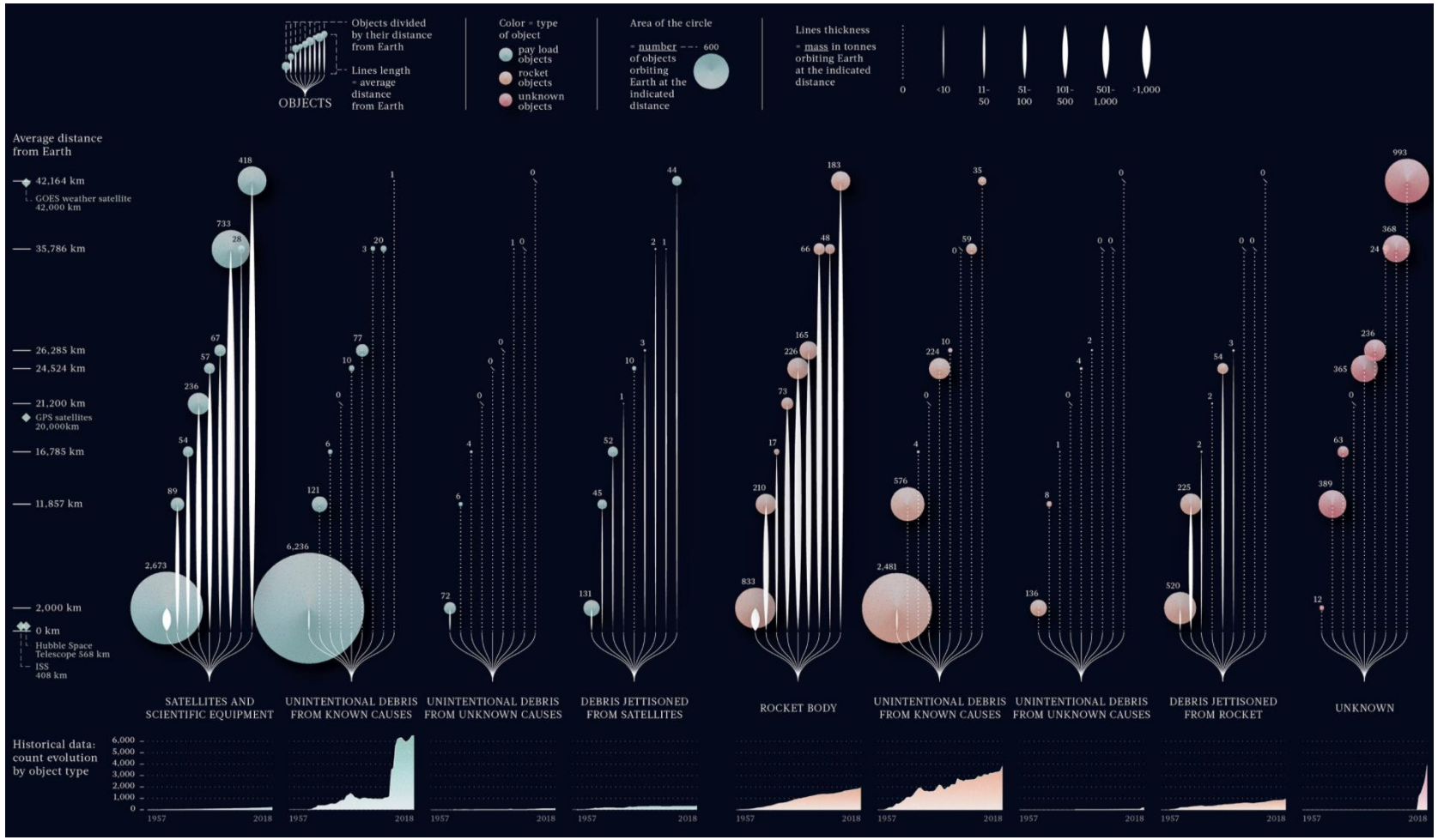
- “Space junk has been increasing over the last few decades, and collisions could increase if the problem is not kept in check.”

- What exactly is the scale of the space junk problem?

- What to represent:
  - number of objects
  - mass of objects
  - types of object
  - distance from Earth



Example: Space junk



Source: https://www.behance.net/gallery/81688575/Space-Junk-BBC-Science-Focus / Data set: https://sdup.esoc.esa.int/

Recap

- Data visualization is any attempt to make data more easily digestible by rendering it in a visual context.

- We can use data visualization to:
  - perform exploratory data analysis
  - explain data to others

Break

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Getting started with data viz

- We first need to understand the types of data that exist.
- Then, we'll move on to what you need to consider before visualizing.

What is data?

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available

— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published

— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Merriam Webster

Types of data

Qualitative

- Data that approximates or characterizes
- Often obtained through interviews, surveys, focus groups, documents, etc.
- Examples include:
  - name
  - sex (male, female, nonbinary)
  - observations (e.g., the food was salty)

Quantitative

- Data that is measurable and expressed as a number
- Often gathered using metrics, tests, experiments, instruments with a ratings scale, etc.
- Examples include:
  - cost
  - number of applicants
  - weight

Types of data

Discrete

- Data that can only take certain values

- Examples include:
  - the cost of a BMW 7-series
  - your shoe size
  - number of people enrolled in this course

Continuous

- Data that can take any value and usually changes over time

- Examples include:
  - the speed of a car during the morning commute
  - your weight
  - the time your cat wakes up each day

Activity: data types

- Turn to page 5 of your participant guide to find the **data types** activity.
- Read the paragraph and answer the questions that follow.

Executive Summary

Almost all commodity prices saw sharp declines during the past three months as the COVID-19 pandemic worsened. Mitigation measures have significantly reduced transport, causing an unprecedented decline in demand for oil, while weaker economic growth will further reduce overall commodity demand. Crude oil prices are expected to average $35/bbl this year and $42/bbl in 2021—a sharp downward revision from October in both years. Non-energy prices are also expected to fall this year. Metals are projected to decline more than 13 percent in 2020, before recovering in 2021 while food prices are expected to be broadly stable. The risks to the price forecasts are large in both directions and depend on the speed at which the pandemic is contained and mitigation measures are lifted. A Special Focus investigates the impact of COVID-19 on commodity markets and compares it with previous disruption episodes. It finds that the impact of COVID-19 has already been larger than most previous events and may lead to long-term shifts in global commodity demand and supply. A Box examines the impact of international commodity production agreements, with a particular focus on OPEC, and concludes that OPEC+, the last remaining international agreement to manage supply, is subject to the same forces that led to the collapse of its predecessors.

Activity: data types

1. What examples of quantitative data are included in the summary that you read?

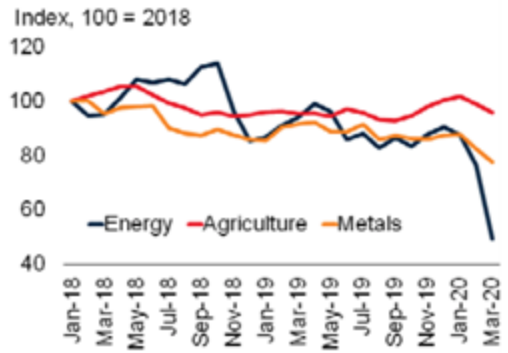
2. Do you see any examples of qualitative data in this summary? If so, what are they?

3. Do you see examples of continuous data are mentioned in this summary? If so, what are they?

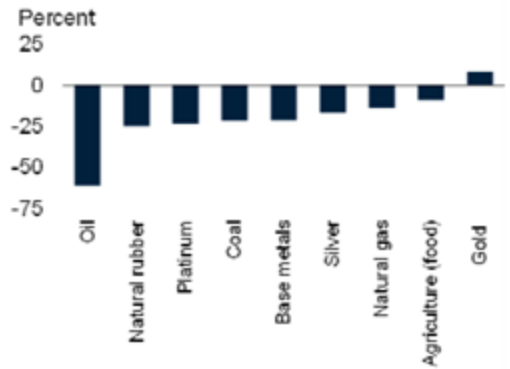
4. What types of visualizations might make this summary easier to digest?

Activity: data types

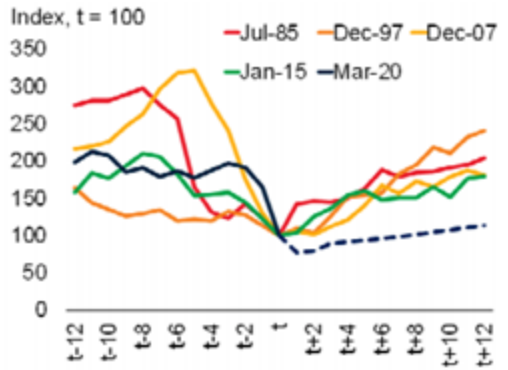
A. Commodity price indexes, monthly



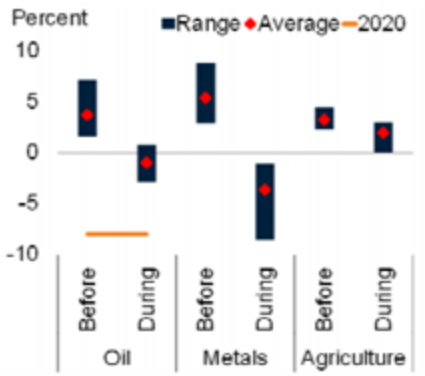
B. Commodity price changes since January 20th



C. Oil prices during collapse and recovery episodes and 2020 forecast



D. Commodity demand growth around global recessions



Source: Bloomberg; BP Statistical Review; IEA; USDA; World Bank; World Bureau of Metal Statistics

http://pubdocs.worldbank.org/en/900511587395260657/CMO-April-2020-Executive-Summary.pdf

To get started with data viz

1. Know your audience and understand how it processes visual information. (Who)

1. Determine what you're trying to visualize and what kind of information you want to communicate. (What)

1. Choose a type of visual that conveys the information in the best and simplest form for your audience. (How)

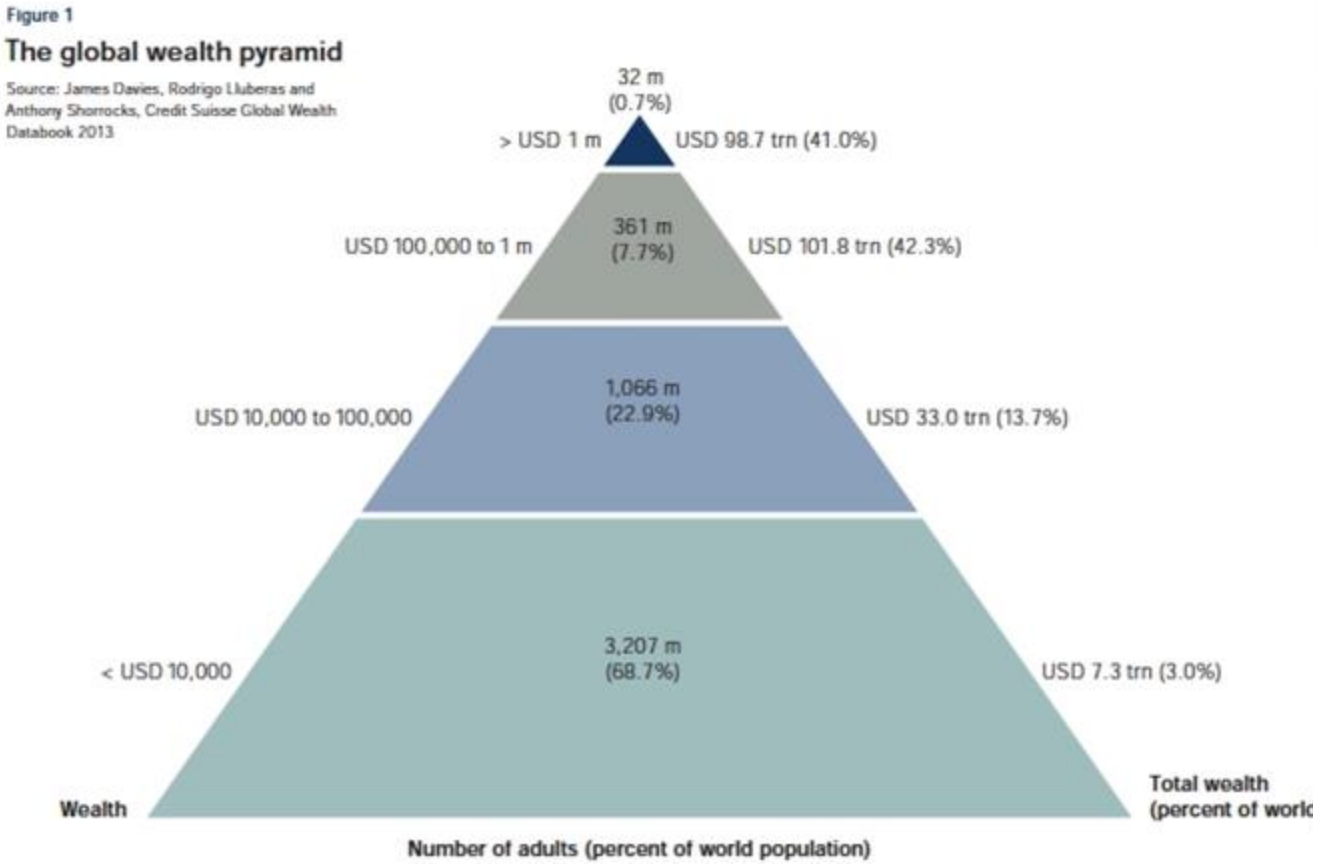
Who

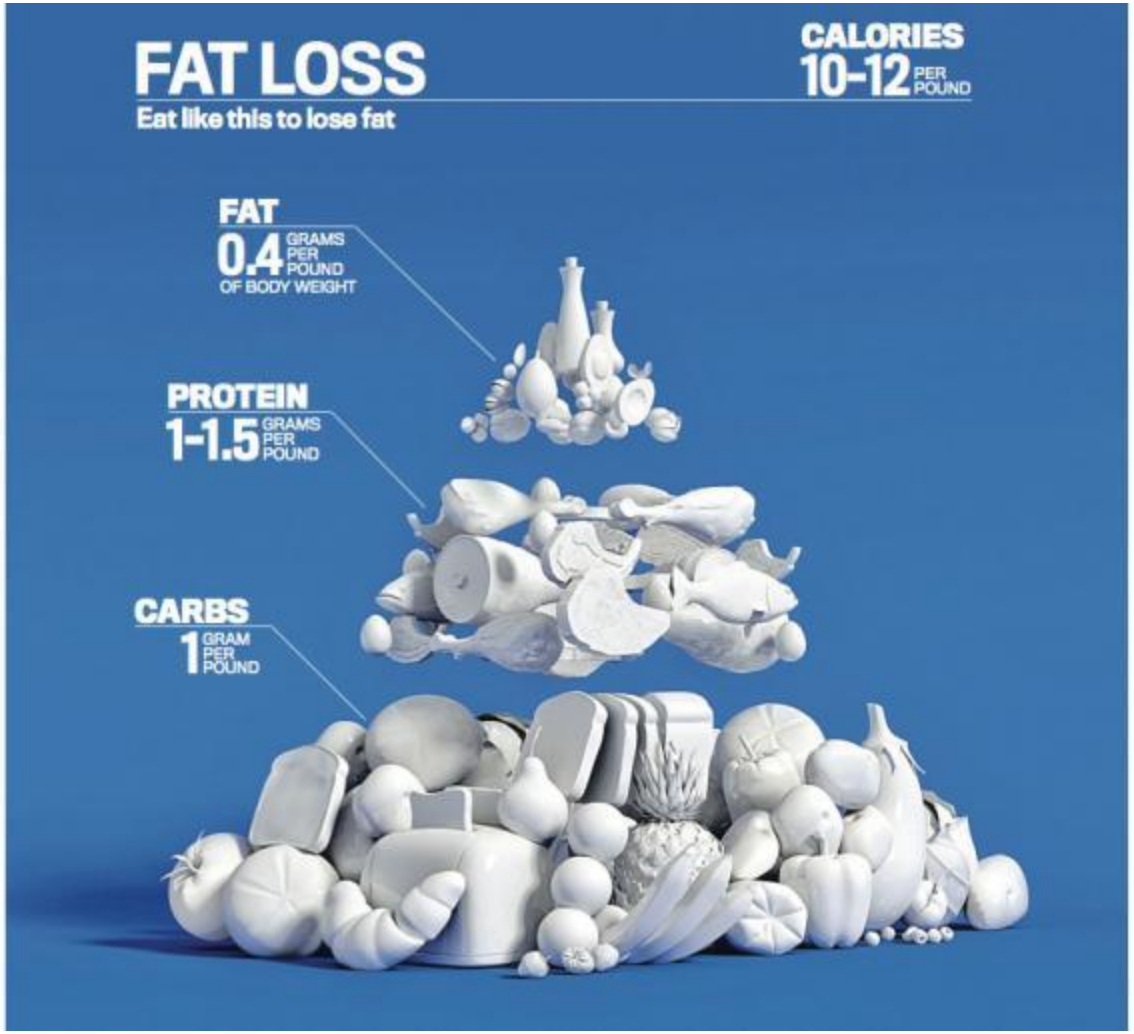
- Know your audience and understand how it processes visual information.

- Consider audience familiarity. For example:
  - High-level executives are generally well-versed in visual data, so use a variety of methods to stand out
  - Less-experienced audiences will want it kept simple (e.g., pie charts, bar graphs, and word maps)

- Consider how the visualization will be used by the audience:
  - *Is it for executives to use to make decisions?*
  - *Is it to inform the public?*

How might the audience differ?





https://www.oreilly.com/library/view/learning-highcharts-4/9781783287451/ch08s04.html

http://www.mensfitness.com/nutrition/what-to-eat/mens-fitness-food-pyramid

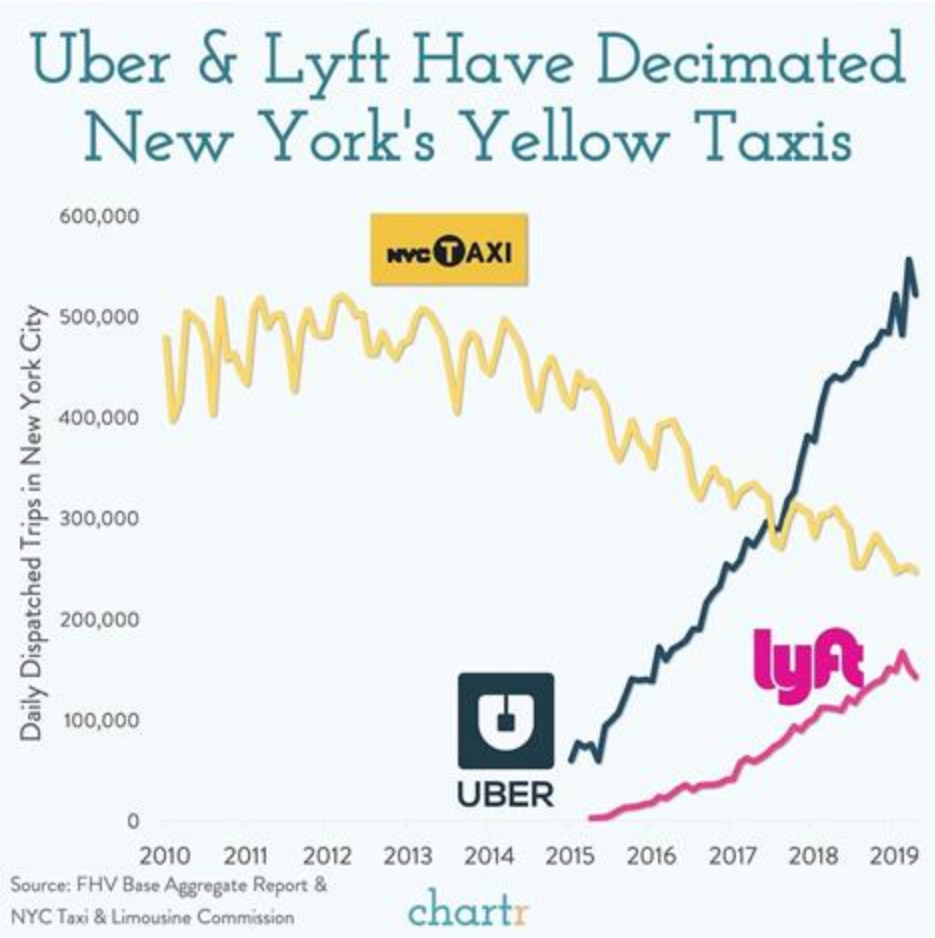
What

- Determine what you're trying to visualize and what kind of information you want to communicate.

- Remember, the audience only knows as much as you tell them:
  - *Do you want them to explore the data on their own? (exploratory analysis)*
  - *Do you want to tell a specific story about the data? (explanatory purposes)*

- If the message is explanatory, consider:
  - *What type of data you have on which to base the analysis?*
  - *What are the audience's topmost concerns or requirements?*
  - *What decisions can be made based on the results you provide?*

Polling question



Is the message of this data visualization clear?

- Yes
- No
- I'm not sure

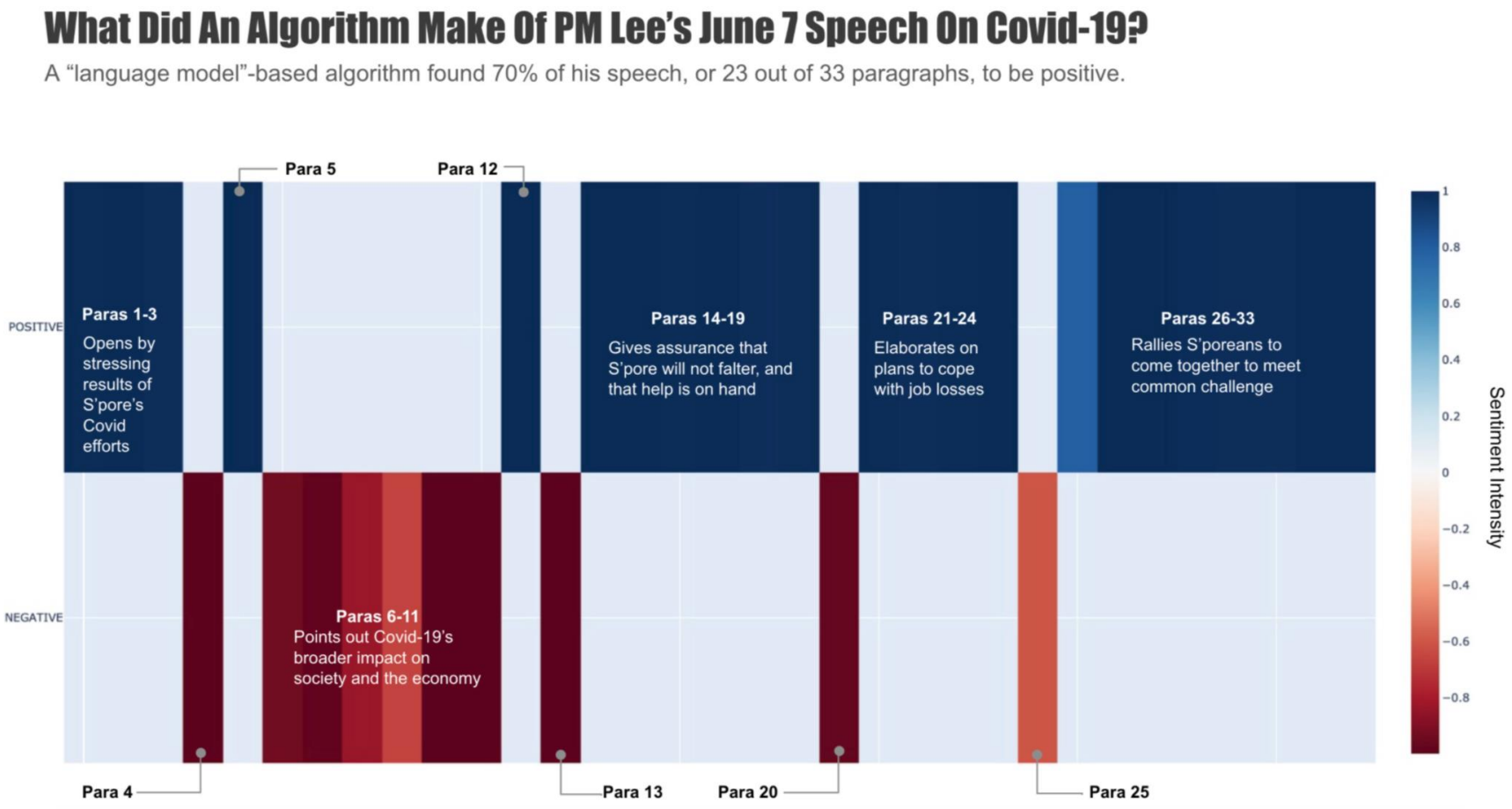
How

- Choose a type of visual that conveys the information in the best and simplest form for your audience.

- The type of visual you use depends primarily on two things:
  1. the data you want to communicate
  2. what you want to convey about that data

- Then, choose the visual that will be easiest for your audience to read.
  - Aim for them to "get it" in 30 seconds or less.

Poll question



Did the algorithm determine this speech to be largely positive or largely negative?

Source: https://towardsdatascience.com/sentiment-analysis-of-political-speeches-using-hugging-faces-pipeline-feature-3109c121d351

Poll question

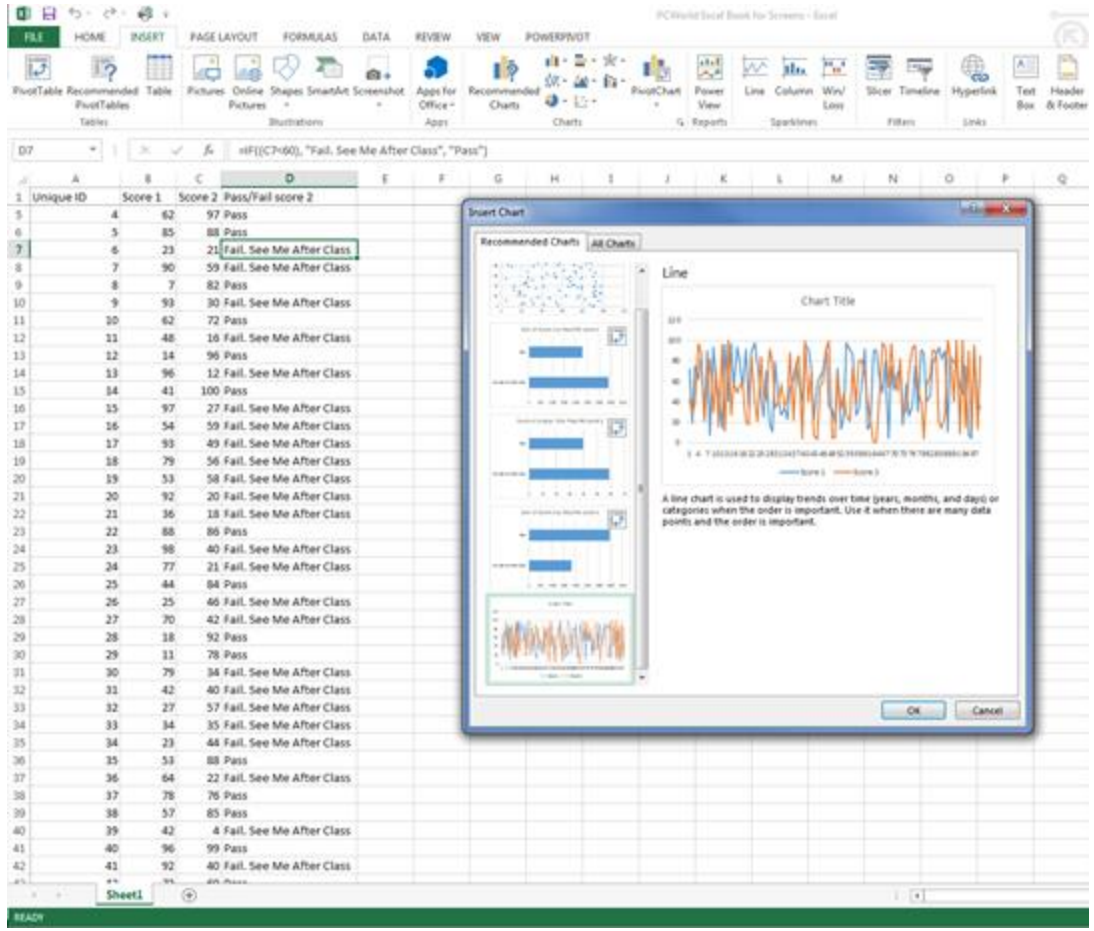
Which data visualization tools do you use most frequently?

- Microsoft Excel
- Google Charts
- Tableau
- R and R Studio
- Python
- Power BI

Microsoft Excel

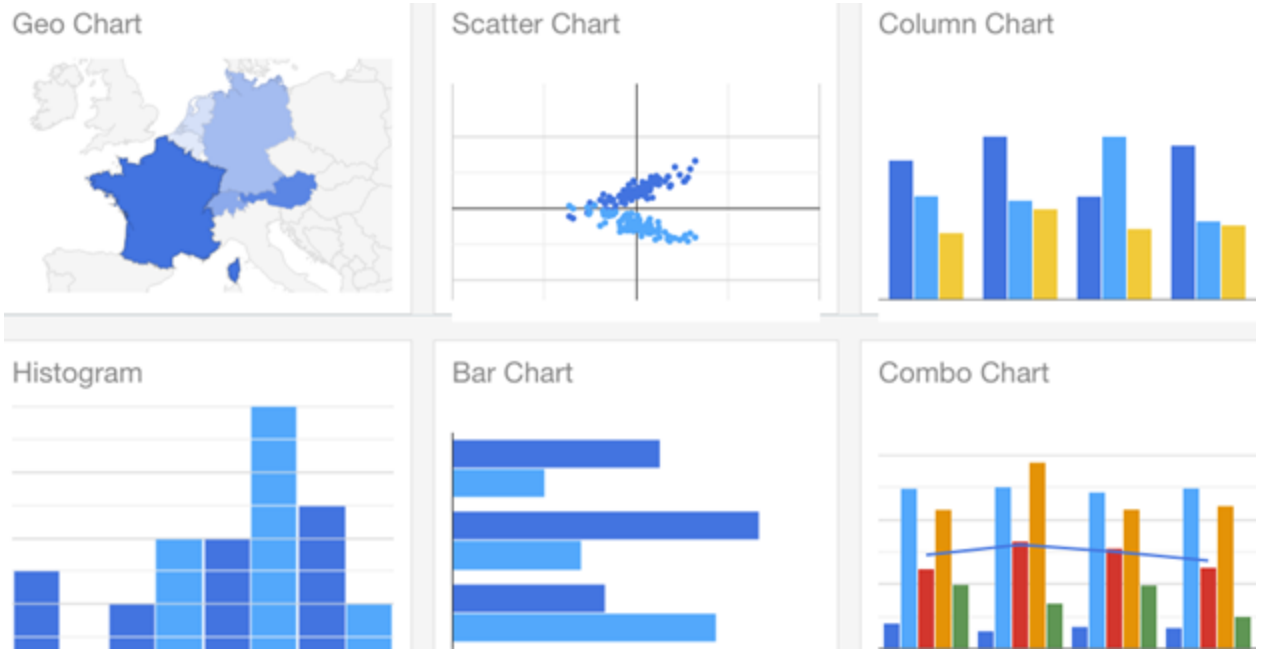
- Microsoft Excel allows you to create basic chart types such as pie, line, bar, scatter, and more.

- Charts created in Excel can easily be ported to PowerPoint and Word.



Google Charts

- Google Charts is free and open source.
- Its chart gallery has many ready-to-use chart types, which can be customized and organized in dashboards.
- It allows you to create interactive, animated, and geospatial graphics.
- Charts are easy to export for web use.

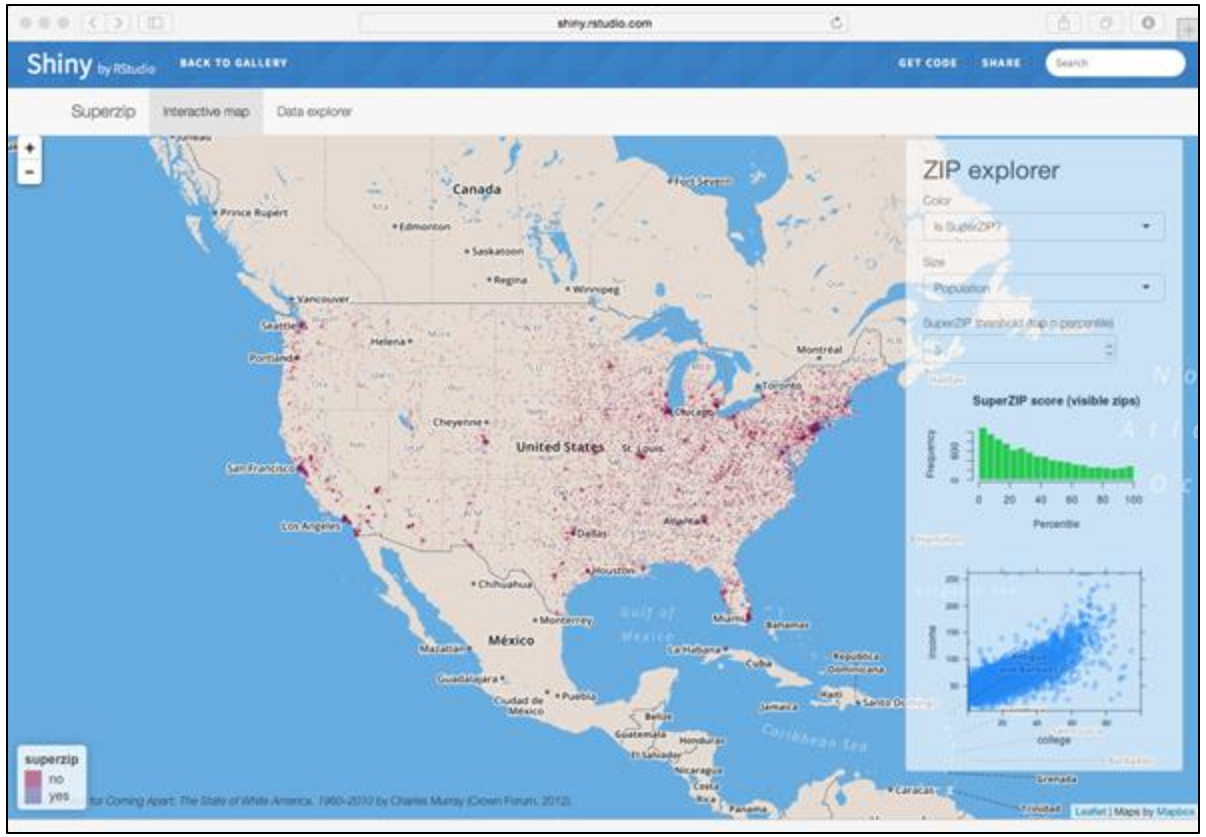


Tableau

- Tableau is a tool for creating powerful and insightful visuals.
- No programming is required; you can drag and drop.
- It allows you to share and collaborate in the cloud for use department- or organization-wide.

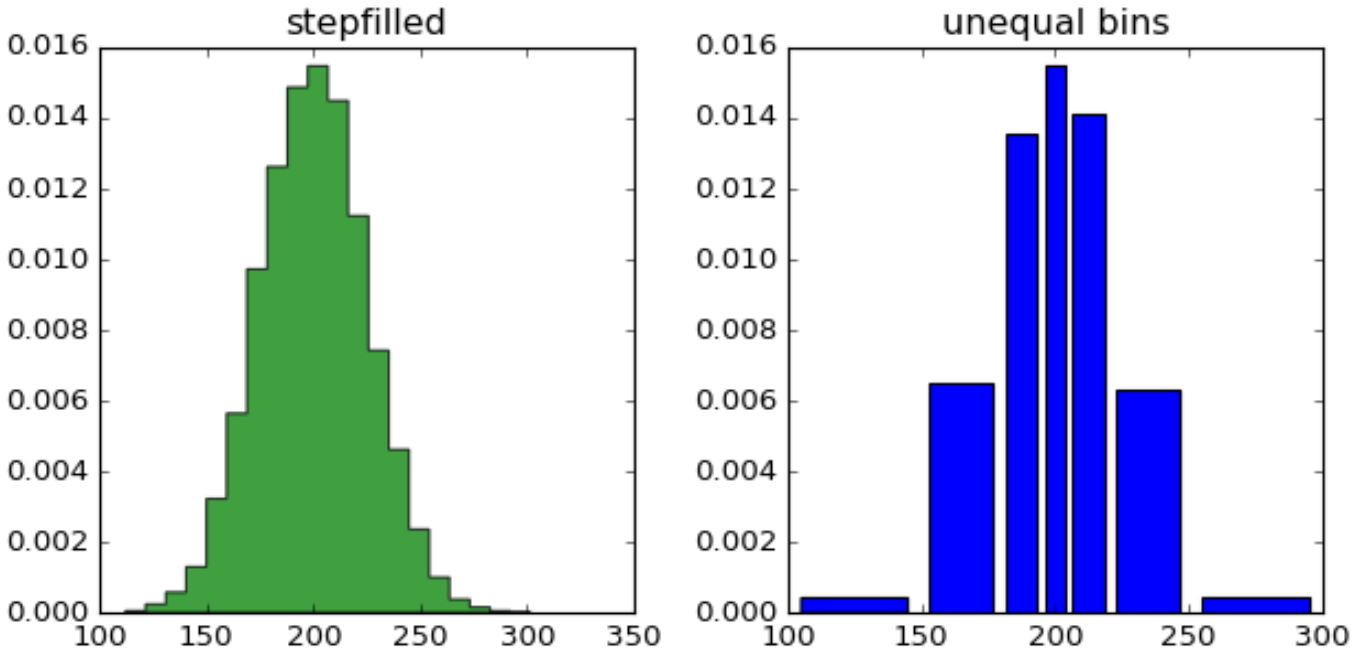
R and RStudio

- RStudio is a programming tool that is mainly used for statistical analysis.
- It offers functions and libraries to build visualizations and present data.
- It is open source and free.



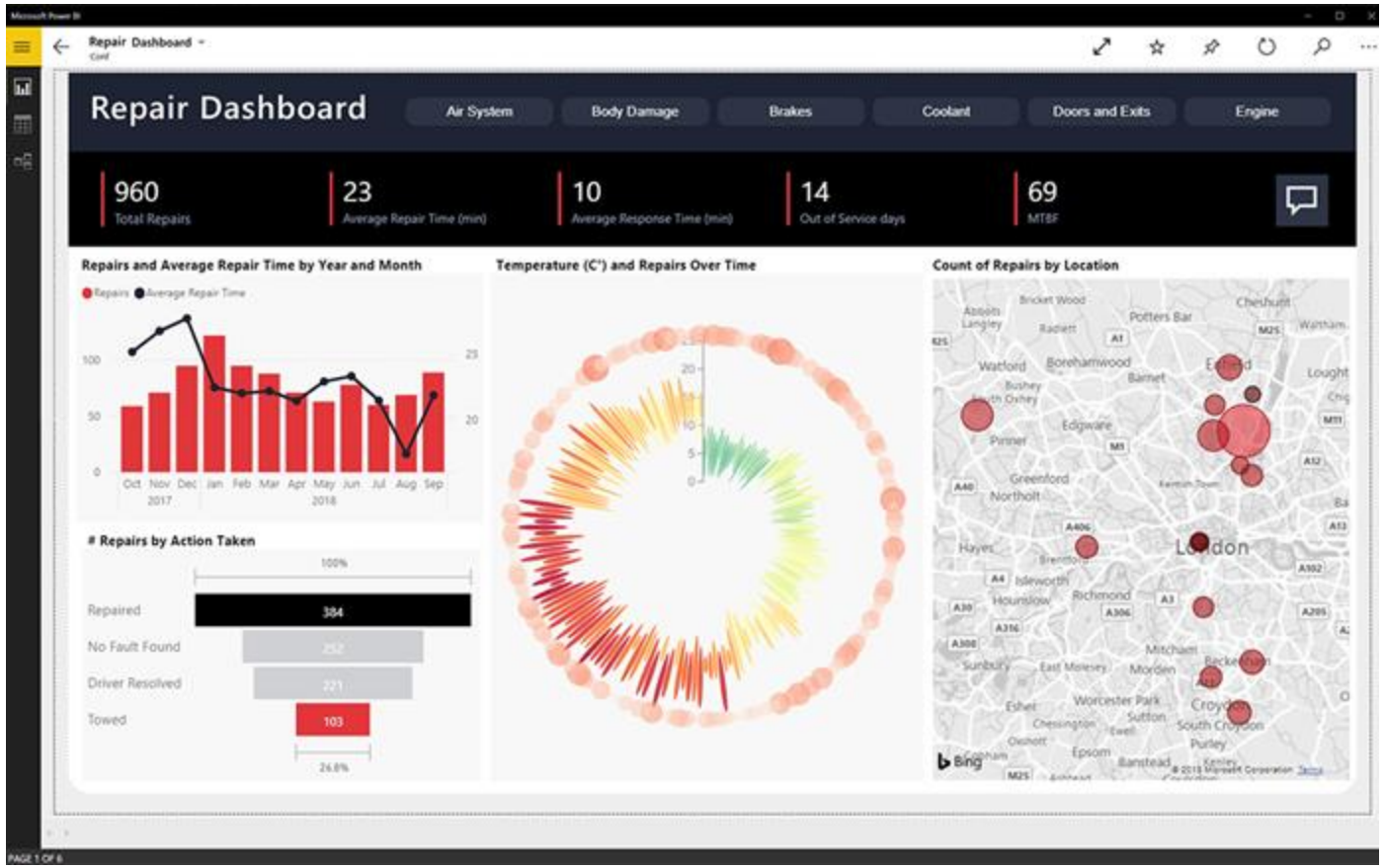
Python

- Python is a programming tool.
- You'll find libraries for practically every data visualization need.
- It is free and open source.



Power BI

- Power BI can create interactive visualizations and has business intelligence capabilities.
- It has a simple interface.
- Many use it to create dashboards.



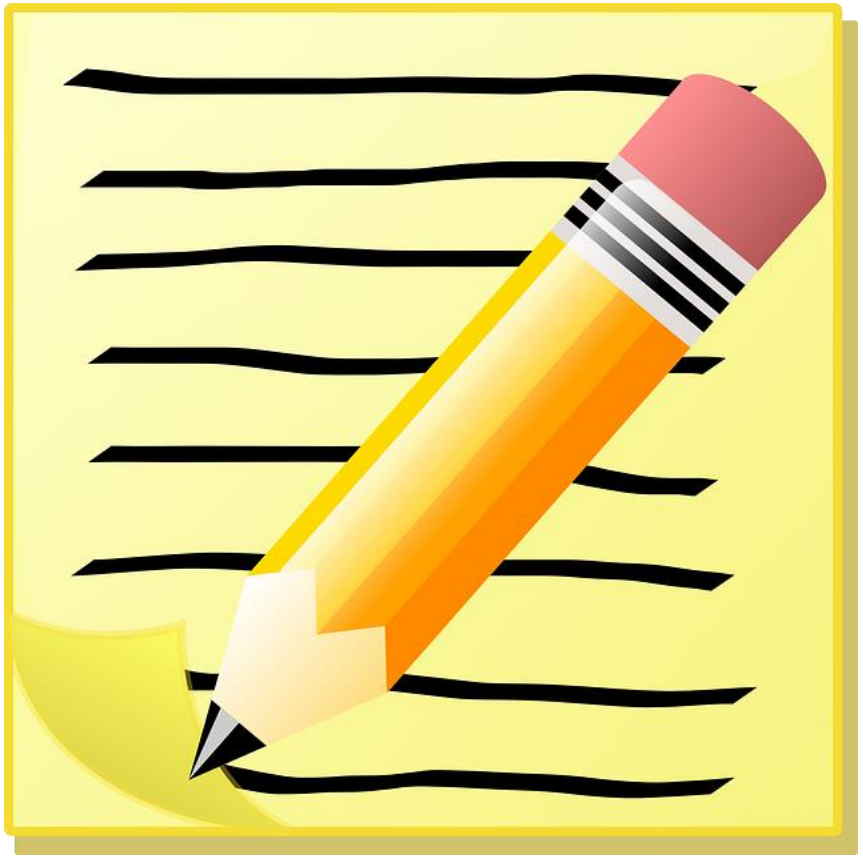
Infographic tools

- An infographic is a collection of imagery, charts, and text that gives an easy-to-understand overview of a topic.
- There are also many specialized tools on the market for creating infographics These include Piktochart, Adobe Spark, and Canva.
- General tools such as Visio and Microsoft PowerPoint can also be used.

https://www.freepik.com/free-vector/great-infographic-template-with-traveler-checklist_1040519.htm

Getting started on paper

- Many great ideas begin with pencil and paper.
- Rapid prototyping visualizations using analog tools allows you to experiment with and compare multiple options.
- Avoid sinking time and energy into a tool at the beginning of the process.



How do you pick the best tool?

- Which tool you should select depends on your specific needs.

- Consider:
  - budget
  - organizational restrictions
  - compatibility with existing systems (e.g., operating systems, hosting platforms)
  - simplicity of use
  - ability to create and customize the types of visualizations you need
  - ability to share and collaborate
  - security and maintenance

See page 29 of the participant guide for a tool selection checklist

Recap

- Before you begin a visualization consider:
  - Who the audience is
  - What your message is
  - How you can best communicate

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

The anatomy of a chart

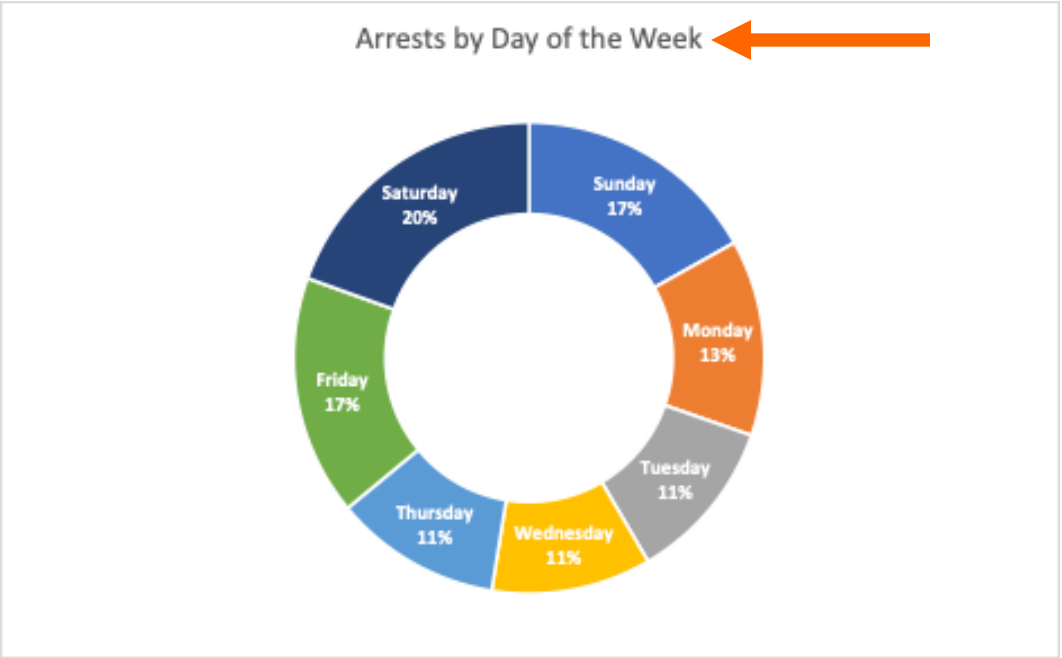
- The elements of a chart will vary by chart type, but most have a similar anatomy.

- Title
- “Y” axis
- “X” axis
- Axis titles and labels
- Gridlines
- One or more data series
- Legends
- Markers or data labels
- Trendlines

Title

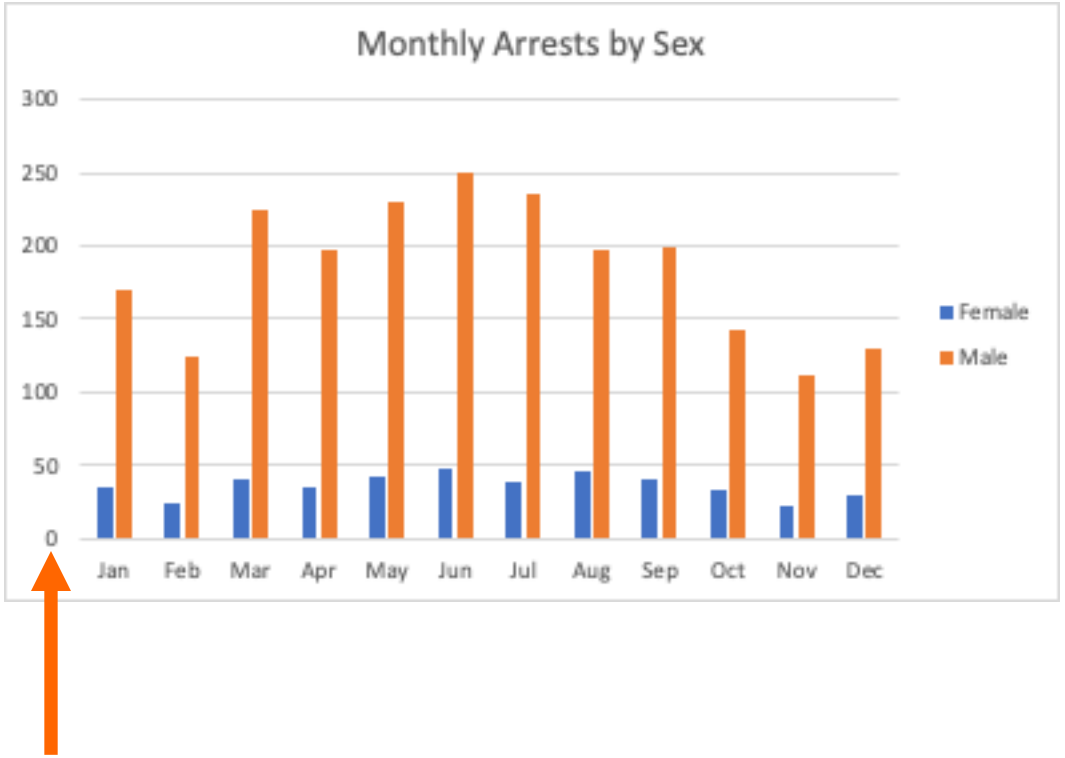
- The title explains the chart.

- It's good practice to:
  - summarize the main point using plain language
  - include units of representation (e.g., $ in millions)
  - include the time period (e.g., FY 2019)
  - keep it short and precise (avoid using articles and adjectives)



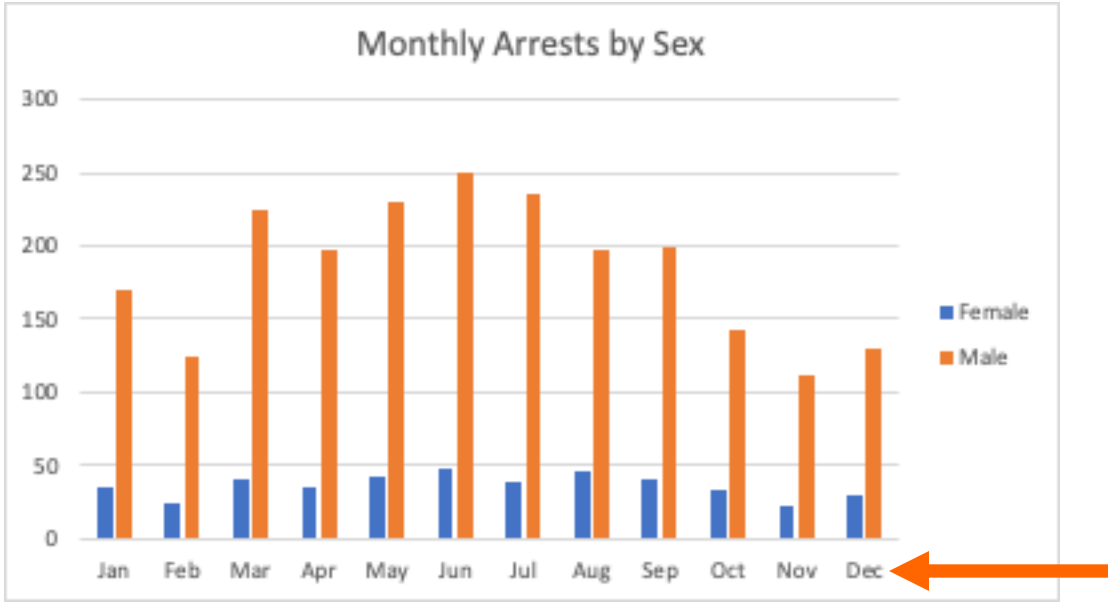
“Y” axis

- The Y axis displays values.
- It is vertical on most charts.
- It is segmented into ticks, usually of equal size.



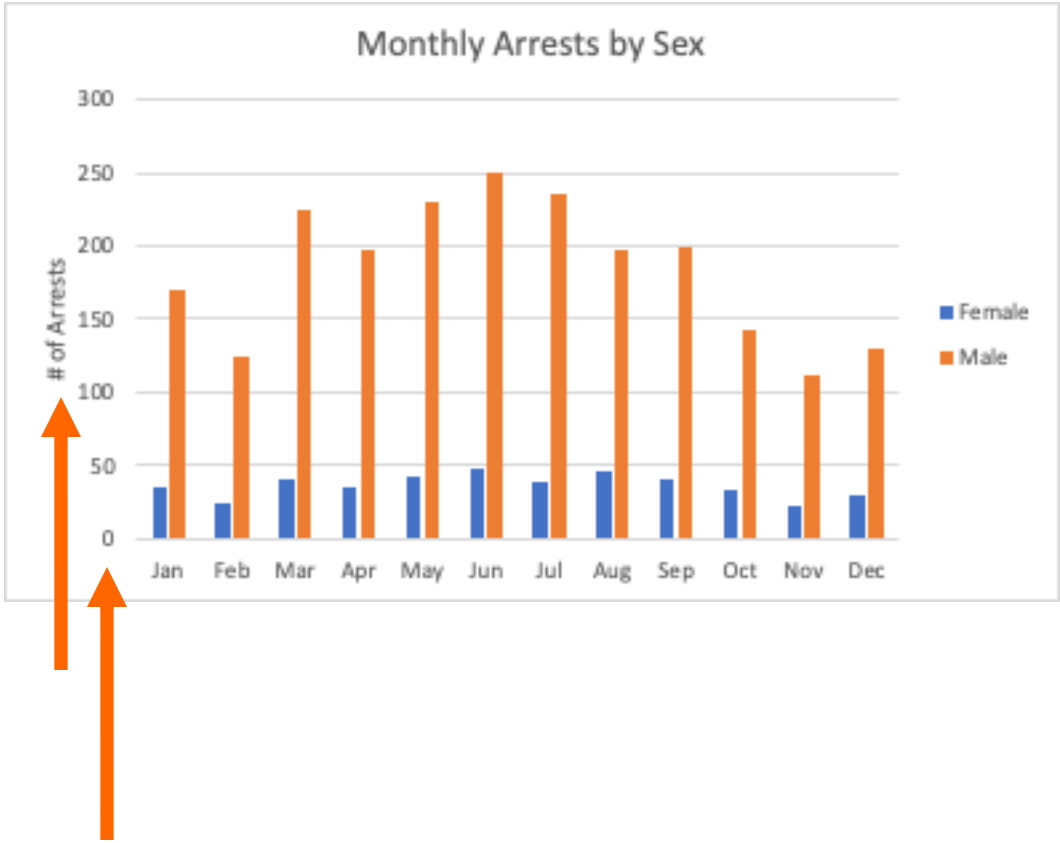
“X” axis

- The X axis displays category names.
- It is horizontal on most charts.
- It is segmented into ticks, usually of equal size.



Axis titles and labels

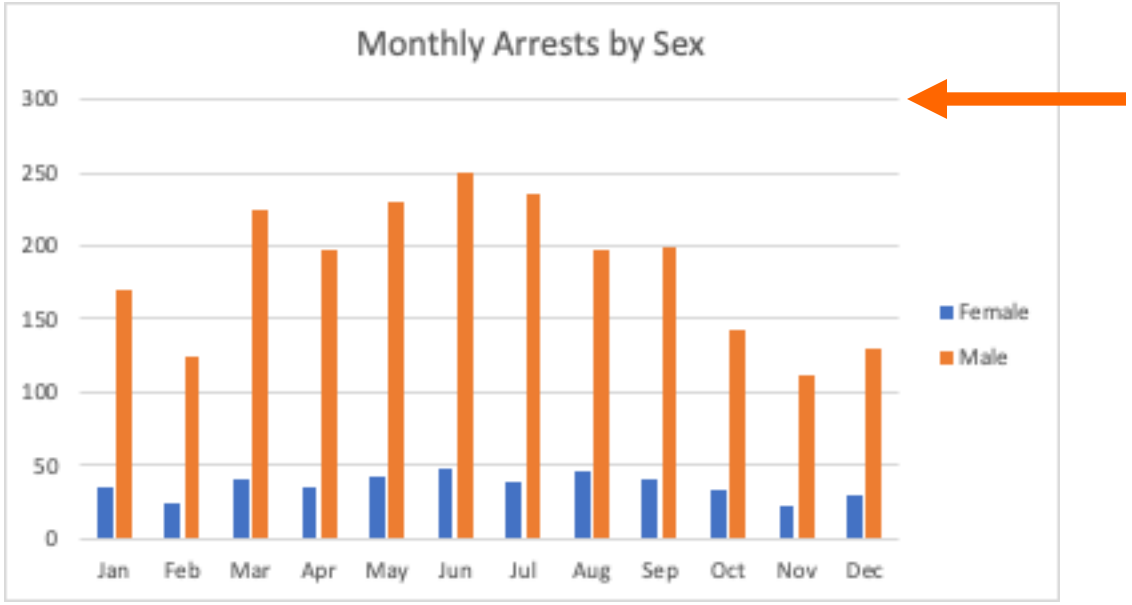
- The axis title uses words to explain the entire axis.
- The axis labels are words or numbers that mark the different axis ticks.



Gridlines

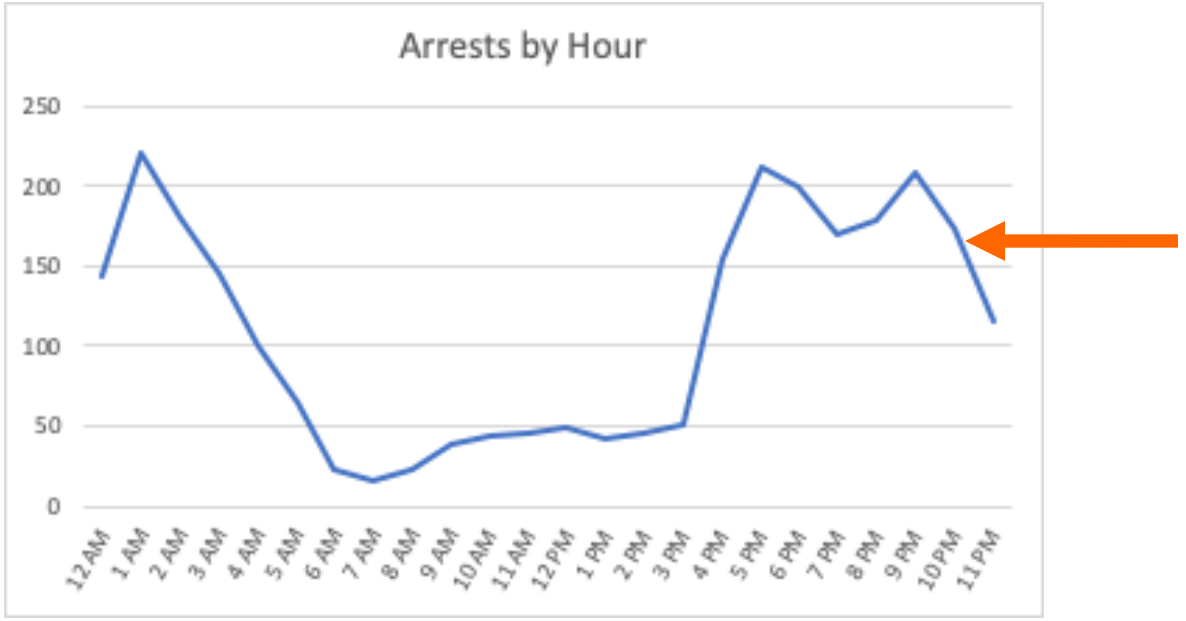
- Gridlines are horizontal or vertical lines that extend from the axis.

- They can be used to make data easier to read (but sometimes make a chart too cluttered).



Data series

- A data series is a set of related data.
- A chart can have one or more series.
- Each chart type displays series differently.
- Often series correspond to a row or column of data.



Legend

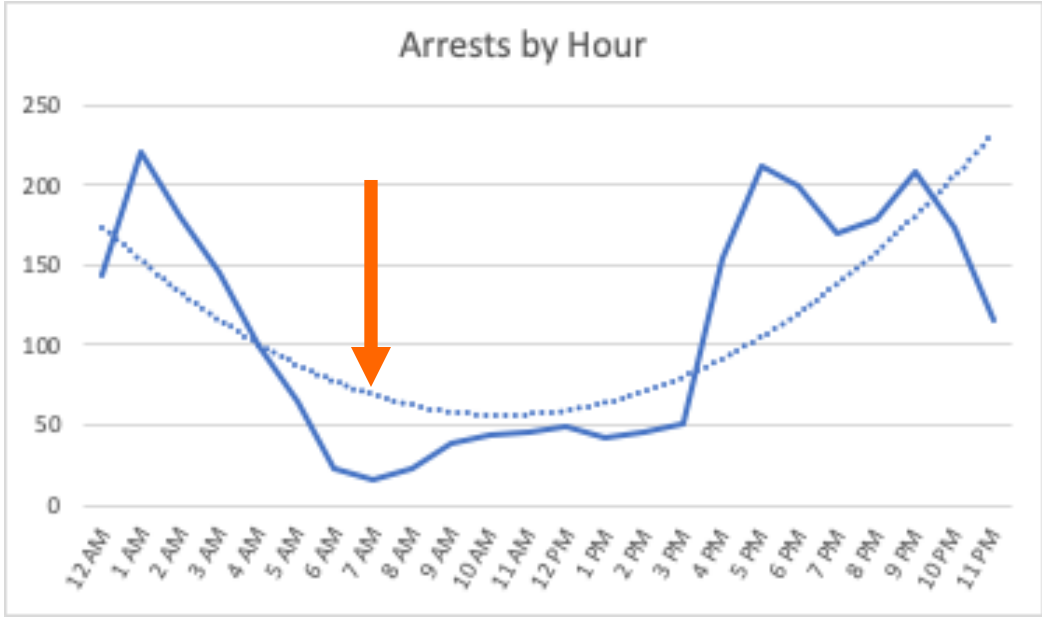
- A legend shows what kind of data is represented in the chart.
- It may identify patterns, colors, or symbols associated with the markers of a chart data series.

Markers / data labels

- Markers, or data labels, identify a single data point.
- They can be used on all data points in a series or only some.
- They can be a good way to emphasize a particular piece of data on the chart.

Trendline

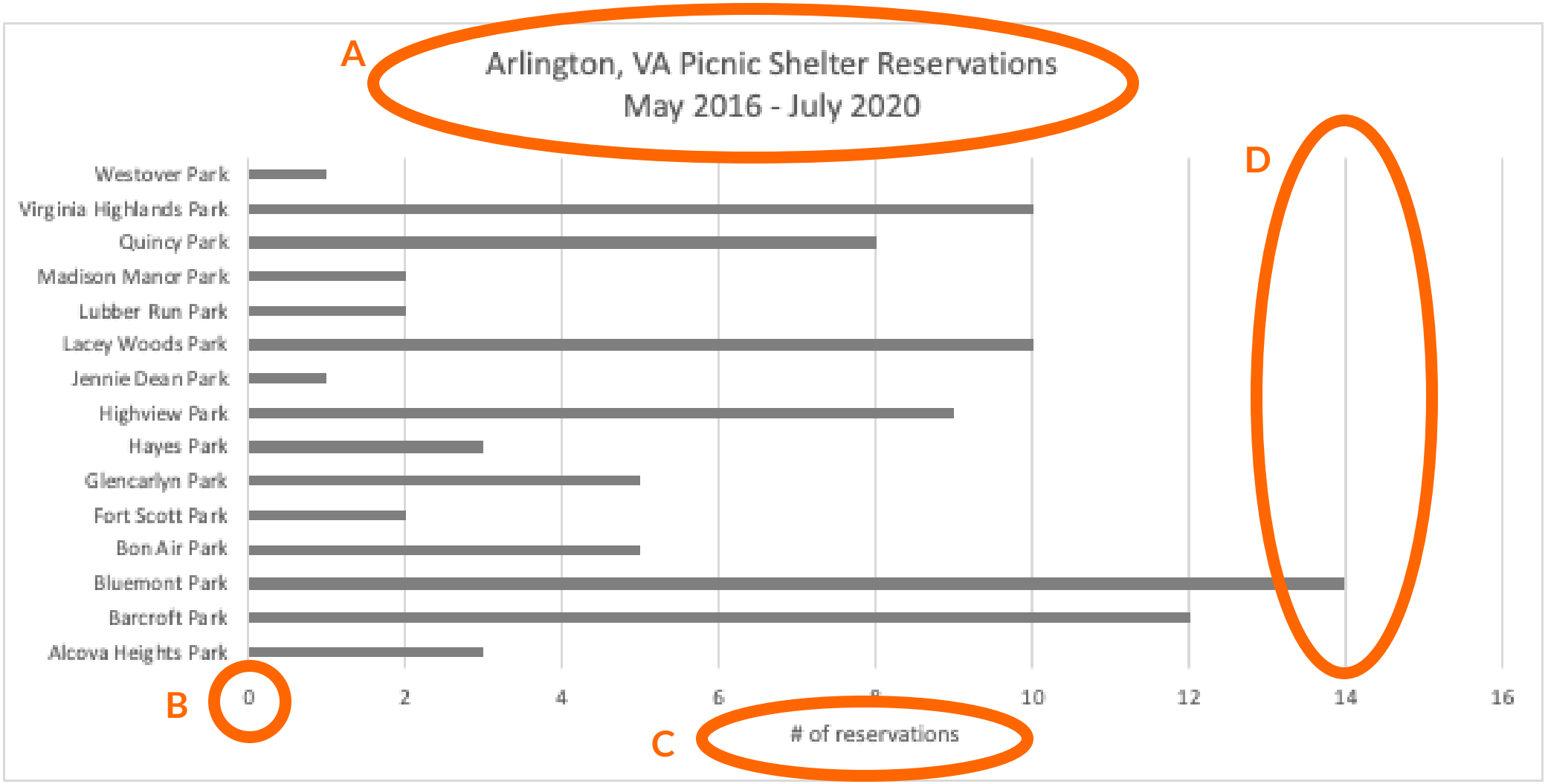
- A trendline is a line that shows the general pattern or overall direction of the data.
- It can be straight or curved.



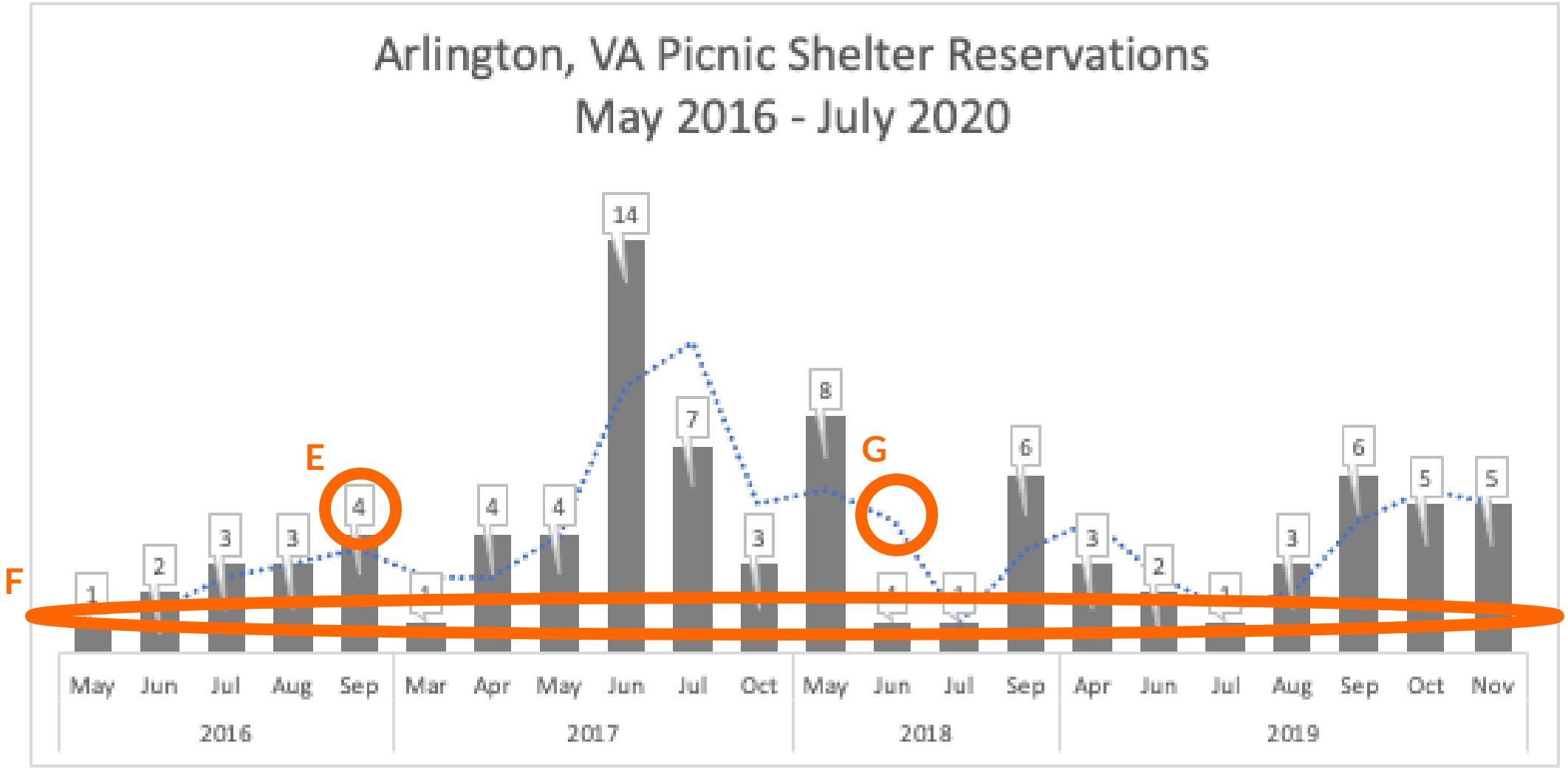
Activity: anatomy of a chart

- Turn to page 7 of your participant guide to find the **anatomy of a chart** activity.
- Review the charts and name the circled elements from the options provided.

Activity: anatomy of a chart



Activity: anatomy of a chart



Q&A

DATA SOCIETY®

Welcome back
Day 2

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

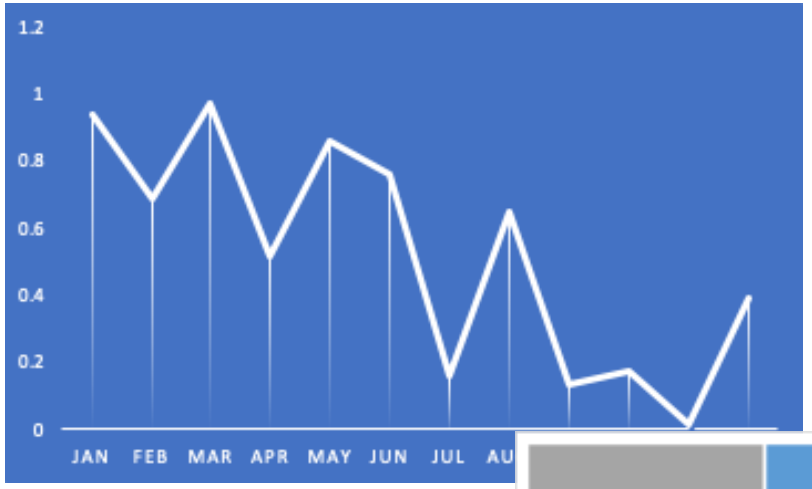
- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

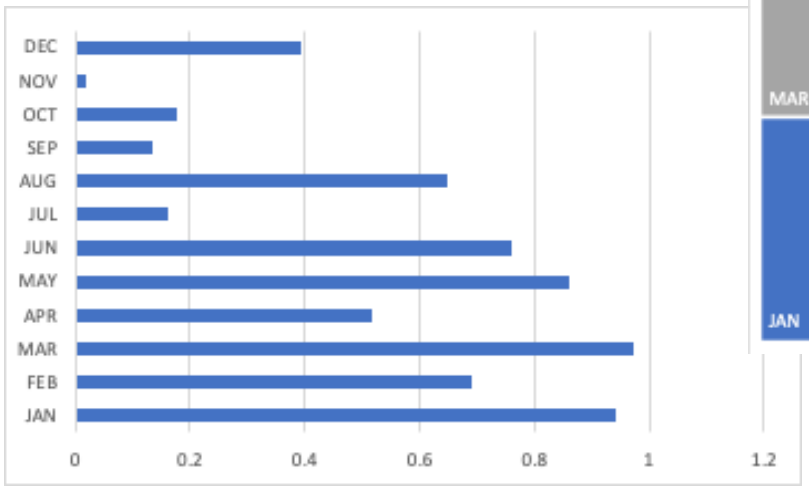
Day 4

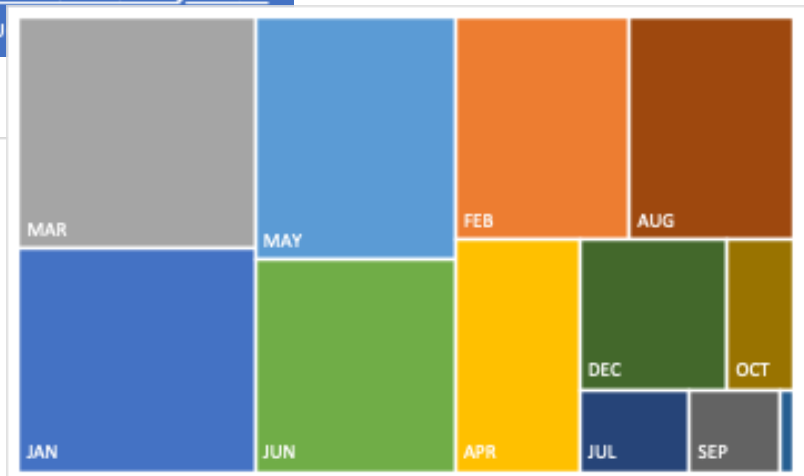
- Data storytelling

Using appropriate visuals

- We'll review the most common charts and graphs in this section.
- Later, we'll talk about which ones to use in specific situations.







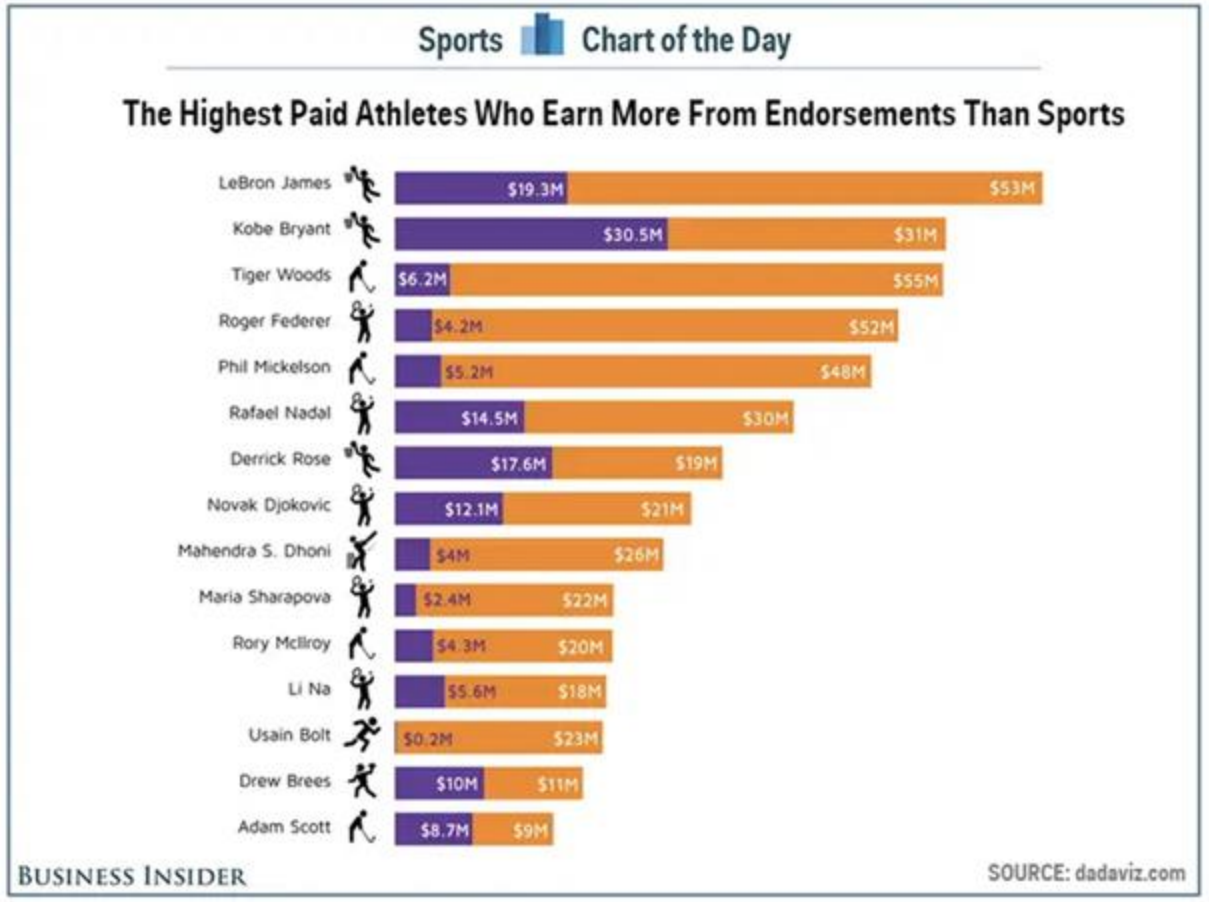
Polling question

What types of charts and graphs do you frequently create or review?

- Bar charts, line charts, or pie charts
- Maps or heatmaps
- Scatter plots or bubble charts
- Boxplots or histograms
- Something else
- I don't do this frequently

Bar chart

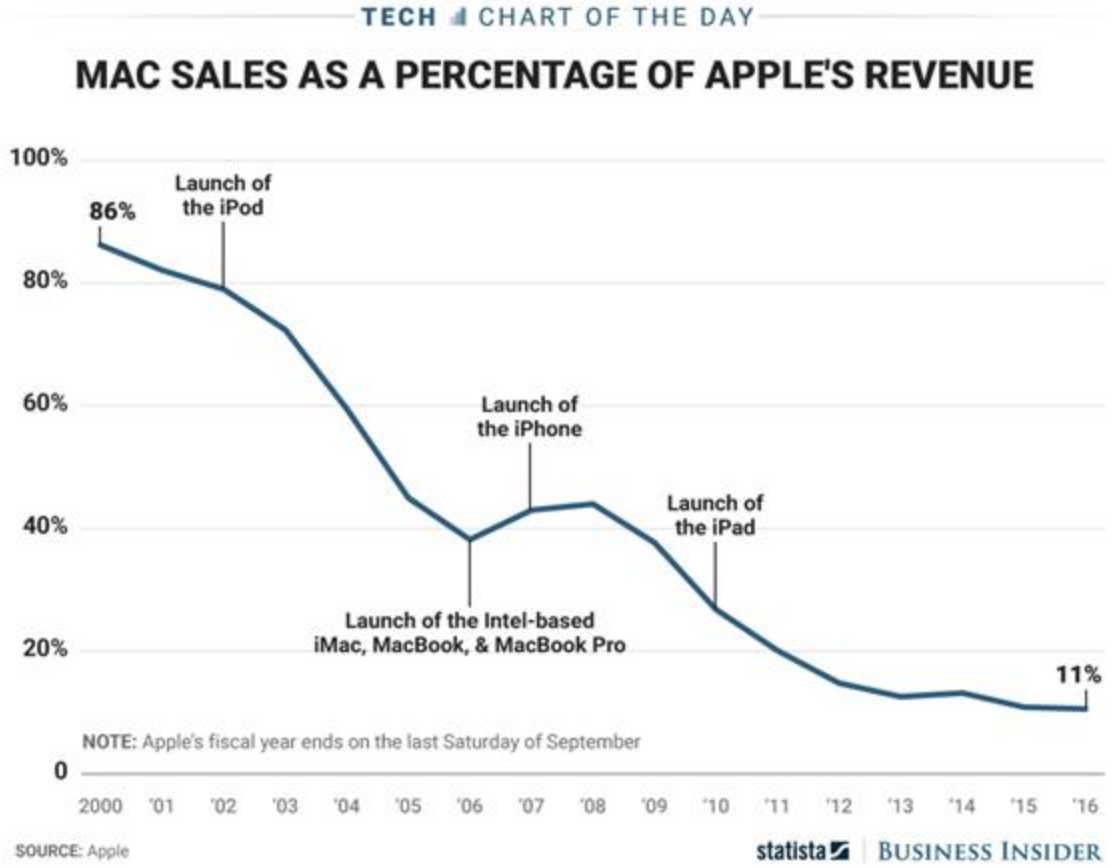
- A bar chart is a chart with rectangular bars with lengths proportional to the values that they represent.
- One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value.
- Vertical bar charts are also called column charts.



Line chart

- A line chart displays information as a series of data points called ‘markers’ connected by straight line segments.

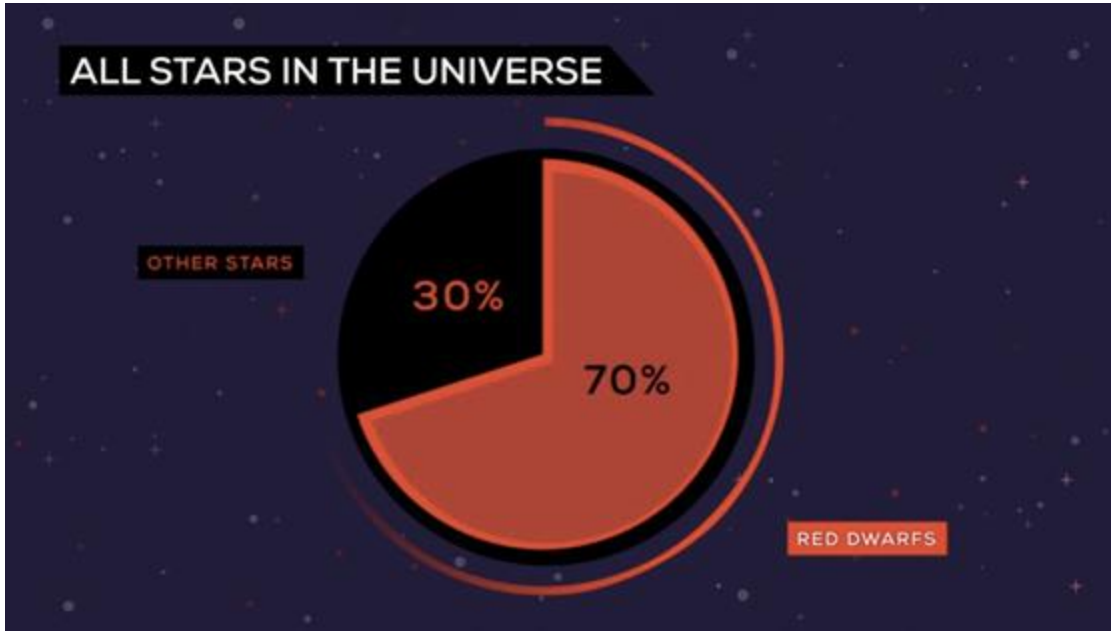
- An area chart is a line chart with the area below the lined filled with colors or textures.



http://www.datavizdoneright.com/2017/04/mac.html

Pie chart

- A pie chart is divided into sectors, illustrating numerical proportion.
- In a pie chart, the area of each sector is proportional to the quantity it represents.
- A doughnut chart is a pie chart with a blank center that can be used to display additional data.

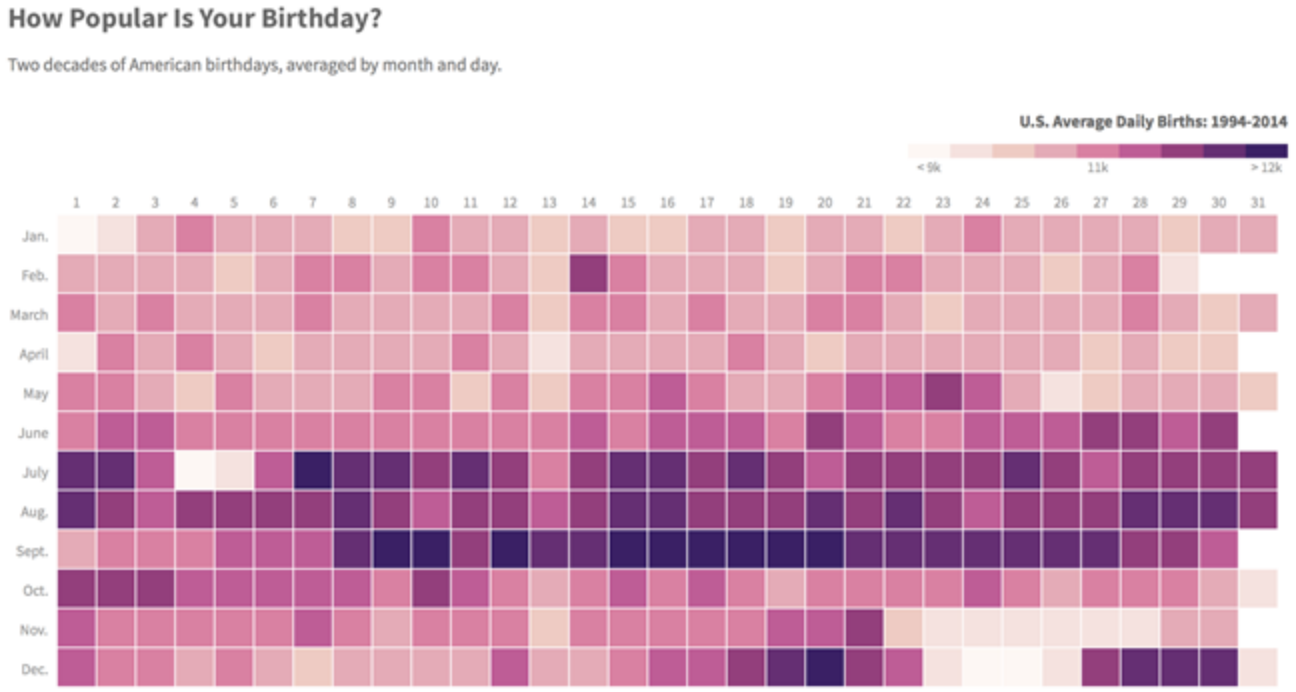


https://www.informationisbeautifulawards.com/showcase/1195-the-last-star-in-the-universe-red-dwarfs-explained?utm_content=buffer0e946&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Heatmap

- In a heatmap, individual values are contained in a matrix with variations in coloring.

- Heatmaps are useful for visualizing variance across multiple variables to display patterns in correlations.

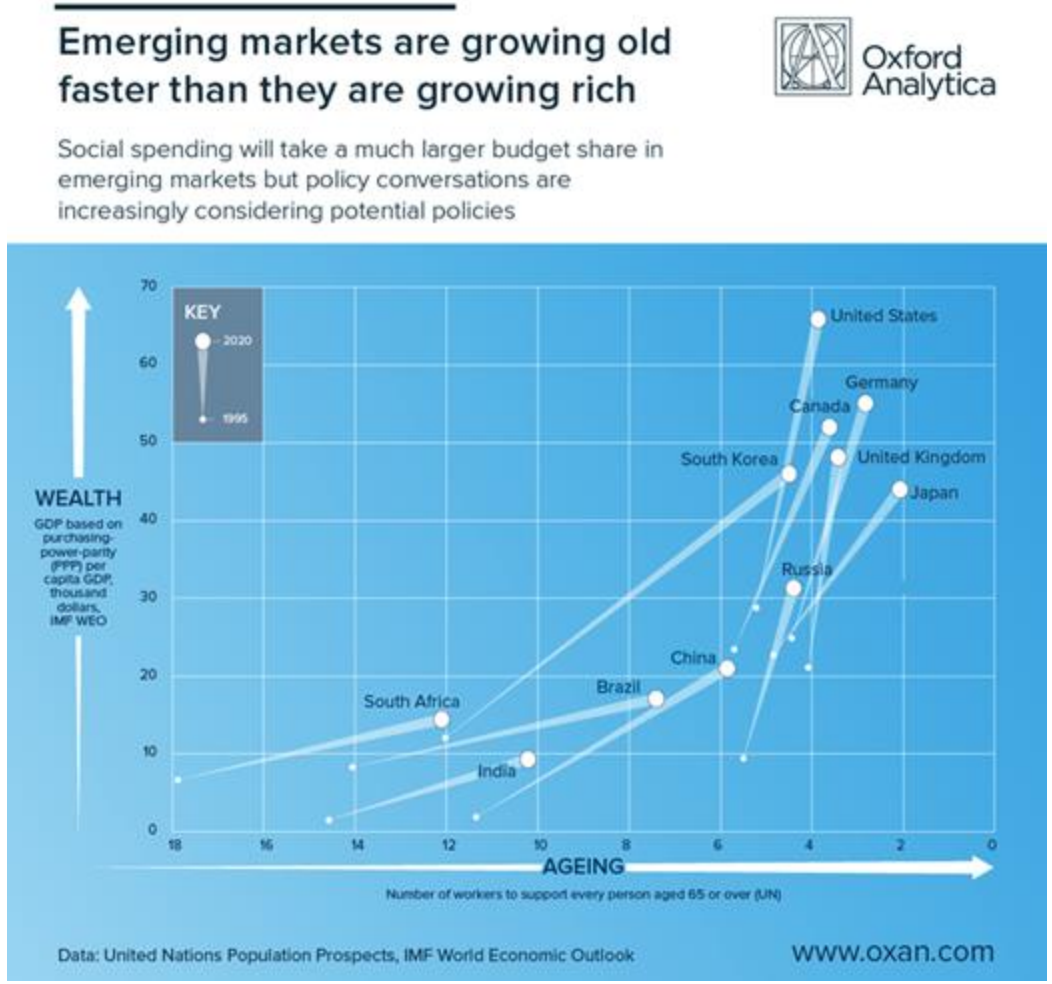


http://thedailyviz.com/2016/09/17/how-common-is-your-birthday-dailyviz/

Scatter plot

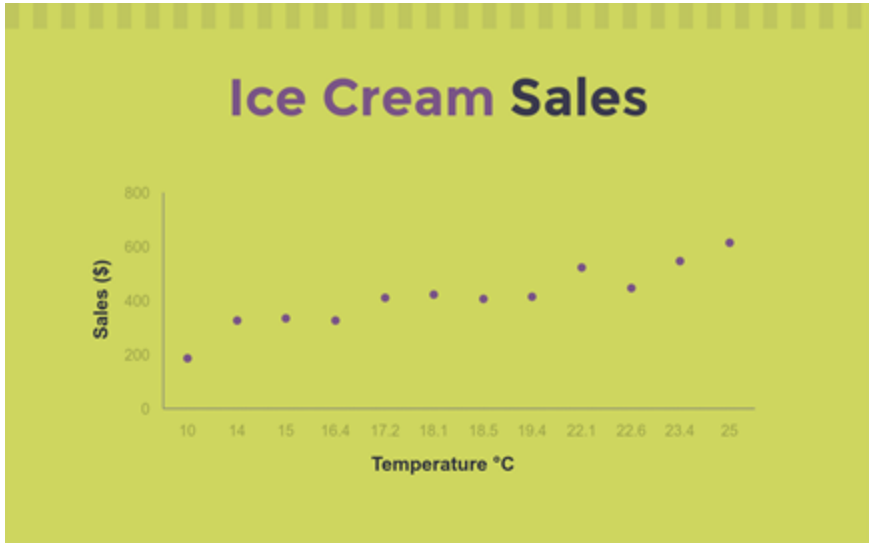
- A scatter plot is a type of diagram that uses coordinates to display values for two variables for a set of data.

- The data is displayed as a collection of points, each having the value of one variable determining the positions on the X and Y axes.



How to read a scatter plot

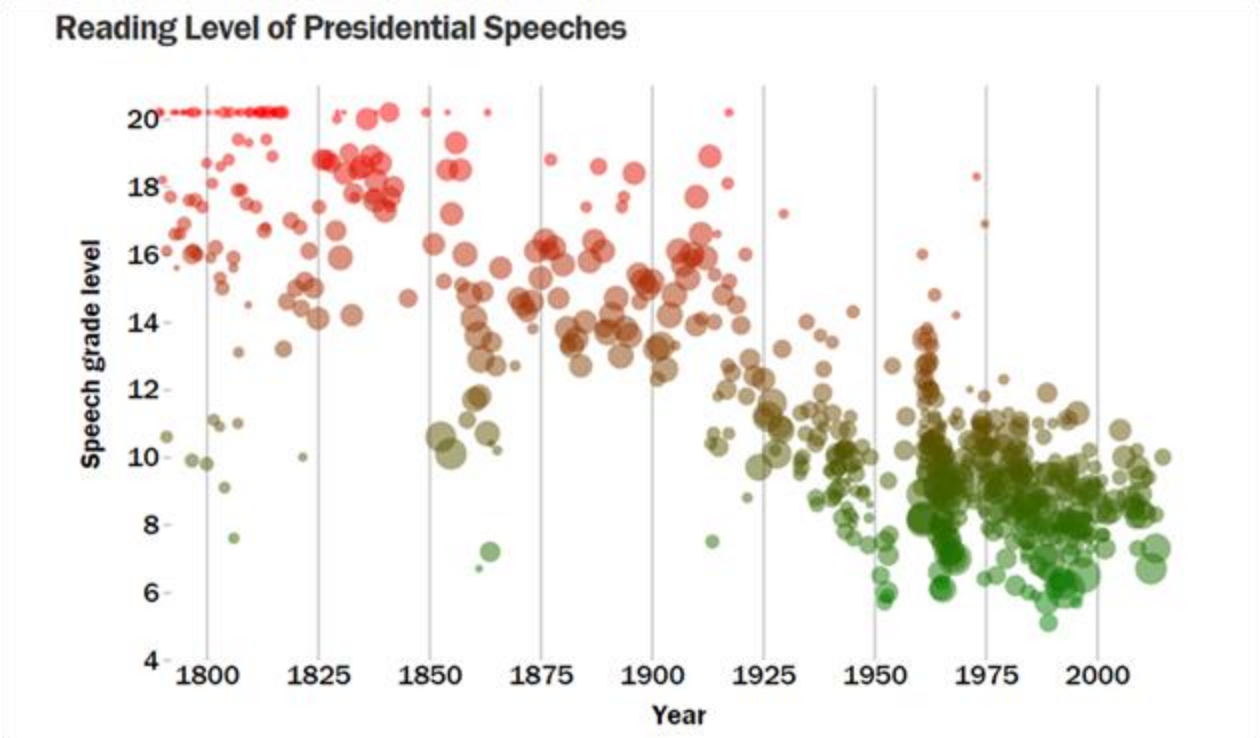
- If the data moves uphill there is a positive relationship between X and Y (as temperature increases, ice cream sales tend to increase)
- If the data moves downhill there is a negative relationship between X and Y (as months worked increases, internet usage tends to decrease)
- If the data doesn't have a pattern, then no relationship exists between X and Y.



https://www.visme.co/scatter-plot-maker/

Bubble chart

- A bubble chart displays three dimensions of data.
- Each entity is plotted as a bubble that expresses two of the values through the xy location and the third through its size.

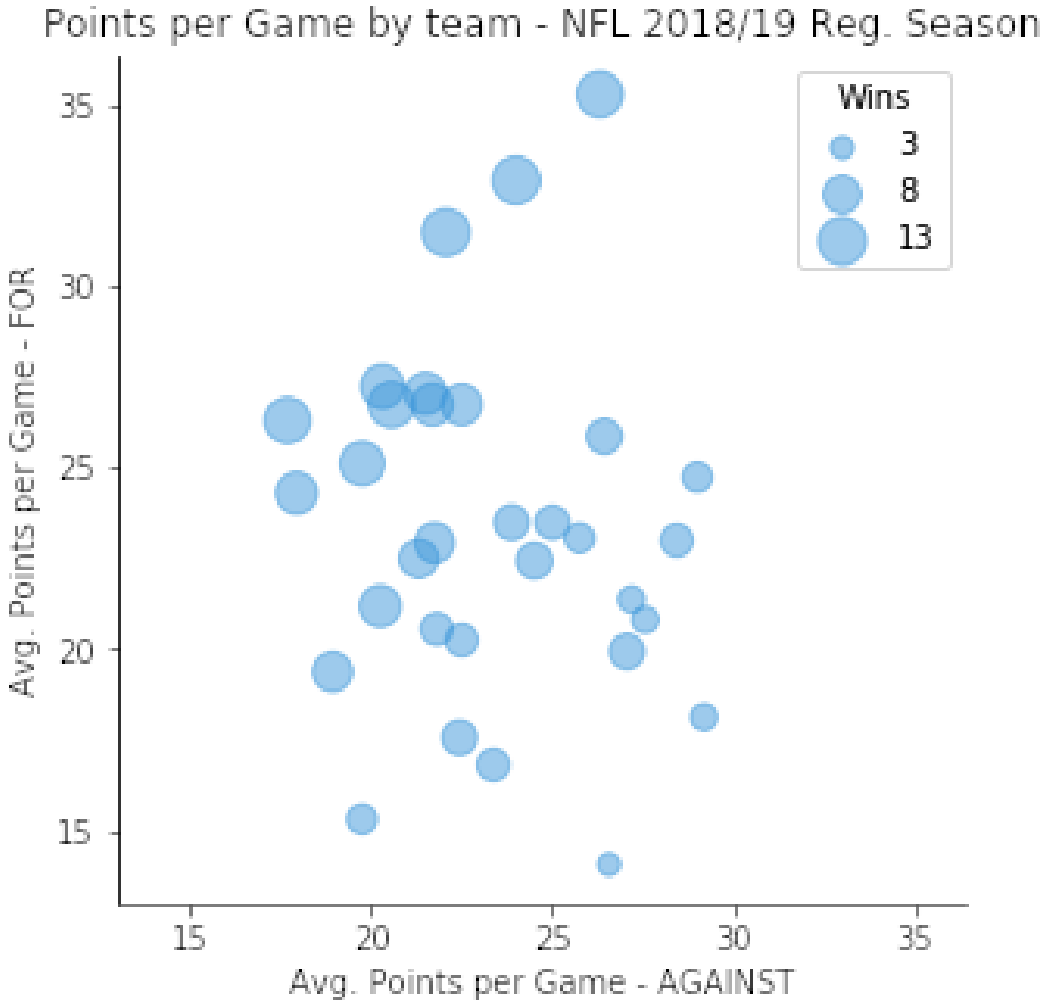


http://dadaviz.com/i/1276

How to read a bubble chart

1. Look for a relationship between the x and y variables
   - Positive
   - Negative
   - No relationship

2. Look for patterns in bubble size
   - consistency in bubble size means there is less variation in the relationship between the variables

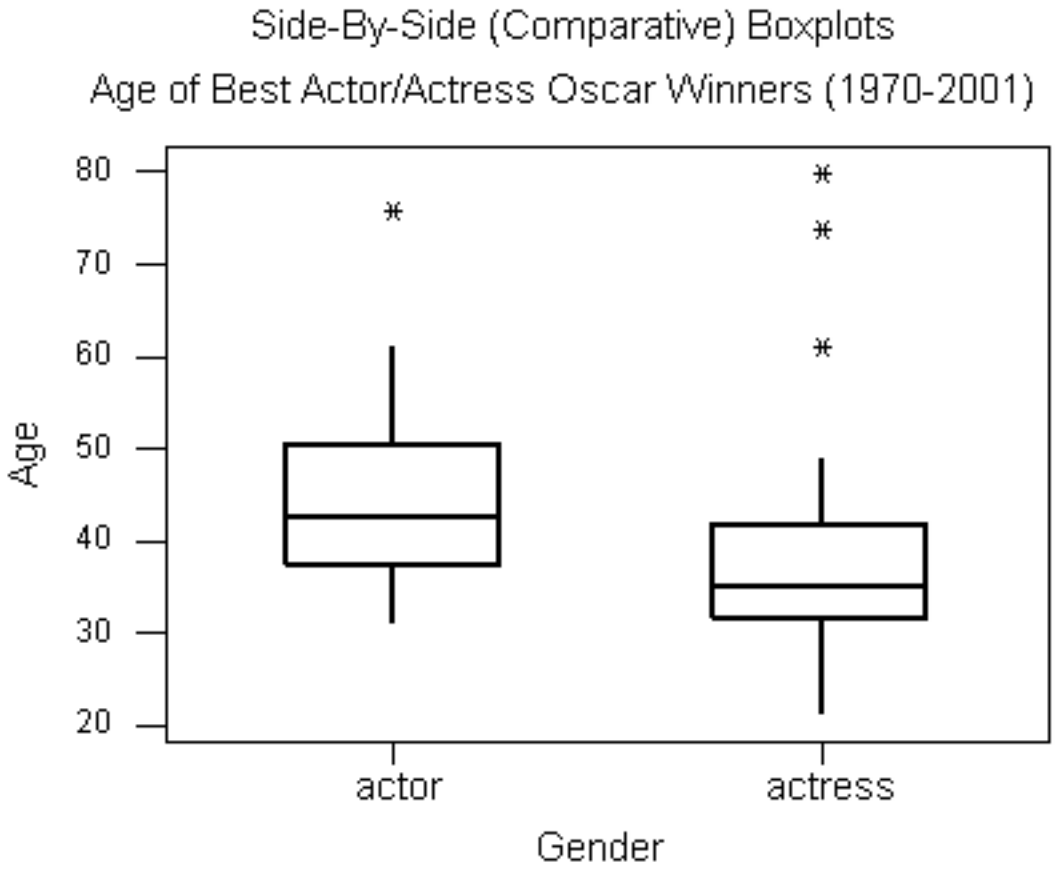


https://chartio.com/learn/charts/bubble-chart-complete-guide/

Boxplot

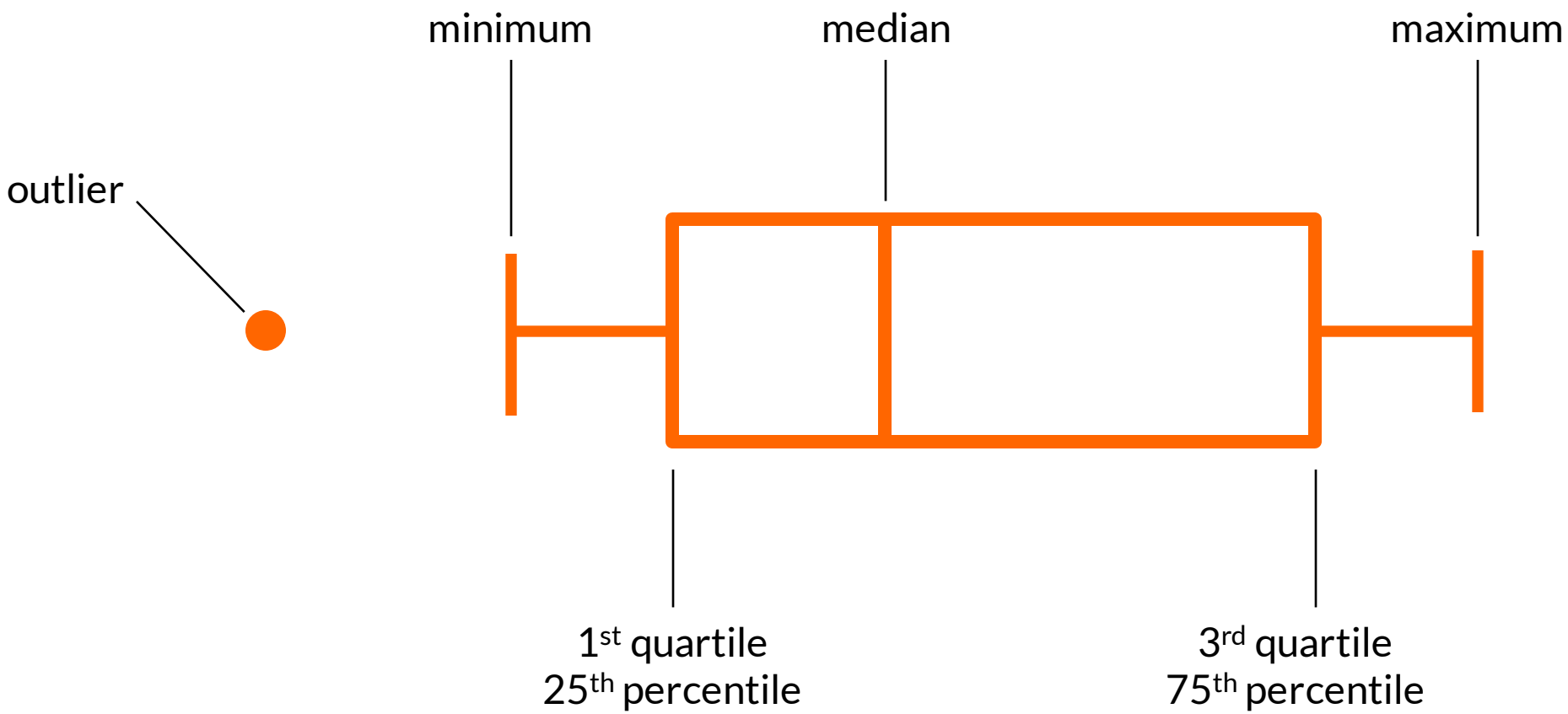
- A boxplot displays median, higher/lower quartiles and maximum/minimum.

- The spacings between the different parts of the box indicate the degree of dispersion and skewness in the data and show outliers.



https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot

How to read a boxplot

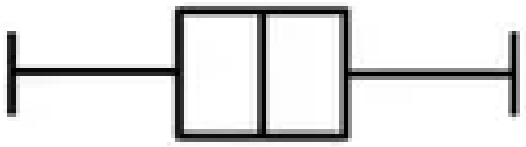


How to read a boxplot

Left-skewed

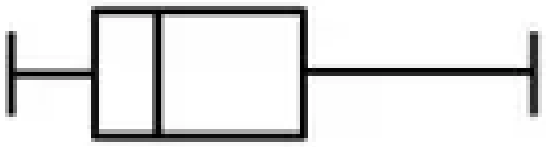
- When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

Symmetric



- When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.

Right-skewed



- When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).

How to read a boxplot

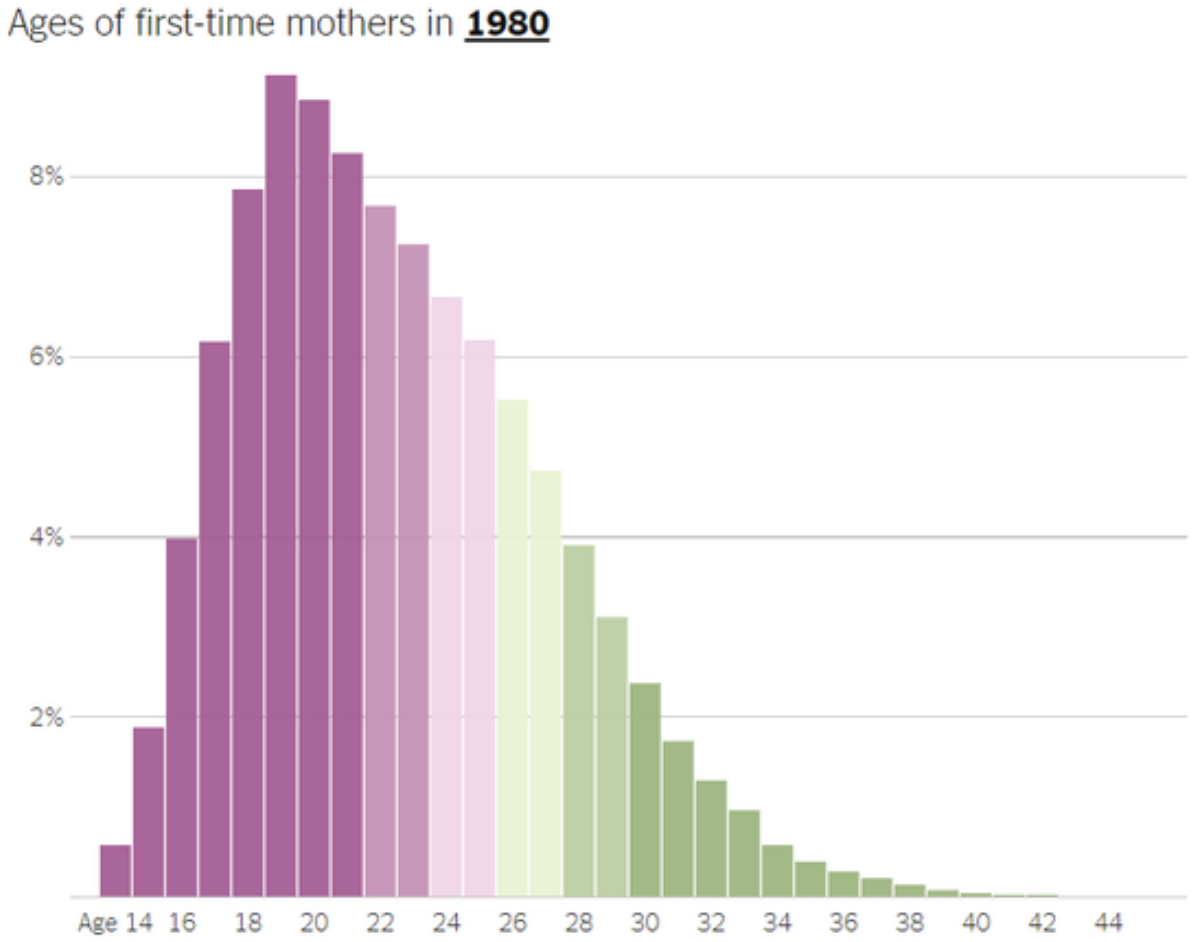
https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot

	Actors	Actresses
Minimum	31	21
Q1	37.23	32
Median	42.5	35
Q3	50.25	41.5
Maximum	76	80

- Takeaways:
  - In general, actresses win the Best Actress Oscar at a younger age than actors do.
  - The actors' ages are more alike than the actresses' ages. However, the middle 50% of the age distribution of actresses is more homogeneous than the actors' age distribution.
  - We have outliers in both distributions.

Histogram

- A histogram groups numeric data into bins, displaying the bins as segmented columns.
- They're used to depict the distribution of a dataset: how often values fall into ranges.

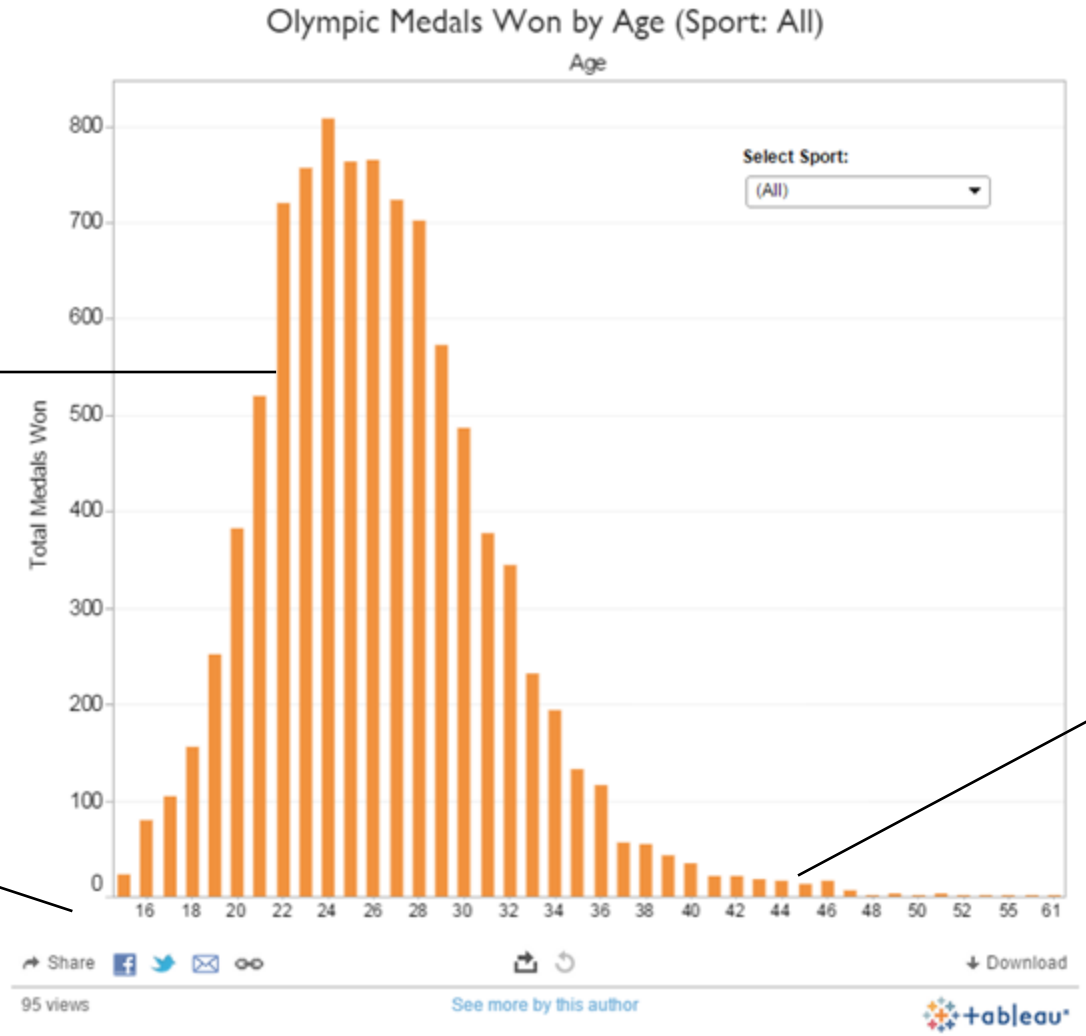


https://www.nytimes.com/2018/11/22/learning/whats-going-on-in-this-graph-nov-28-2018.html

How to read a histogram

The area of the bar (height x width) tells the number of occurrences

Data is split into intervals called bins, which should be neither too small nor large to see patterns



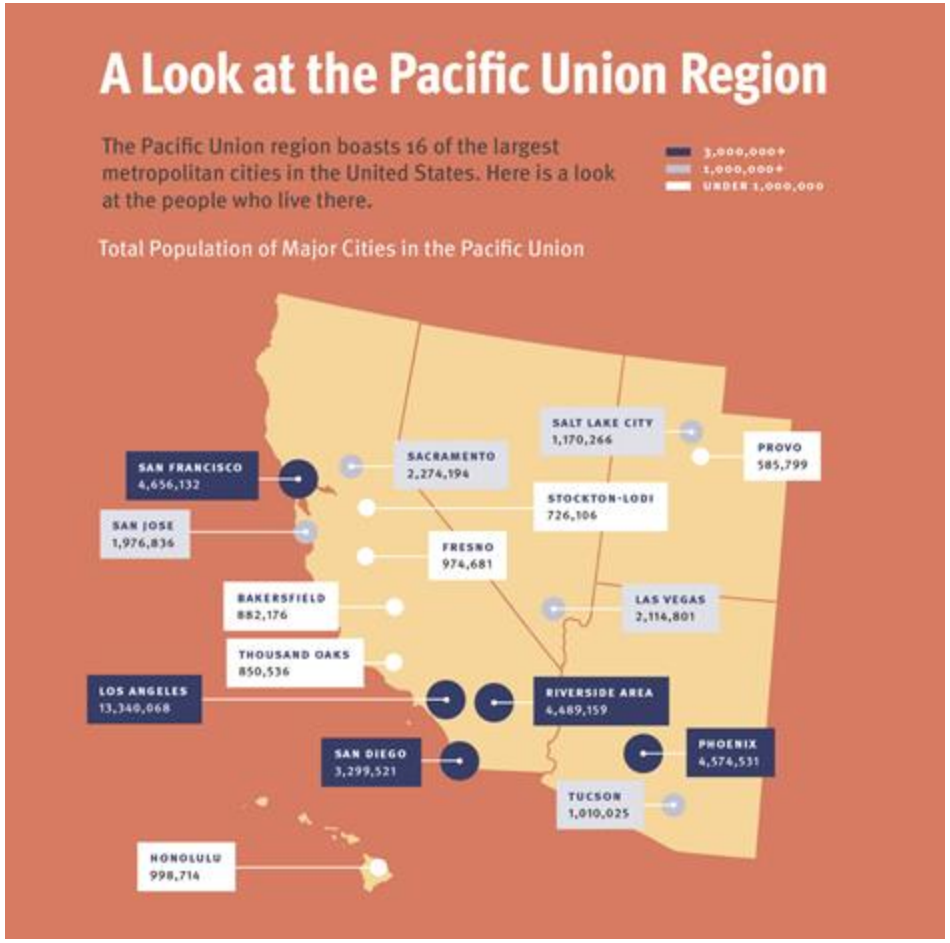
Generally there won't be a gap between the bars because histograms work well with continuous data

https://www.tableau.com/about/blog/2014/12/5-chart-types-youve-never-tried-tableau-35281

Map

- Maps present geographically-related data in a clear and intuitive manner.

- Maps are often combined with points, lines, bubbles, and more.



https://dribbble.com/shots/2881738-Infographic/attachments/595499

Activity: common charts and graphs

- Turn to page 9 of your participant guide to find the **common charts and graphs** activity.

- Review the charts and answer the questions that follow.

Break

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Choosing the type of visual

- Remember, we first define the who and what, and then the how.

- The type of visual you choose should depend on the data you want to communicate and the message.

- Consider:
  - How many variables do I want to show?
  - How many data points are there?
  - Should values be displayed over time?
  - Should similar items be grouped?

Just a few numbers

- Don't overcomplicate!
- Simple text works well when there is just a number or two to share.

…we spent only $75,000 of our $125,000 budget…

…therefore, it is not surprising that only 29 percent of the applications were accepted…

…product A ($12.99) was much more affordable than product B ($59.99)…

Unique data

- Don’t overcomplicate!
- Tables are great when communicating to a mixed audience who will look to a particular row of interest or when you need to show different units of measure.

Name	Total Hours	Billable Hours
Aiello, Francisco P.	1,880	1,504
Eidson, Virginia D.	2,300	1,280
McVay, Dorothy	1,905	1,086
Ramos, Emilio Pabón	2,037	1,426

Product	Weight	Price
Toaster (UK)	1.05 kg	£17,49
Toaster (US)	3.13 lbs	$29.99
Toaster (South Africa)	1.07 kg	R239,00

Comparisons

- Comparisons help us evaluate and compare values between two or more data points.

- Examples include:
  - Total number of visitors per month, grouped by country of residence, to see where most visitors come from and where to put more efforts
  - Quarterly expenditures for a particular project, to spot trends or performance issues
  - Number of COVID-19 patients by city, highlighting the prevention efforts undertaken in that area

- Go-to visualizations: bar charts, pie charts, & line charts.

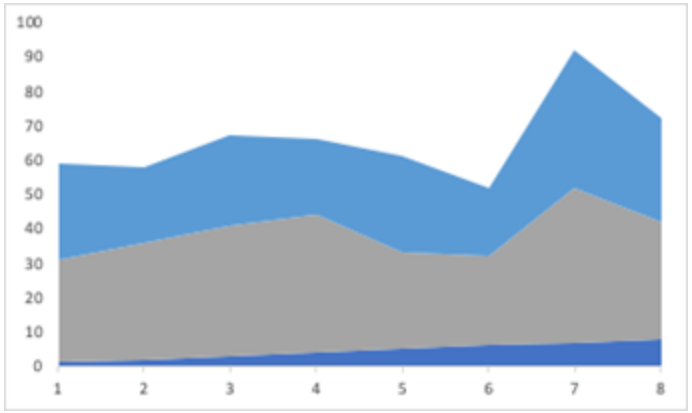
Composition

- Composition will show how individual parts make up the whole.

- Examples include:
  - Advertising spend, by medium, for a given year
  - Total country population by religions, languages, or ethnical groups
  - Total budget, by strategic objective, department, or region

- Go-to visualizations: bar charts, pie charts, waterfall charts, & stacked area charts.

Waterfall chart



Stacked area chart

Distributions

- Distributions combine comparison and composition.

- Examples include:
  - The distribution of ages in a group of people
  - Identifying problems or constraints in quality control systems

- Go-to visualizations: histograms, line charts, area charts, scatter plots, & maps.

Relationships

- Sometimes we want to see the relationships, correlation, or connection of two or more variables and their properties.

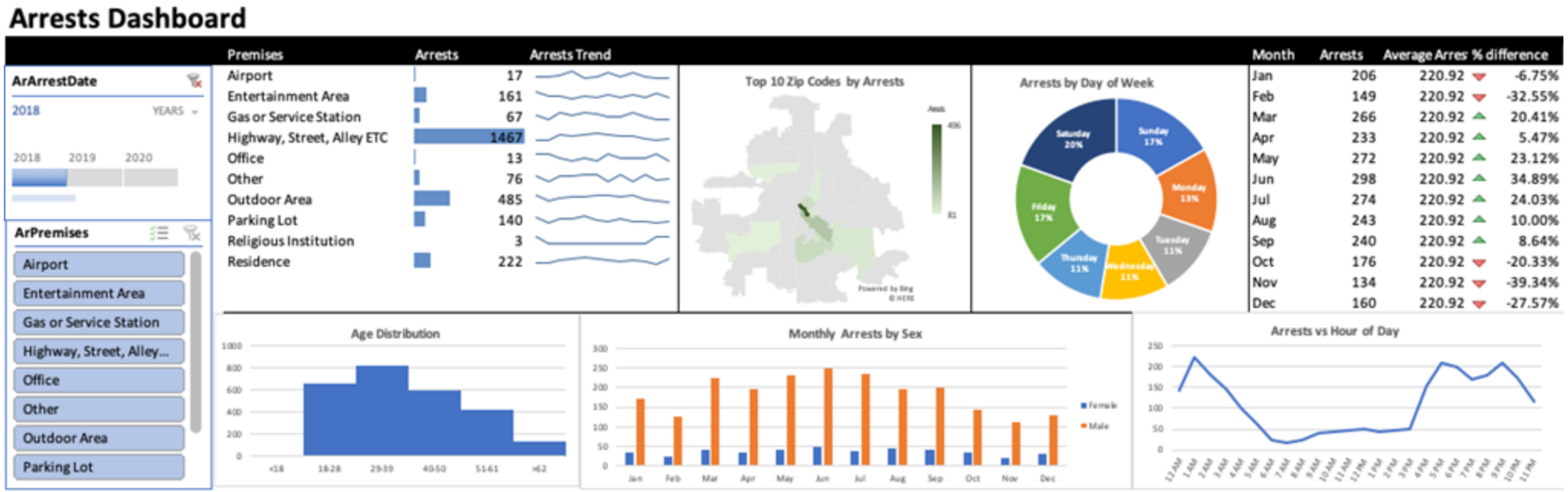
- Examples include:
  - Estimating how expenditures in advertising affect sales
  - Spotting trouble areas by evaluating budget vs. expenses by department or region
  - Answering questions such as, “Does income level depend on education level?”

- Go-to visualizations: scatter plots, bubble charts, & line charts.

Metrics and KPIs

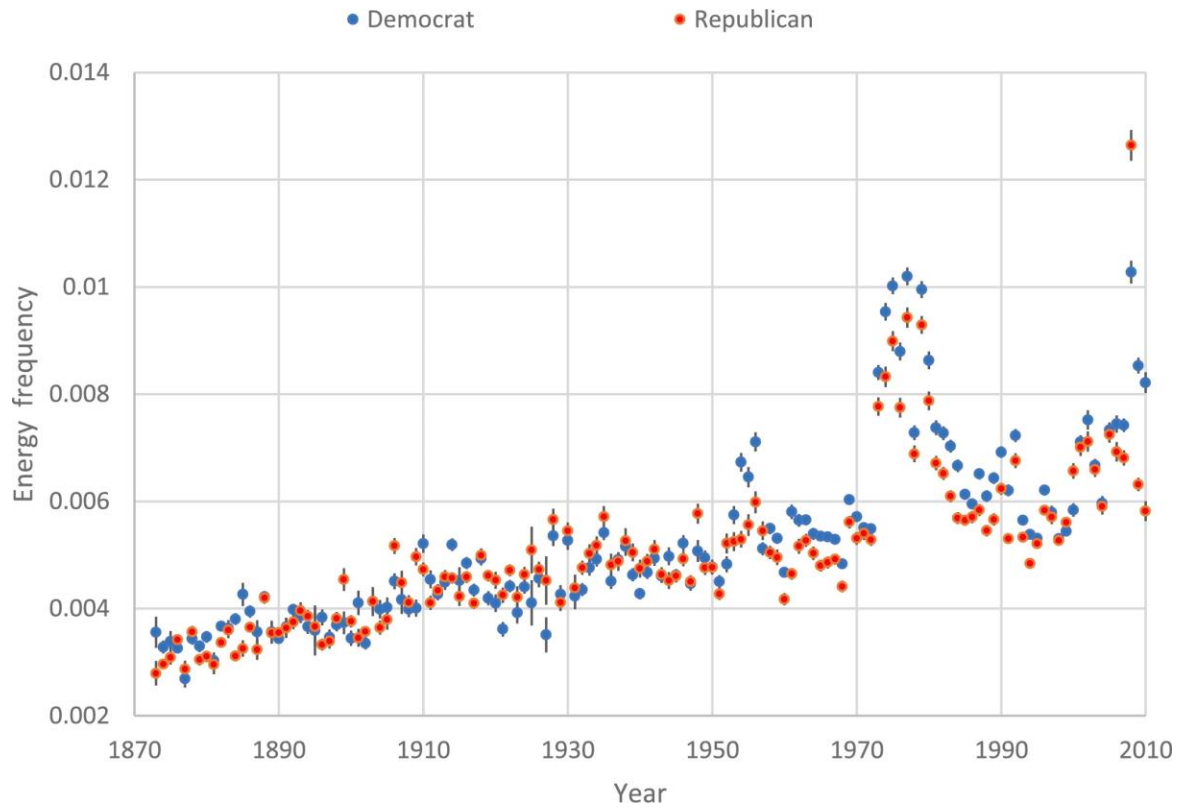
- Use a dashboard to display important metrics and KPIs.

- A dashboard is a collection of data visualizations assembled into a single, unified view. It often presents real-time data.



Poll question

Mean frequency of terms related to energy (e.g., "oil", "gas", "electric", etc.) in congressional speeches

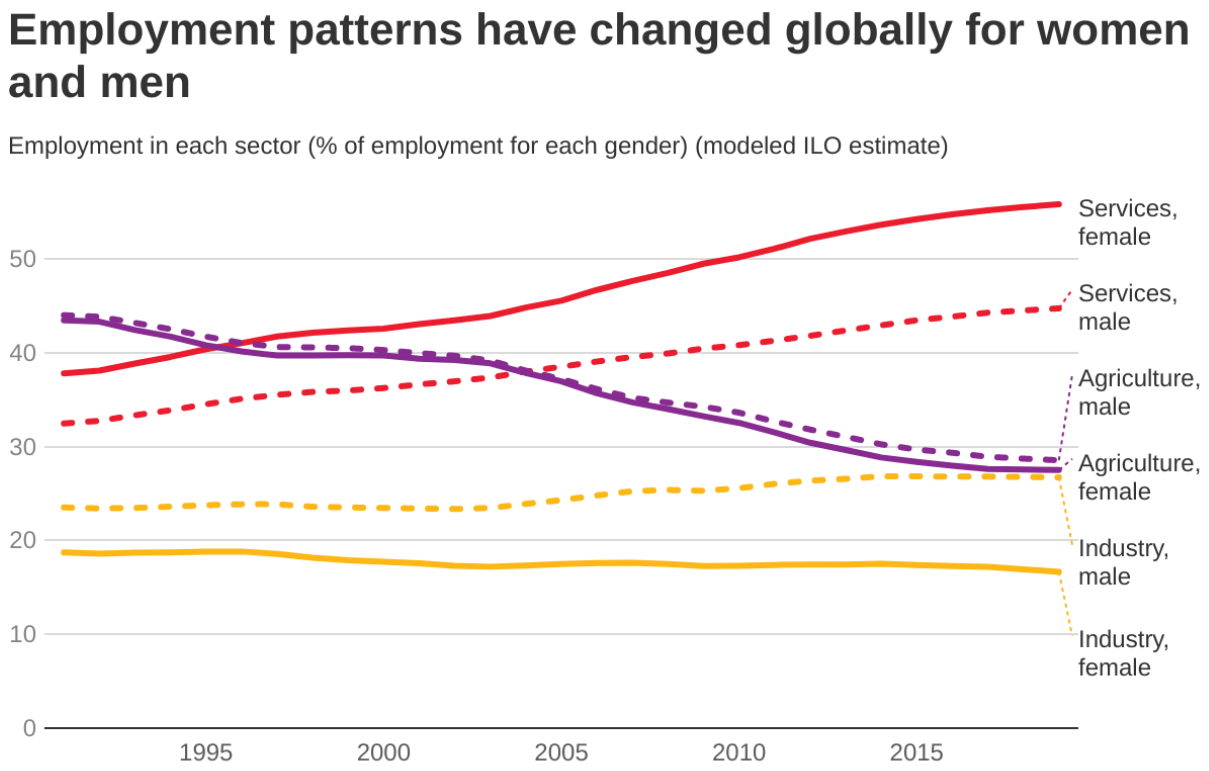


https://www.sciencedirect.com/science/article/pii/S2405844020312615

What kind(s) of function is this chart performing?

- Comparison
- Composition
- Distribution
- Relationship

Poll question



https://www.worldbank.org/en/news/feature/2019/12/20/year-in-review-2019-in-charts

How about this chart?

- Comparison
- Composition
- Distribution
- Relationship

Poll question

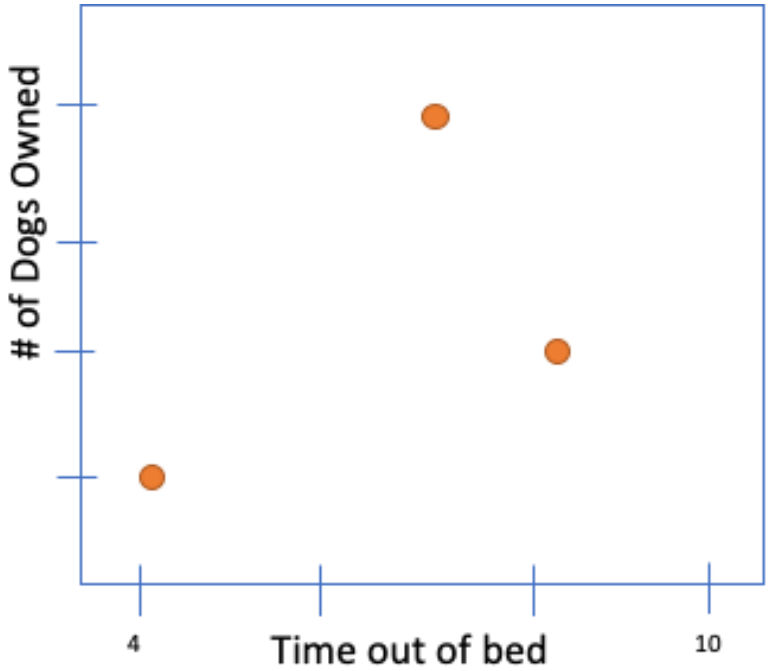
DailyFX Global Commodities Interactive Tool
https://www.dailyfx.com/research/global-commodities/

Lastly, what about this interactive tool?

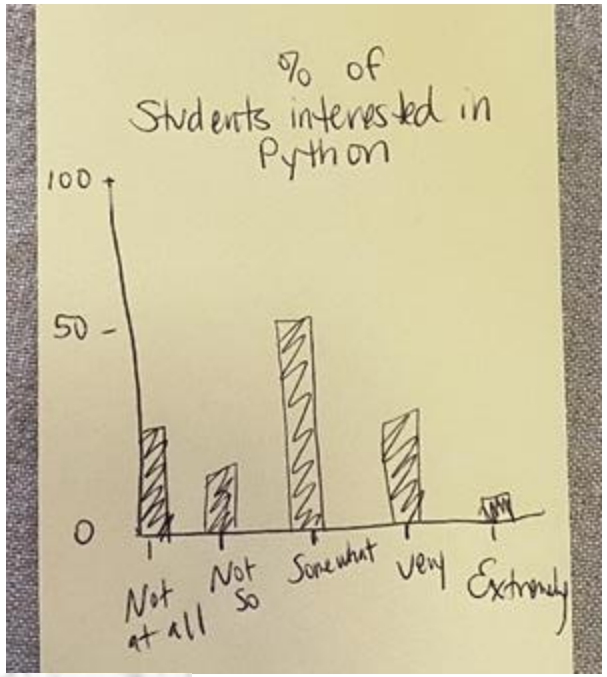
- Comparison
- Composition
- Distribution
- Relationship

Examples

- *"I want to know if my classmates that own dogs got up earlier today. I created my visualization using PowerPoint."*



- *"I want to show my bosses that we should offer a Python programming class. I mocked this up on a scrap of paper."*

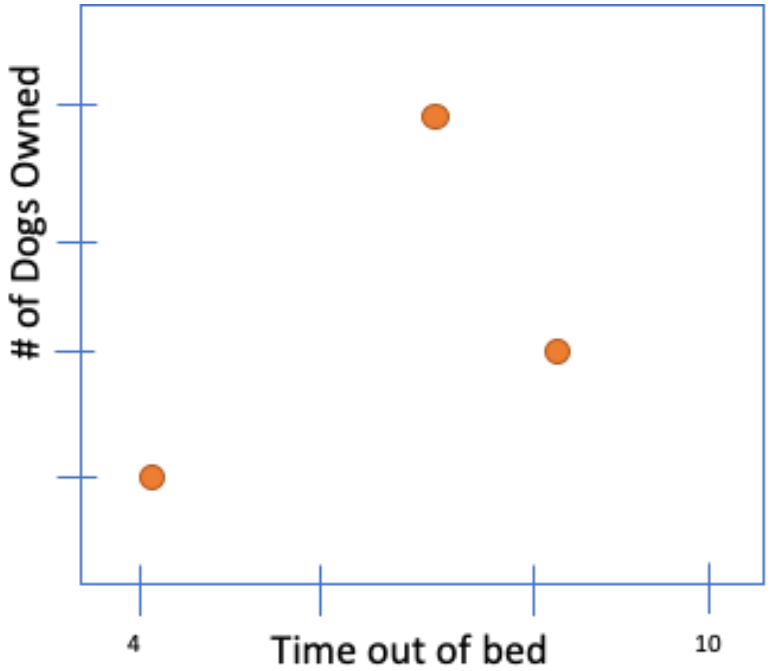


Activity: class survey

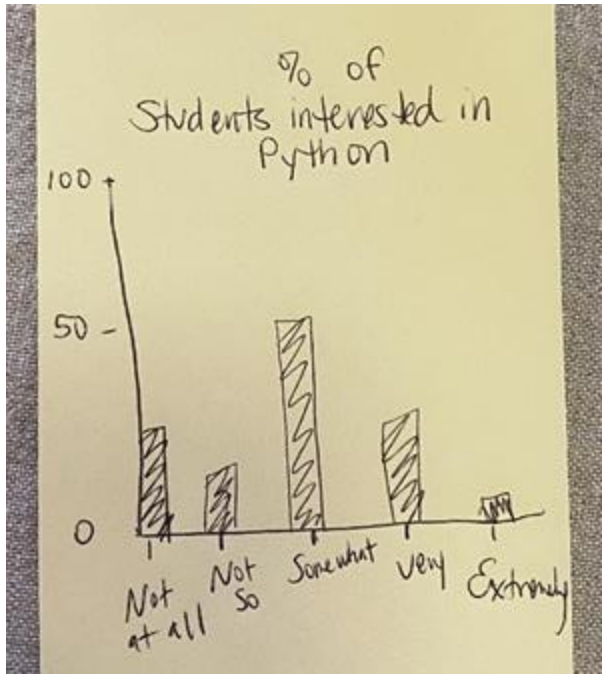
- Turn to page 14 of your participant guide to find **class survey – part II**.
- Review the questions from the class survey that you all took earlier.
- Think about what story you could uncover or tell with the results and who you want to tell it to.
- Draw a quick visualization to express how you'd like to see that story presented.

Wrap up

- *"I want to know if my classmates that own dogs got up earlier today. I created my visualization using PowerPoint."*



- *"I want to show my bosses that we should offer a Python programming class. I mocked this up on a scrap of paper."*



Recap

- The type of visual you use depends primarily on two things:

    1. The data you want to communicate

    1. What you want to convey about that data

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

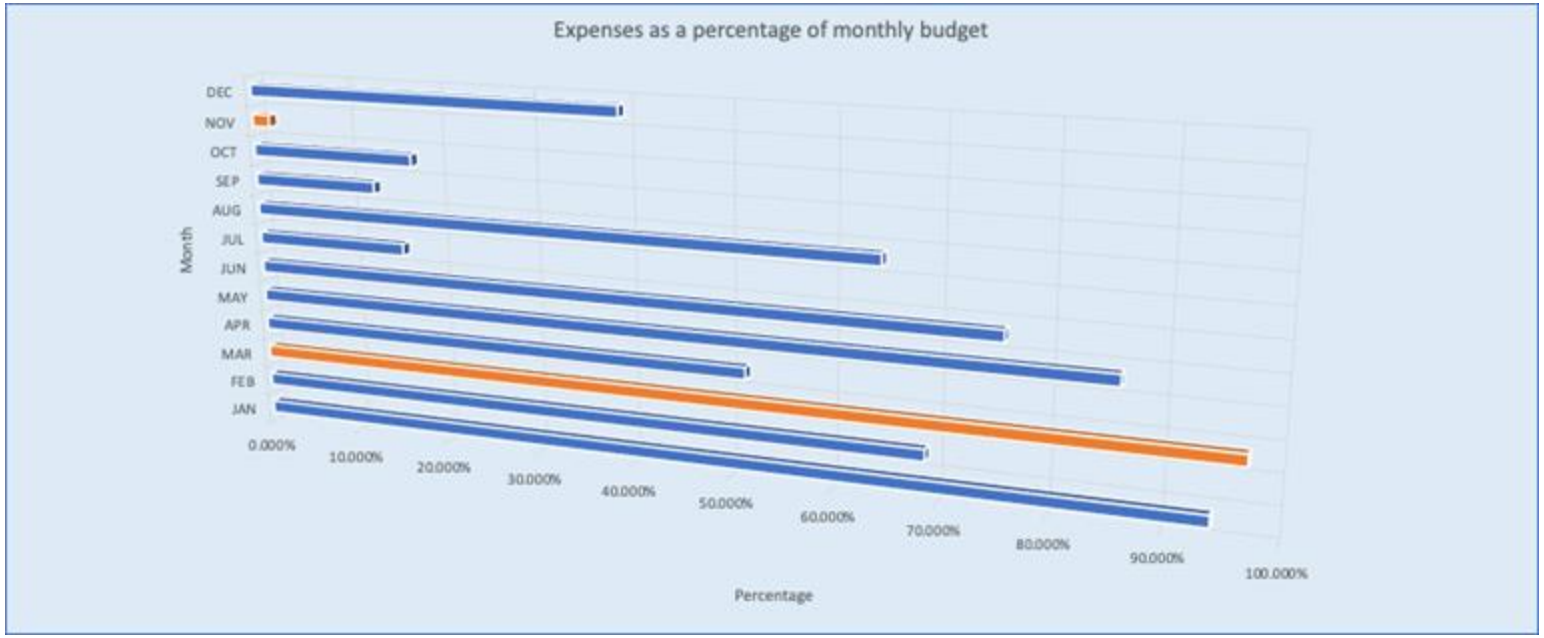
- Data storytelling

Designing compelling visuals

- Picking the right chart type isn't enough.
- There are choices to be made about the elements you include and how they are formatted.
- Data visualization is an art, informed by science.
- We'll start by discussing how to reduce chart clutter.
- In later sections we'll discuss visual design theory and common mistakes.

Reducing chart clutter

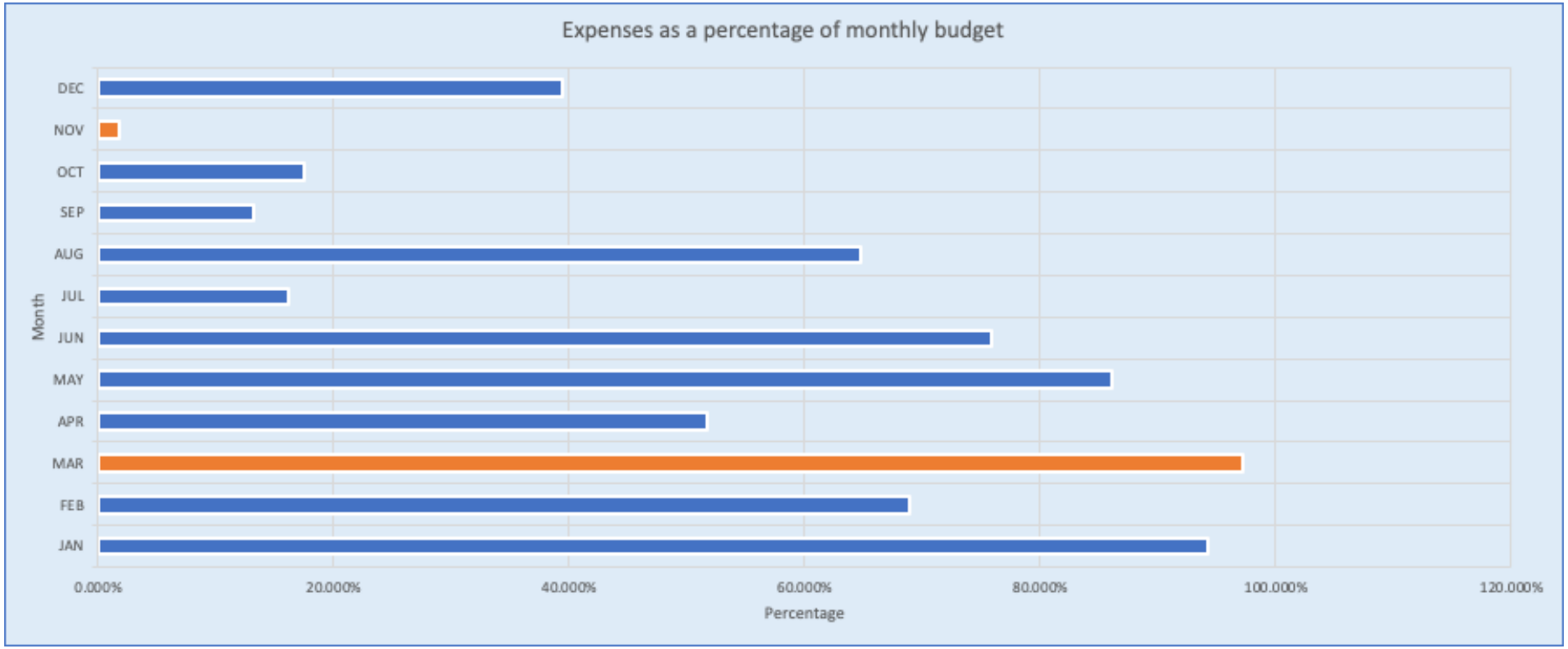
- Small changes can have a big effect on a visualization's impact.
- Let's walk through some tips to make an ugly chart like this look better.



Tip 1

Remove special effects

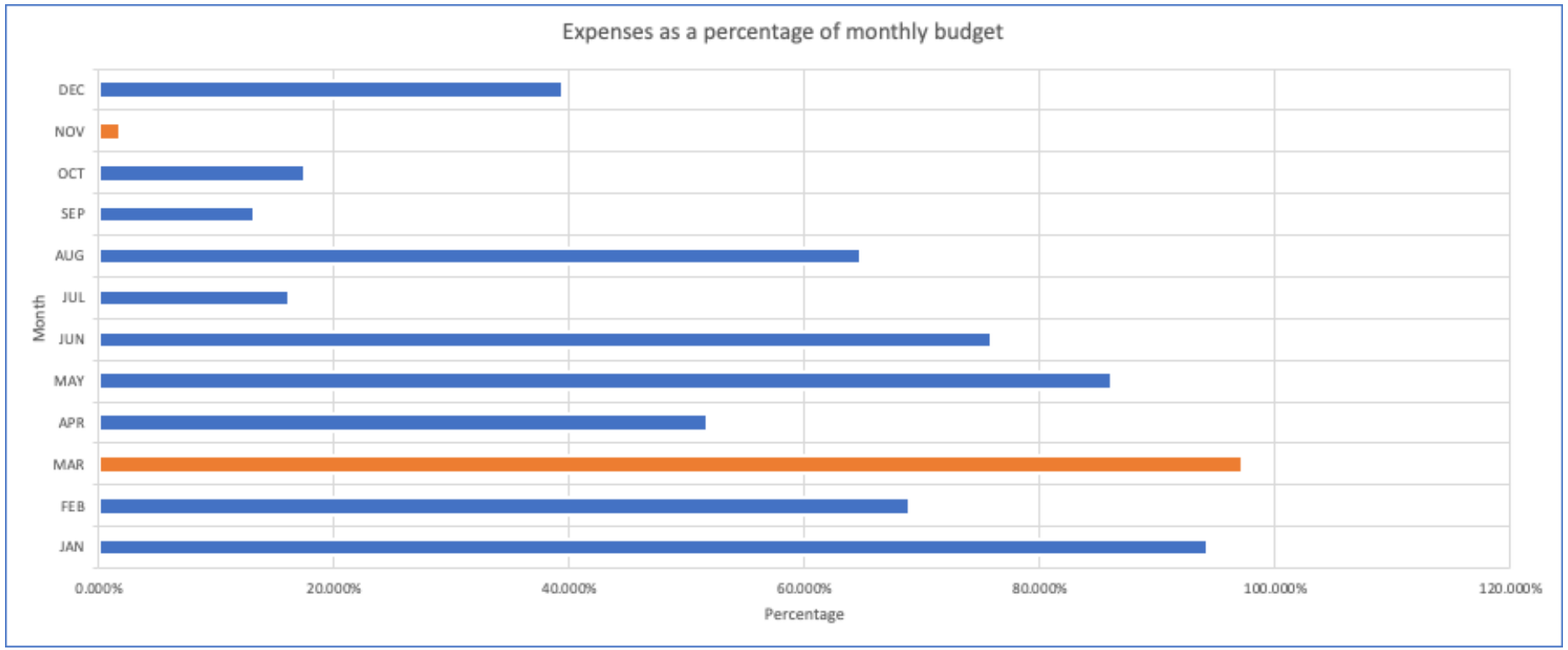
- Special effects, like 3D presentation, can reduce comprehension and make it difficult to compare elements and judge areas.



Tip 2

Lighten the background

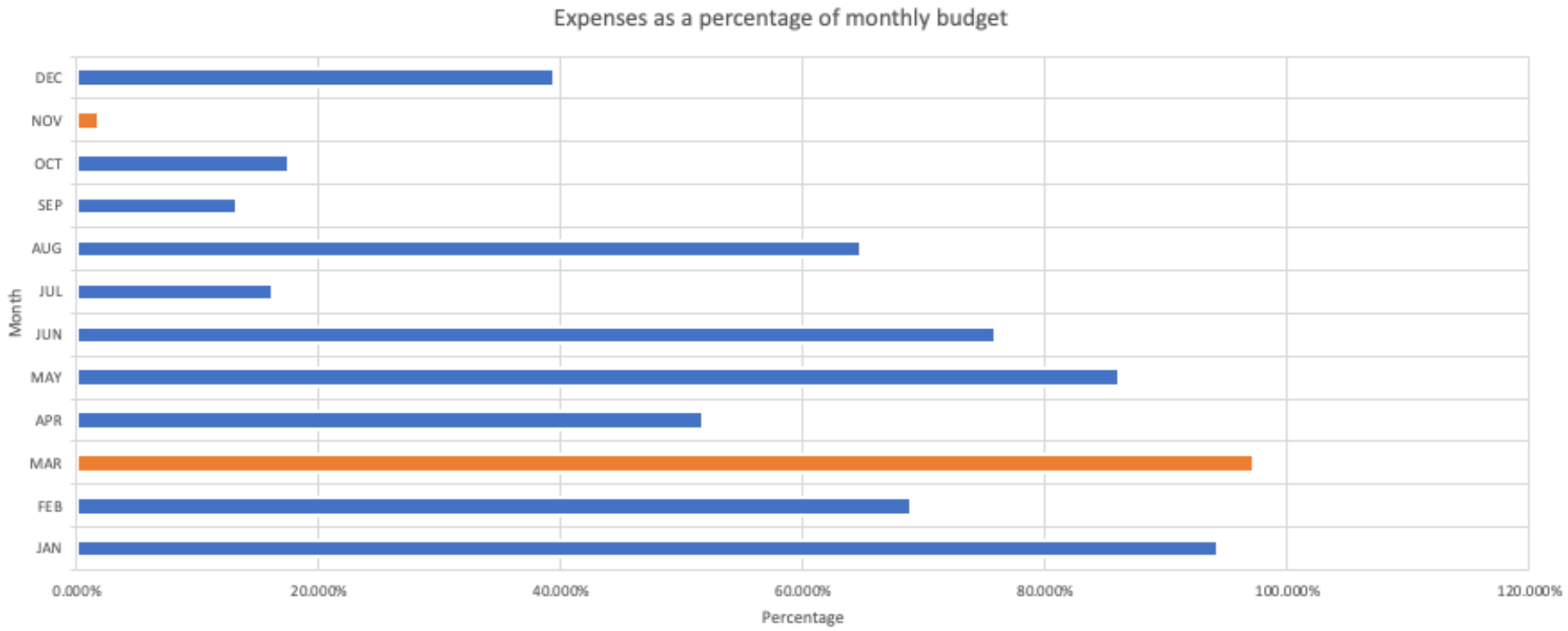
- A dark background can distract from the main message.



Tip 3

Remove chart borders

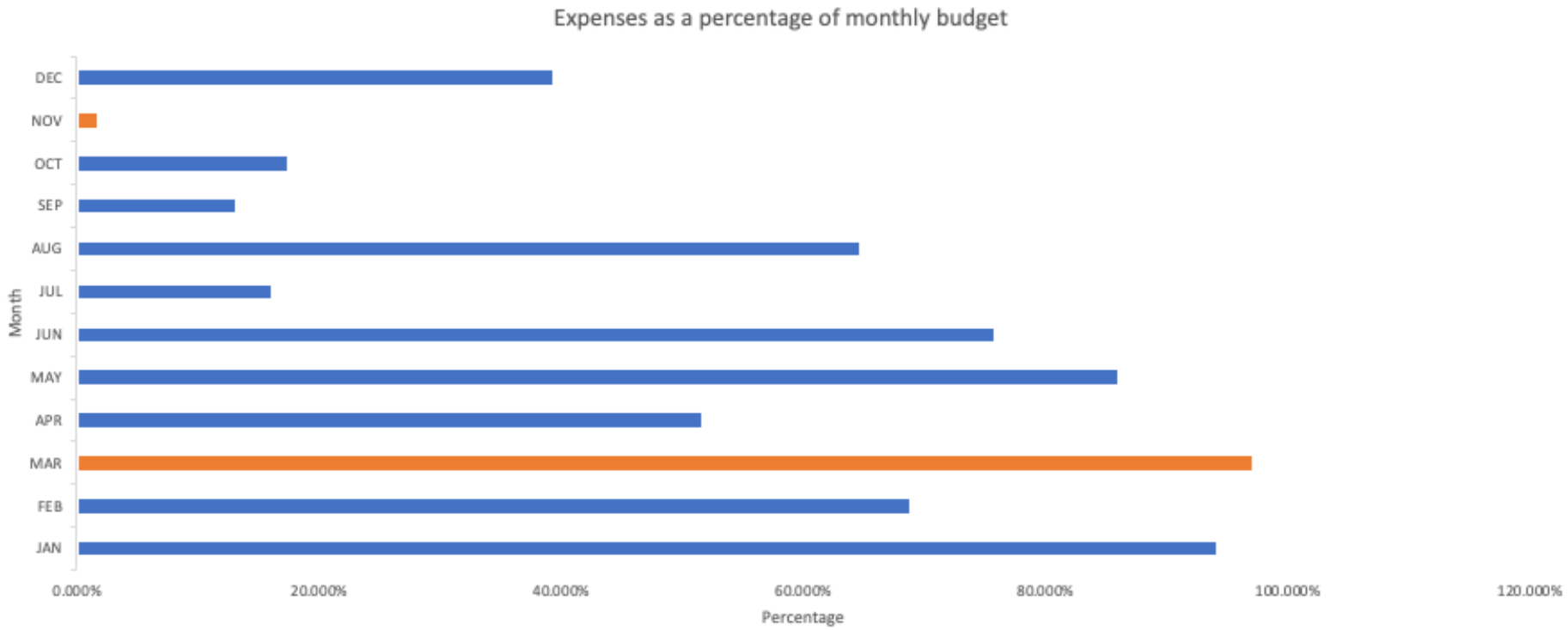
- Simplifying chart elements like borders can help to highlight what is most important, relevant, or interesting.



Tip 4

Remove gridlines

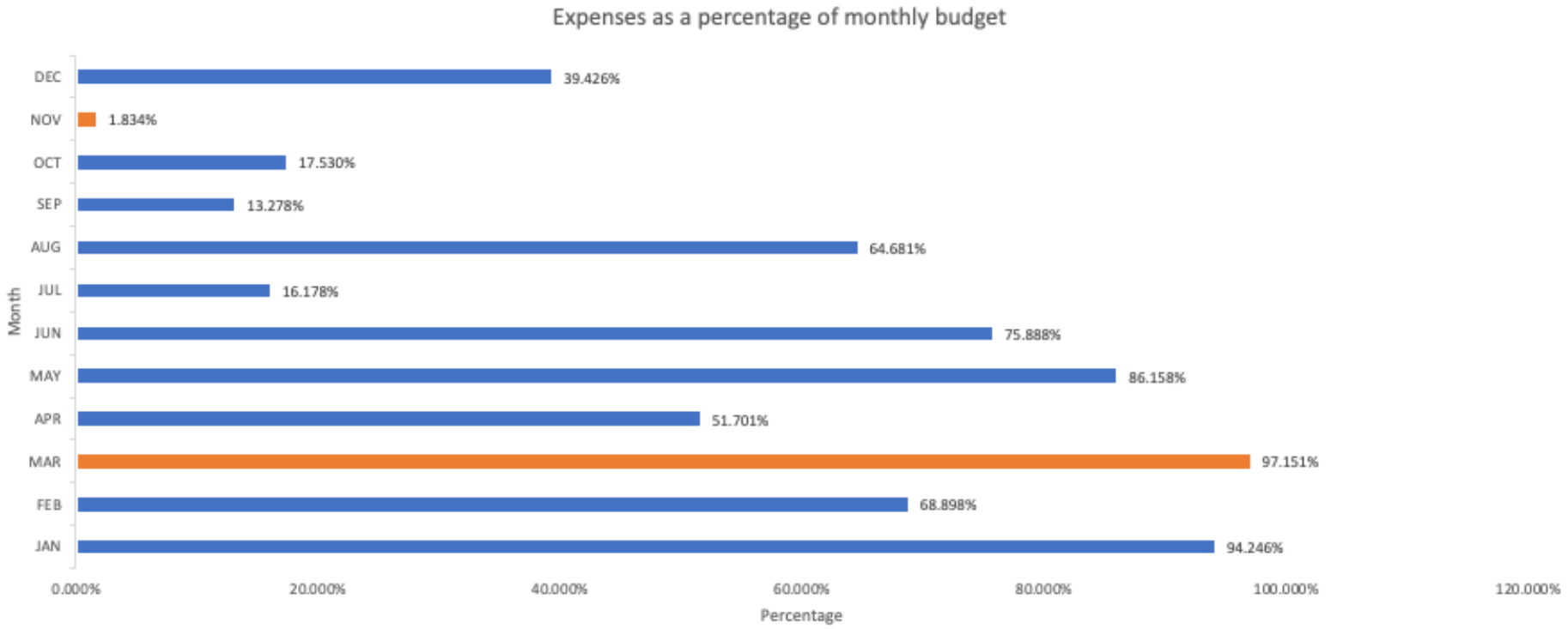
- Gridlines are often unnecessary and distracting.



Tip 5

Direct label

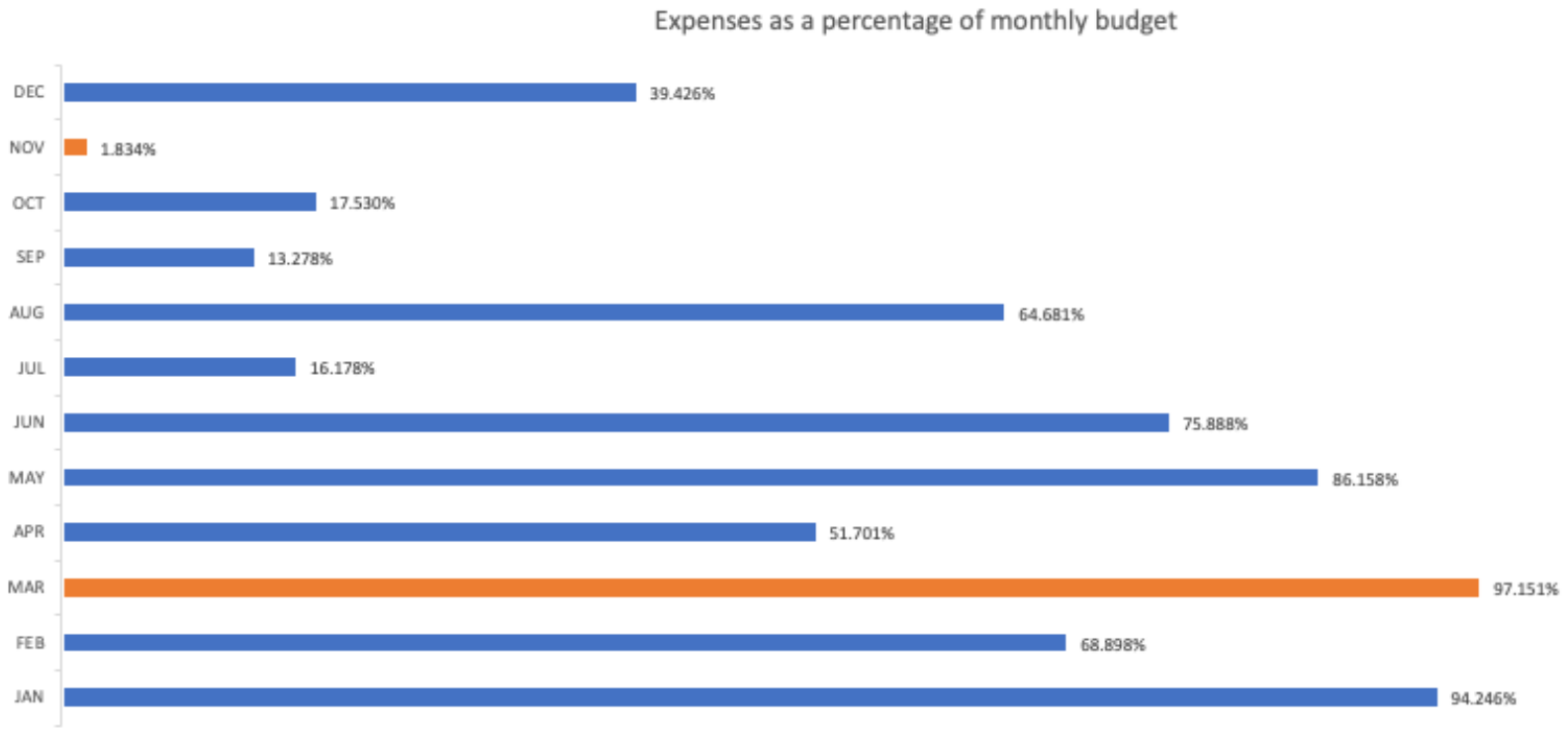
- Your eyes will need to move around the page much less when you can connect the label directly to the data.



Tip 6

Clean up axis titles and labels

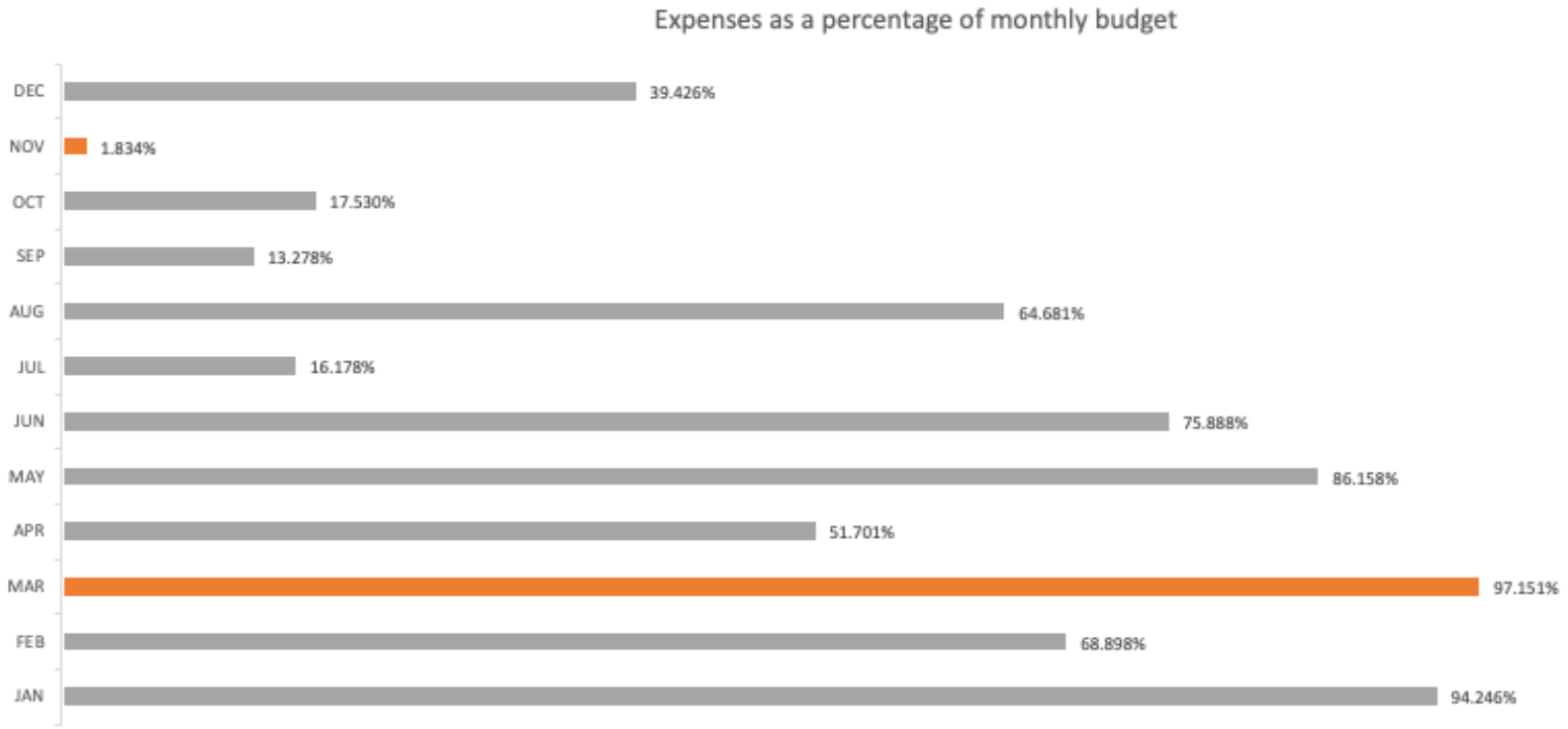
- In this chart the axis titles and some labels were redundant. Eliminating them simplifies the graphic and makes it easier for your reader to process.



Tip 7

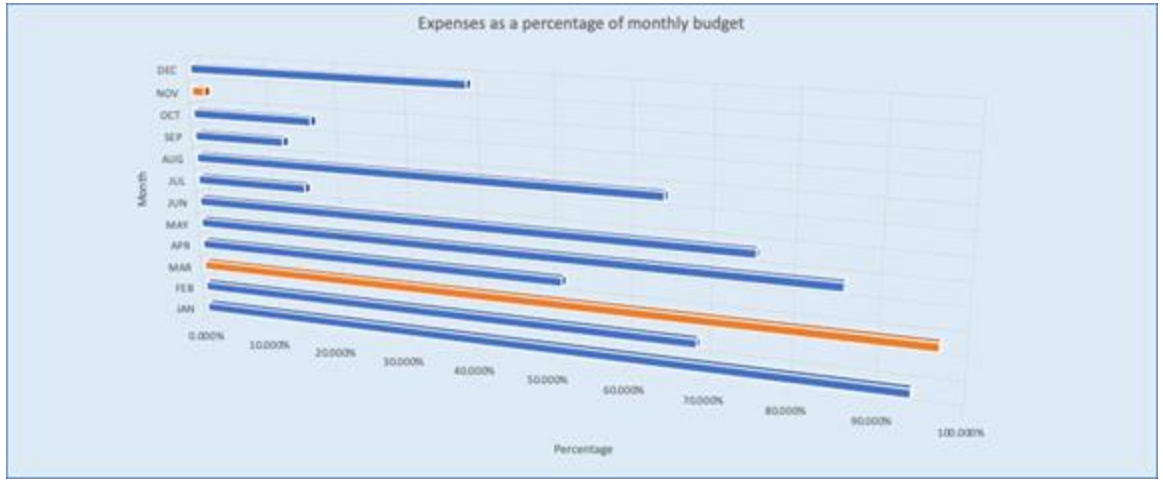
Use consistent colors

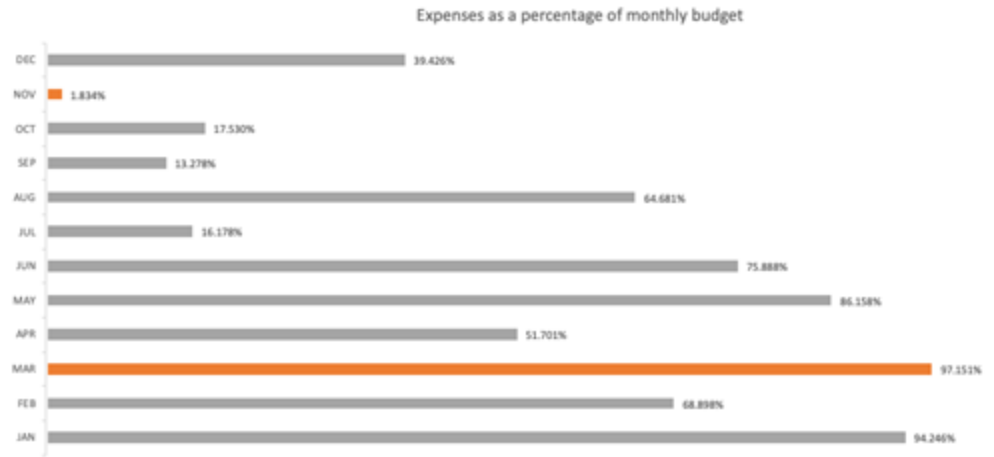
- Using less color also makes charts less complicated and therefore easier to read.
- We selected orange to blend with the look and feel of the slide deck.



Recap

1. Remove special effects
2. Lighten the background
3. Remove chart borders
4. Remove gridlines
5. Direct label
6. Clean up axis titles and labels
7. Use consistent colors





Q&A

DATA SOCIETY®

Welcome back
Day 3

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Visual design theory

- Our eyes “load” information while the brain “processes” it.
- We give the most attention to what looks good and struggle when our working memory is overwhelmed.
- For information to be effective, it should not provide more data than what the human brain can process.

Example: buying oranges

- You want to buy oranges at a new supermarket.
- Our eyes scan the layout of the supermarket, while the brain processes the various sections.
- The brain then instructs the eyes to zone in on the fruit section by sending signals about how fruits look from memory.
- The eyes then break the entire scanned area into parts and scan each part to spot the fruit section.
- The process is repeated until oranges are located.

Designing compelling visuals

- Our eyes and brains work the same way with data visualizations as they did in the oranges example.
  - Use visual clues to make data visualizations easier for the audience.
- However, every piece of information in a visualization also creates cognitive load on the viewer, asking them to use their brain power to process it.
  - Reduce visual clutter to lower the cognitive load and help transmission of the message.

Theory

- The visual design tips we'll review today draw on theory such as:
  - the building blocks of visual design described by the Interaction Design Foundation
  - the four categories of preattentive visual attributes described in Colin Ware's book, *Information Visualization: Perception for Design*
  - the Gestalt Principles of visual perception, which describe how people group similar elements, recognize patterns, and simplify complex images when we perceive objects

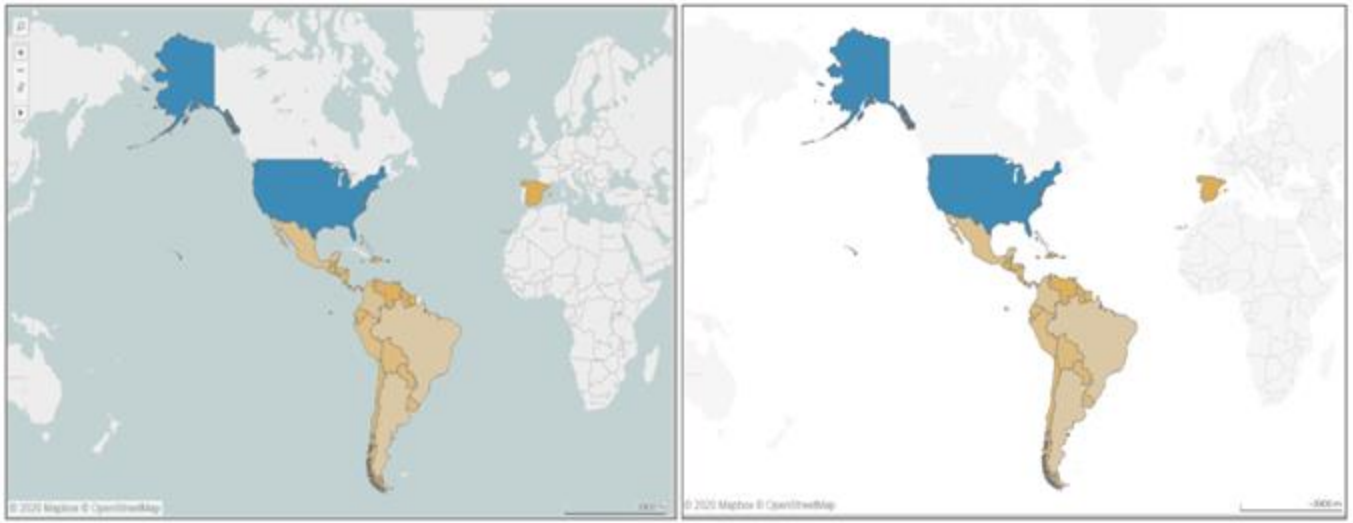
See page 32 of the participant guide for a list of additional resources

Tip 1

Choose a contrasting background

- People instinctively perceive objects as either being in the foreground (important) or the background (less important).

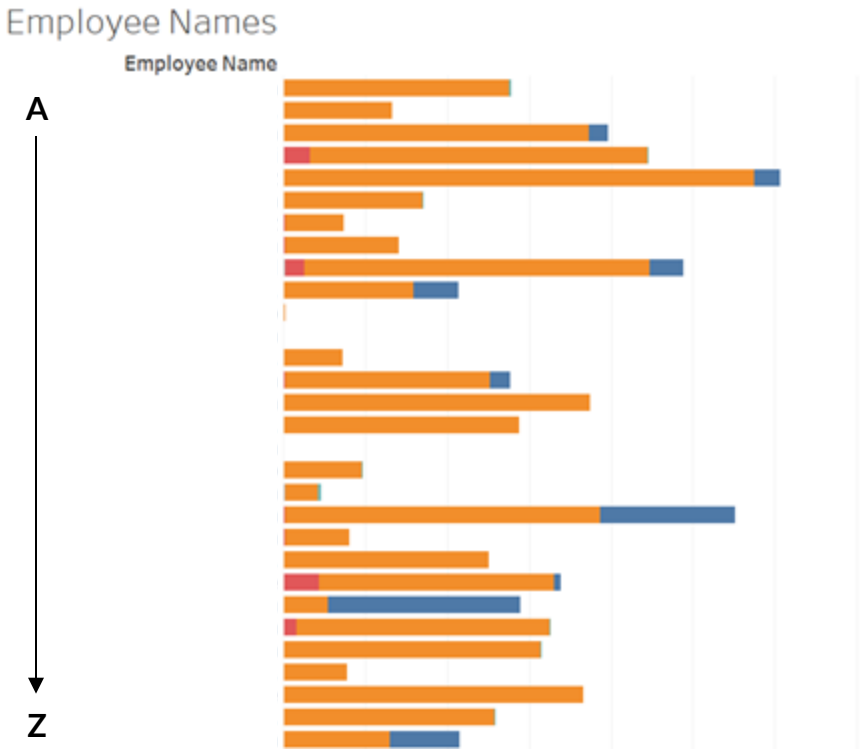
- Therefore, ensure that there is enough contrast between the figure and background to reduce cognitive load and increase readability.

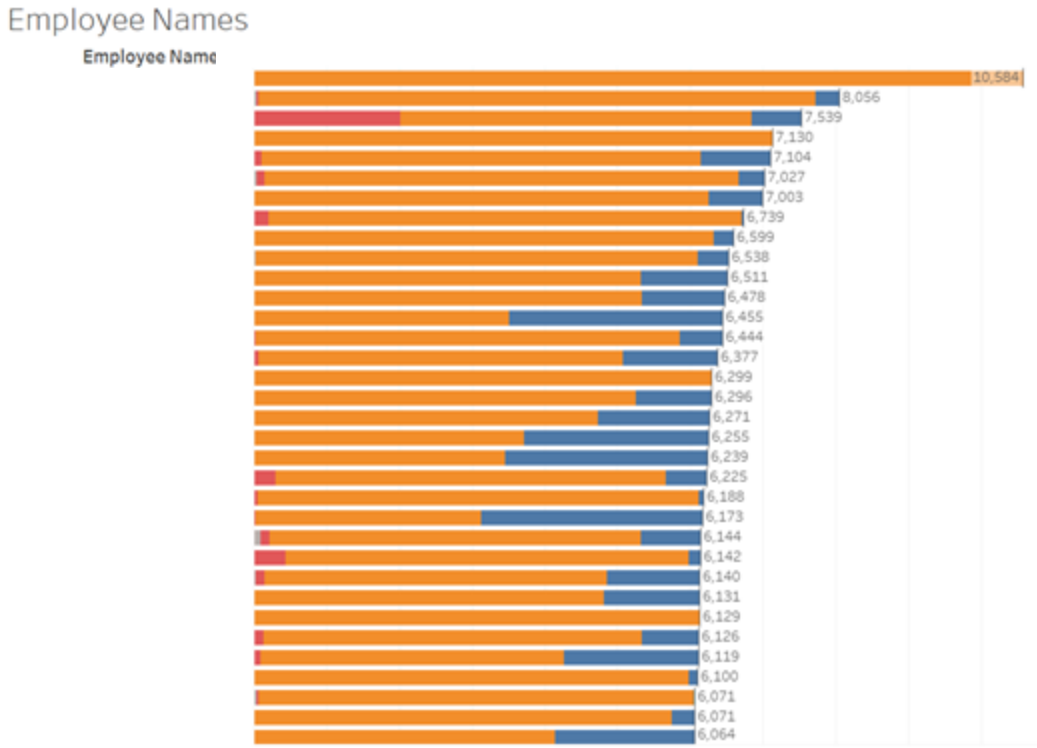


Tip 2

Make position meaningful

- Data should be sorted and placed in the visual in a meaningful way.





The left chart is sorted alphabetically; the right by value.

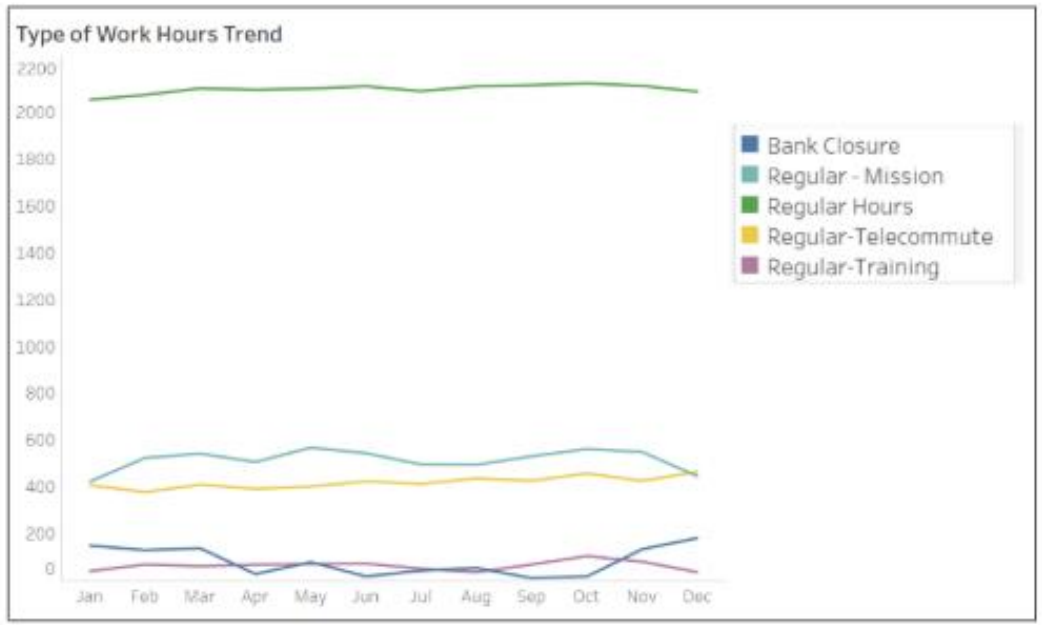
When would you use one over the other?

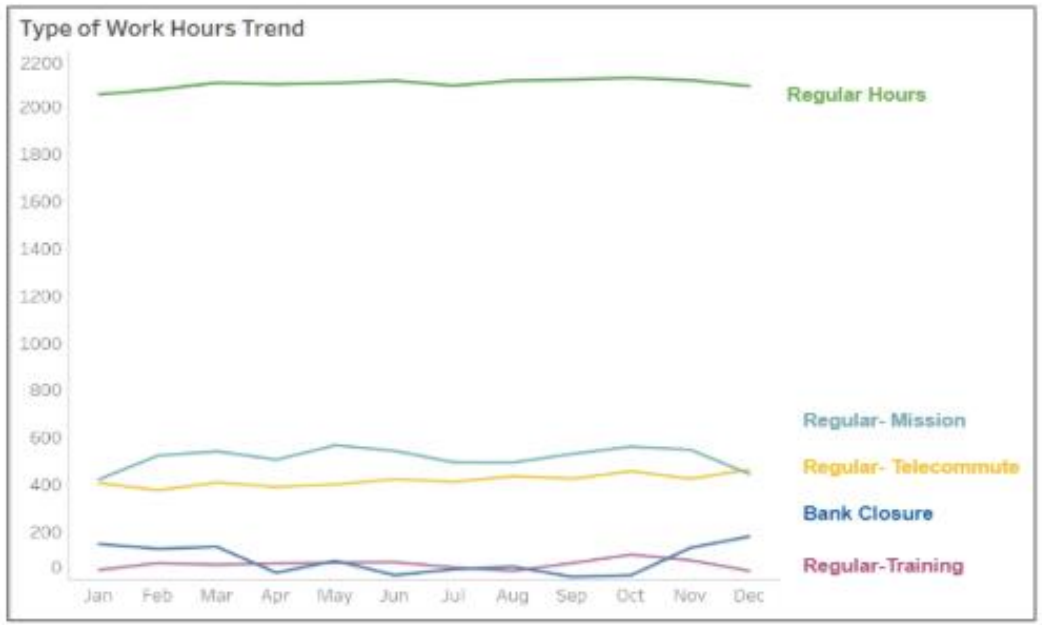
Tip 3

Group related items

- Things that are closer appear to be more related than those that are spaced farther apart.

- In fact, proximity overrules the similarity of other factors (e.g., shape, color).





Tip 4

Used closed regions to group

- Objects located within the same closed region are perceived as grouped together.

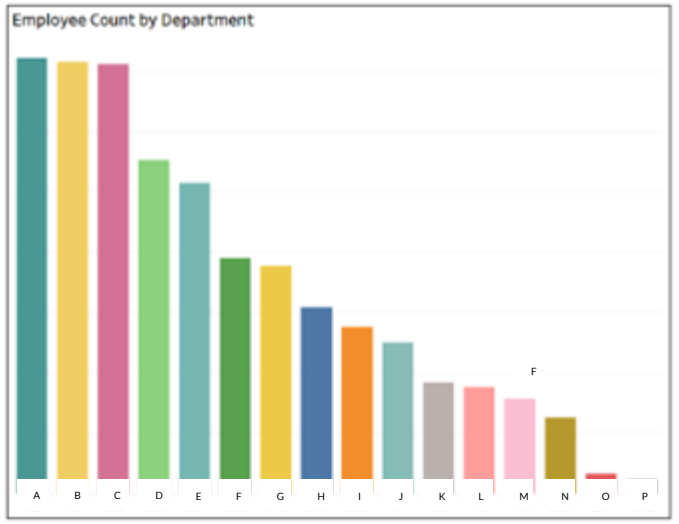
- We can use this principle to get the viewer to focus on a group of objects in the chart.

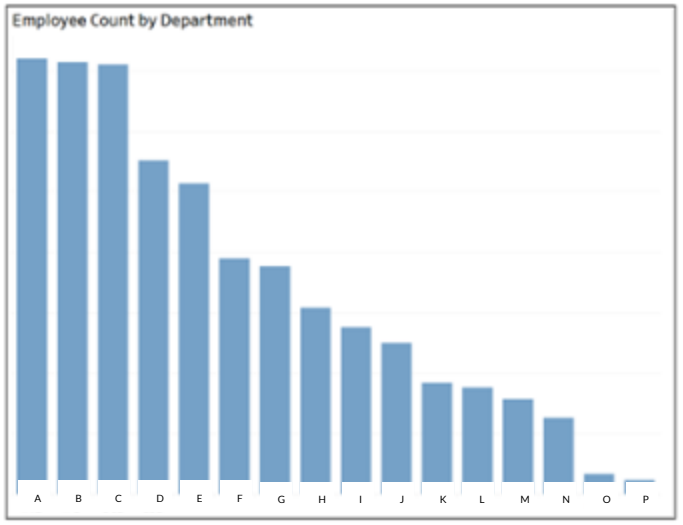
Tip 5

Distinguish different items

- The mind groups together things that look to be similar and assumes they have the same function.

- We can use this principle for:
  - distinguishing different sections
  - differentiating links from regular text
  - showing that elements with certain characteristics serve one purpose and others different





Tip 6

Arrange on a line

- Elements arranged on a line or curve are perceived to be more related than elements not on the line or curve.

Tip 7

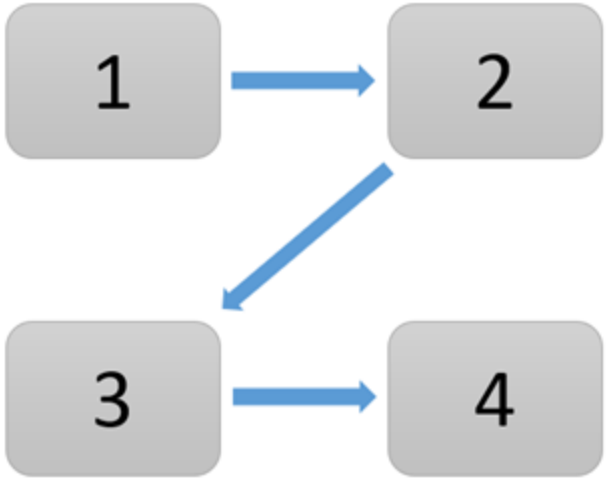
Address data gaps

- When parts of a whole are missing, our eyes fill in the gap as we look for a single, recognizable pattern.
- Therefore, gaps in data should be addressed.

Tip 8

Use natural positioning

- People usually tend to start at the top left of the visual and scan in **zig-zag** motions across the page forming a **Z-pattern**.
- Aim to position elements in a way that will feel natural for users to consume.
- Also, remember that the top of the page is the most precious.



Tip 9

Use labels and legends

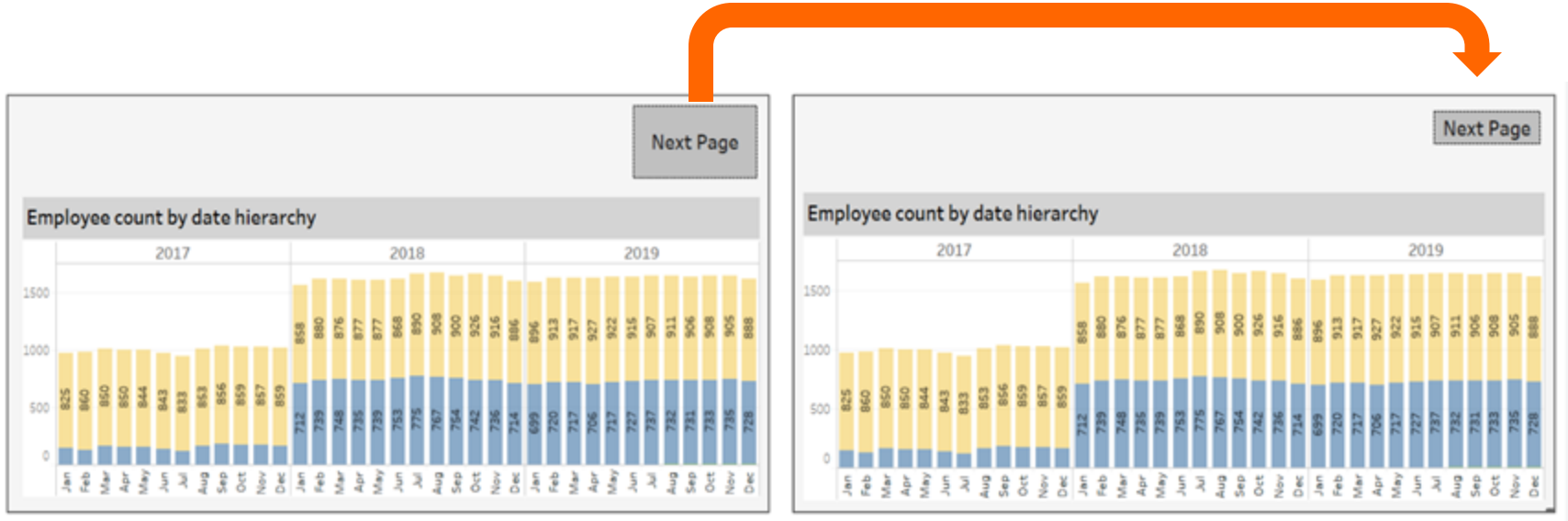
- Labels can be used to show value of datapoint.
- Legends can be used to identify the size, color or any other distinguishing feature in the visual.

The labels and legends used in the bottom chart makes it easier to understand.

Tip 10

Use size to show importance

- Relative size represents relative importance.
- Visuals of almost equal importance should be sized similarly.
- If there's one really important thing, it must be BIG.



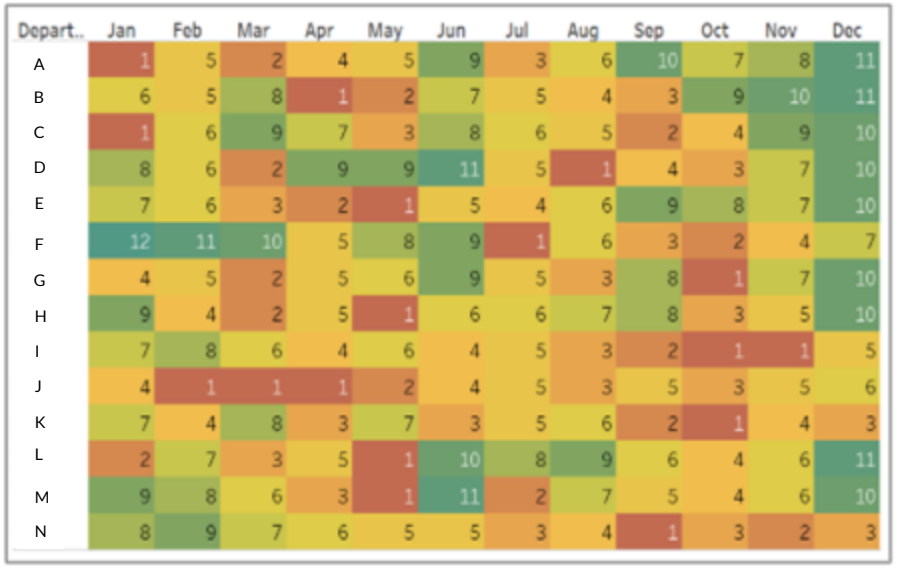
Resizing the "Next Page" button deemphasizes its importance.

Tip 11

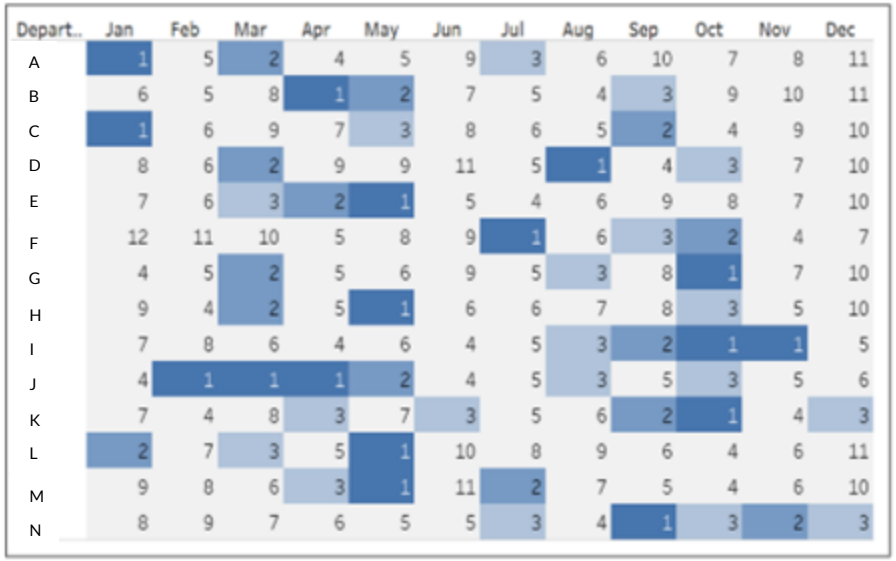
Use color to grab attention

- Color is another powerful tool used to draw the audience's attention
- However, the following must be kept in mind:
  - Use it sparingly: too much variety prevents anything from standing out
  - Use it consistently: a color change can be used to visually reinforce change in topic or tone

Too many colors are used in the image on the left, making it difficult to identify which are the busiest months.



Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1	5
J	4	1	1	1	2	4	5	3	5	3	5	6
K	7	4	8	3	7	3	5	6	2	1	4	3
L	2	7	3	5	1	10	8	9	6	4	6	11
M	9	8	6	3	1	11	2	7	5	4	6	10
N	8	9	7	6	5	5	3	4	1	3	2	3



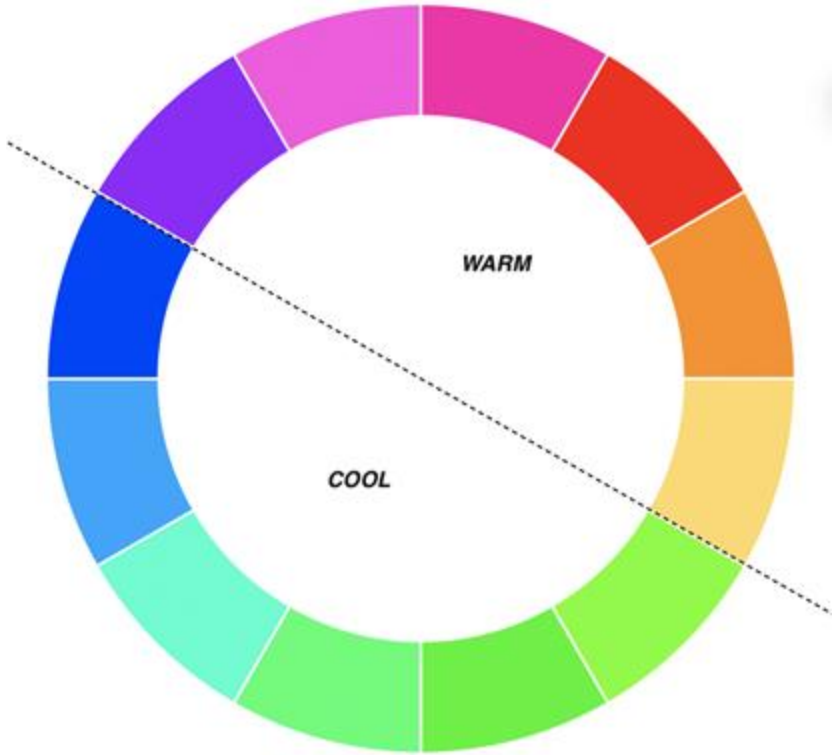
Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A	1	5	2	4	5	9	3	6	10	7	8	11
B	6	5	8	1	2	7	5	4	3	9	10	11
C	1	6	9	7	3	8	6	5	2	4	9	10
D	8	6	2	9	9	11	5	1	4	3	7	10
E	7	6	3	2	1	5	4	6	9	8	7	10
F	12	11	10	5	8	9	1	6	3	2	4	7
G	4	5	2	5	6	9	5	3	8	1	7	10
H	9	4	2	5	1	6	6	7	8	3	5	10
I	7	8	6	4	6	4	5	3	2	1	1	5
J	4	1	1	1	2	4	5	3	5	3	5	6
K	7	4	8	3	7	3	5	6	2	1	4	3
L	2	7	3	5	1	10	8	9	6	4	6	11
M	9	8	6	3	1	11	2	7	5	4	6	10
N	8	9	7	6	5	5	3	4	1	3	2	3

Tip 12

Use color to evoke emotion

- Color evokes emotion, so choose the one that helps reinforce the emotion you want to arouse in your audience.

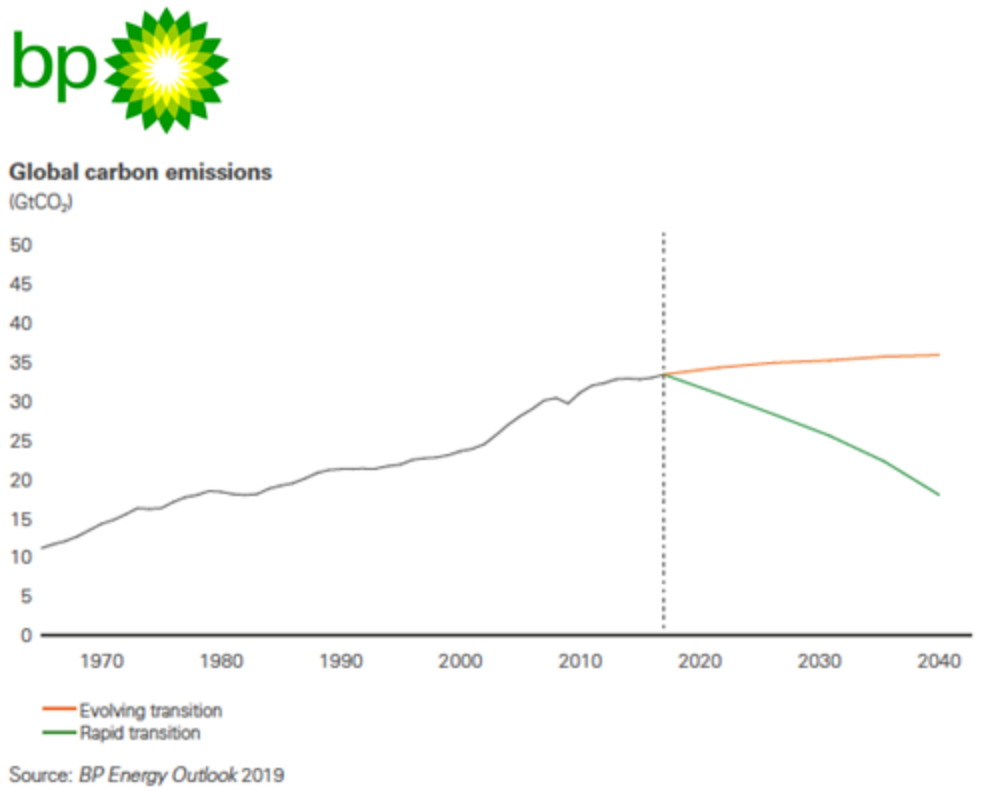
Warm colors	represent energy
Cool colors	represent calmness



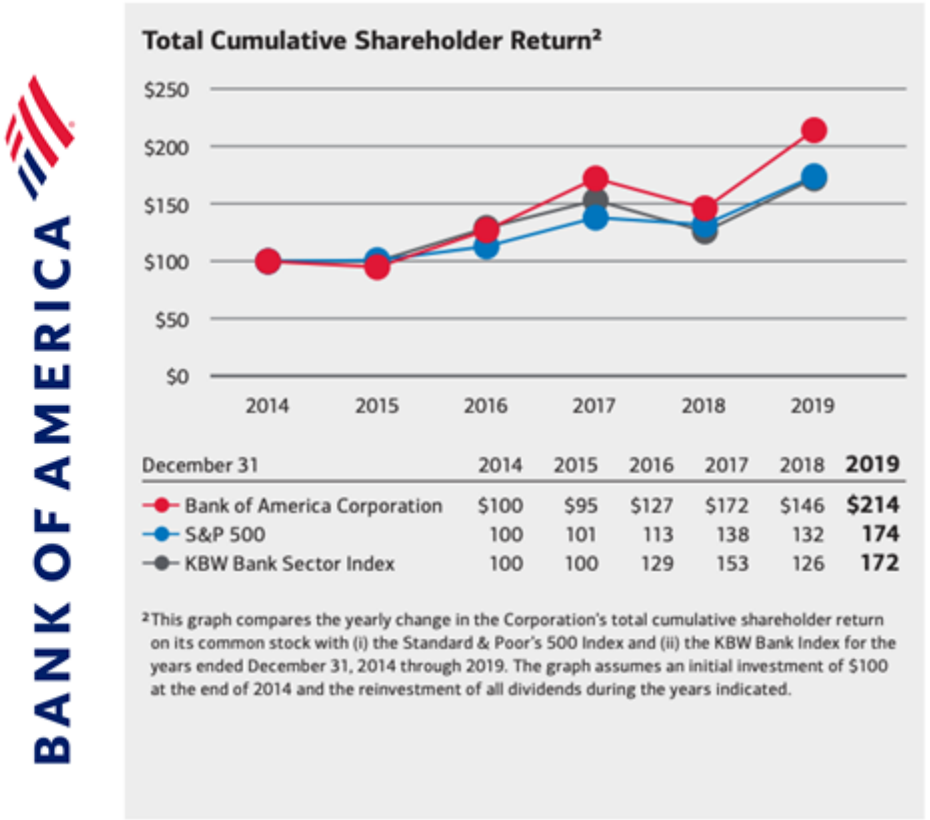
Tip 13

Use color to reinforce branding

- If it's necessary to use brand colors, select just one or two as "user-look-here" clues on an otherwise grey or black scheme.



Source: *BP Energy Outlook* 2019



Bank of America, 2019 Annual Report

Tip 14

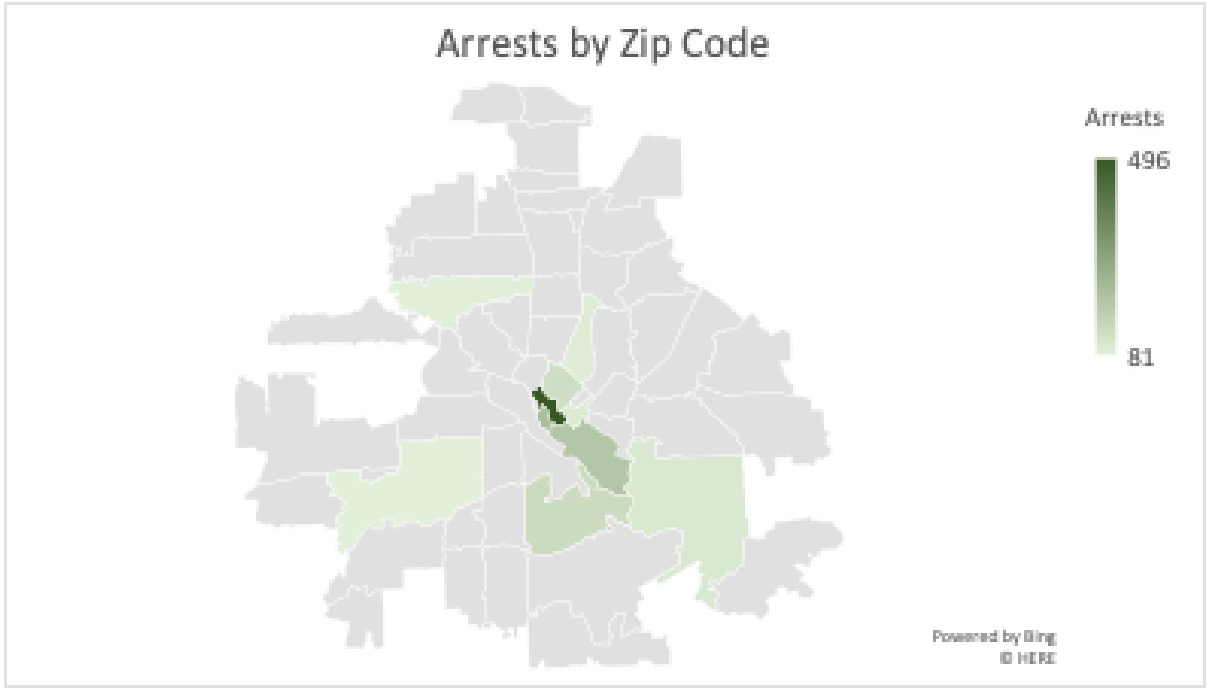
Encode data with color

- Use color schemes to encode data as sequential, diverging, or categorical.

Sequential	Diverging	Categorical
when the order matters	to highlight minimums, maximums, and midpoints	for discrete data values representing distinct categories

Sequential color schemes

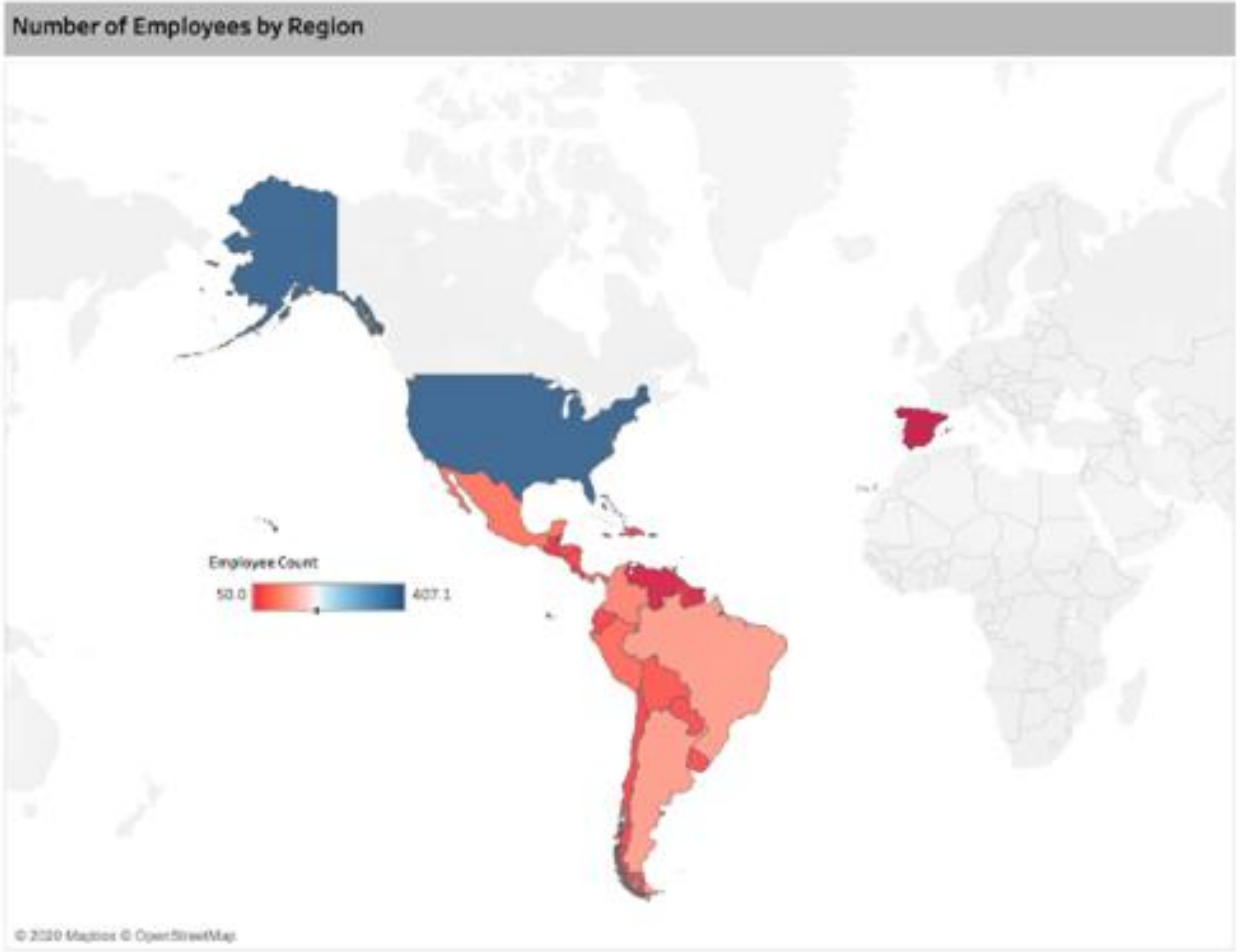
- Use a sequential color scheme when the order matters.
- These schemes range between two colors—usually a lighter shade to a darker one—by varying one or more parameters such as saturation.



Diverging color schemes

- Use a diverging color scheme to highlight minimums, maximums, and midpoints.

- These schemes range between three or more colors with the different colors being quite distinct—usually having different hues.



Categorical color schemes

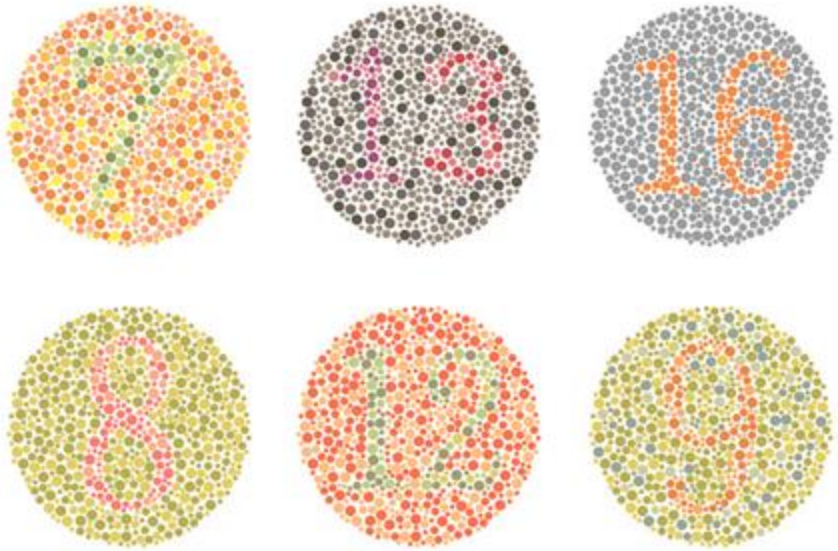
- Use a categorical color scheme for discrete data values representing distinct categories.
- These schemes use different hues with consistent steps in lightness and saturation.

Tip 15

Don’t forget color-blindness

- Color-blindness impacts roughly 8% of men and half a percent of women and results in difficulty distinguishing shades of red and green.

- Design with color-blindness in mind by varying boldness, saturation, or brightness to distinguish colors.



Recap

- These 15 tips will help make data visualizations easier for the audience to read.
- But remember, data visualization is an art. The guidelines may be broken intentionally to make a point!

1. Choose a contrasting background
2. Make position meaningful
3. Group related items
4. Used closed regions to group
5. Distinguish different items
6. Arrange on a line
7. Address data gaps
8. Use natural positioning
9. Use labels and legends
10. Use size to show importance
11. Use color grab attention
12. Use color to evoke emotion
13. Use color to reinforce branding
14. Encode data with color
15. Don't forget color-blindness

Break

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Common mistakes

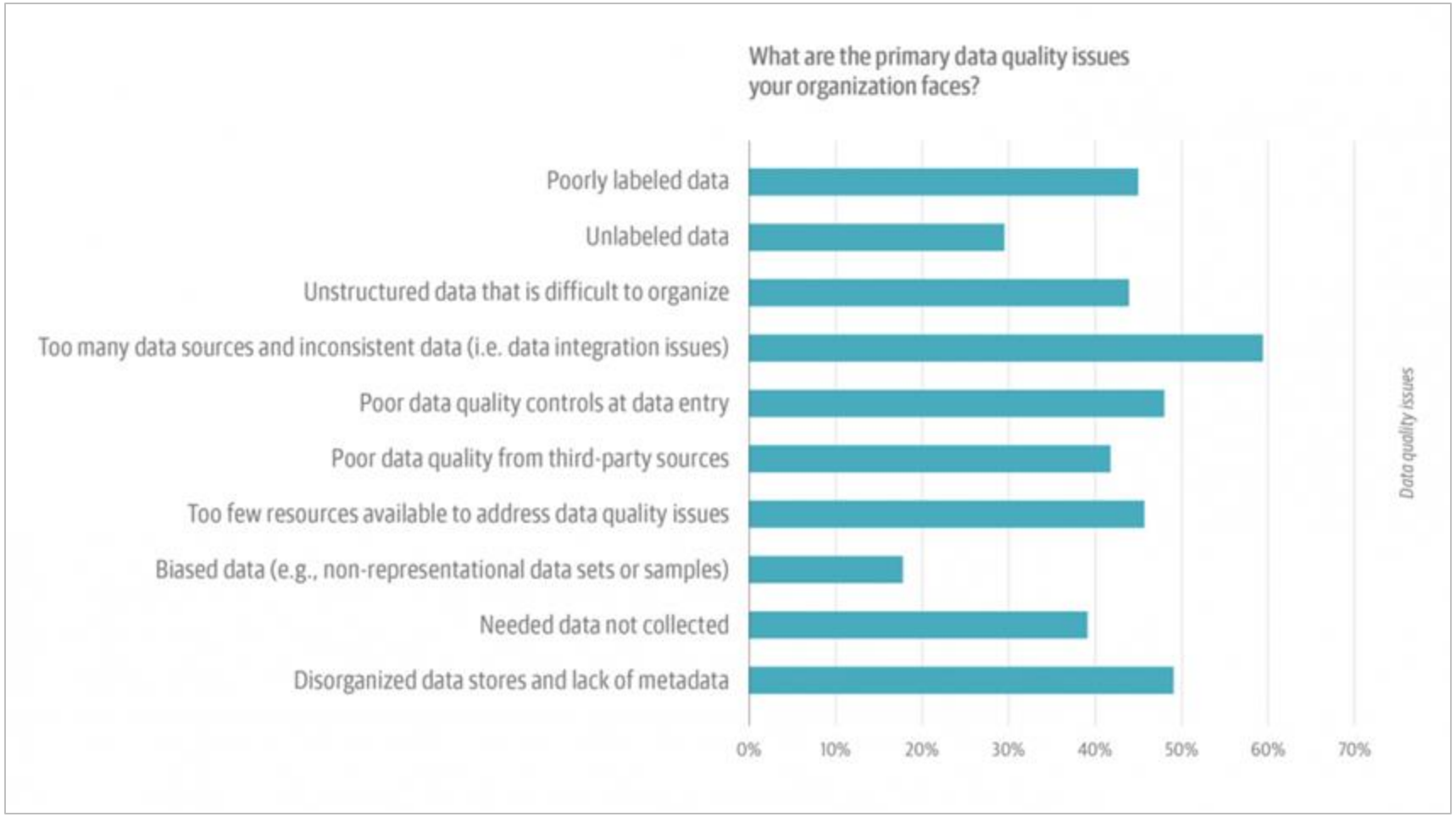
- Poor data visualizations confuse the viewer or, worse, mislead them and cause more harm than good.
- Poor visuals can cause a loss of time and effort and might delay the decision-making process in deadline-driven projects.
- Let's talk about a few common mistakes.

Visualizing bad data

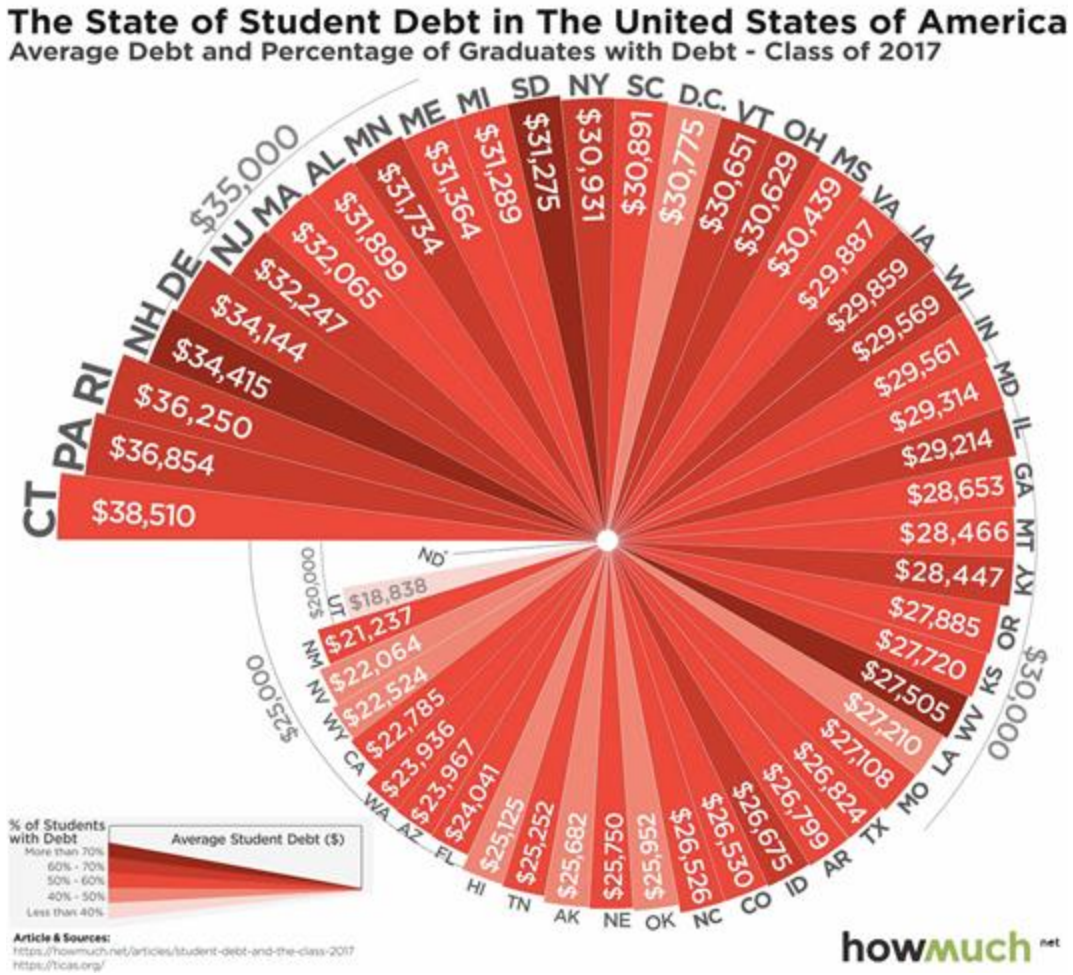
- Visualizations are only as good as the data they represent.
- Garbage in, garbage out!

2019 O'Reilly survey of more than 1,900 leaders and data professionals

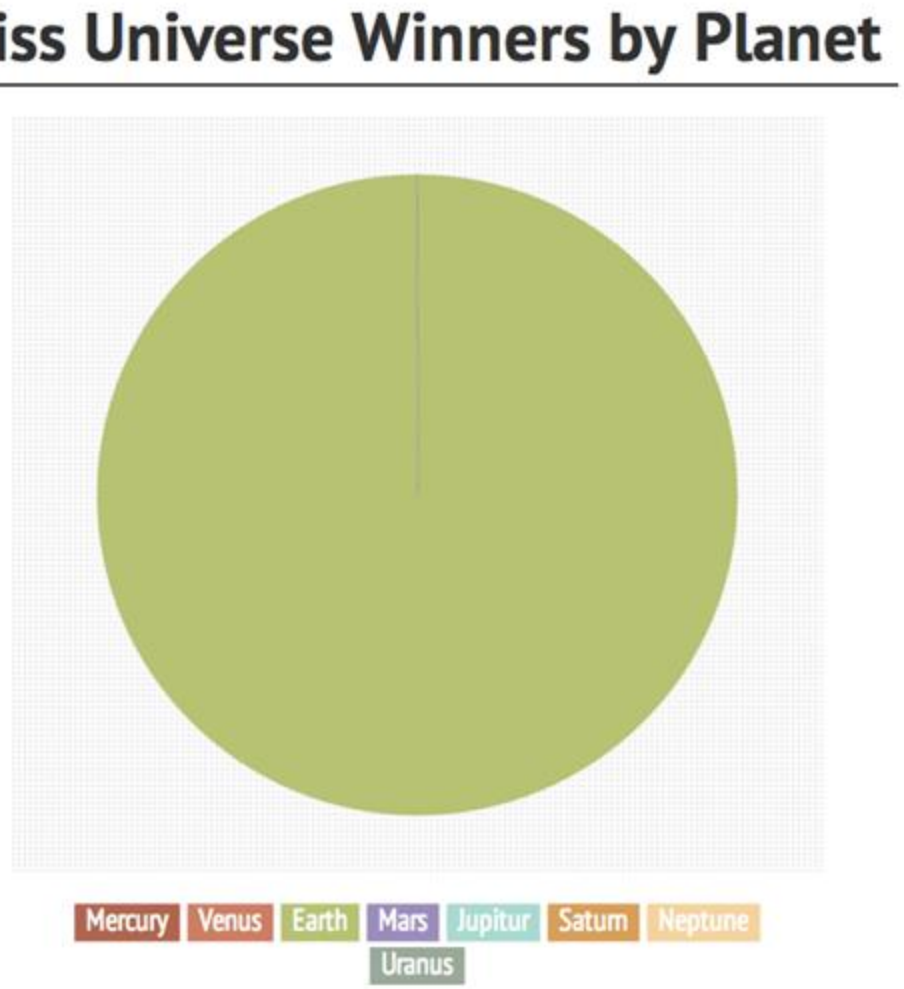
https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/



Misusing pie charts



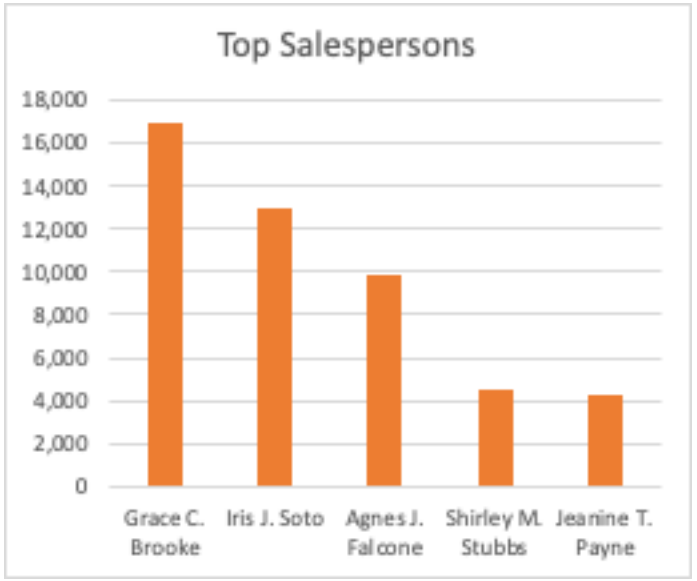
https://howmuch.net/articles/student-debt-and-the-class-2017

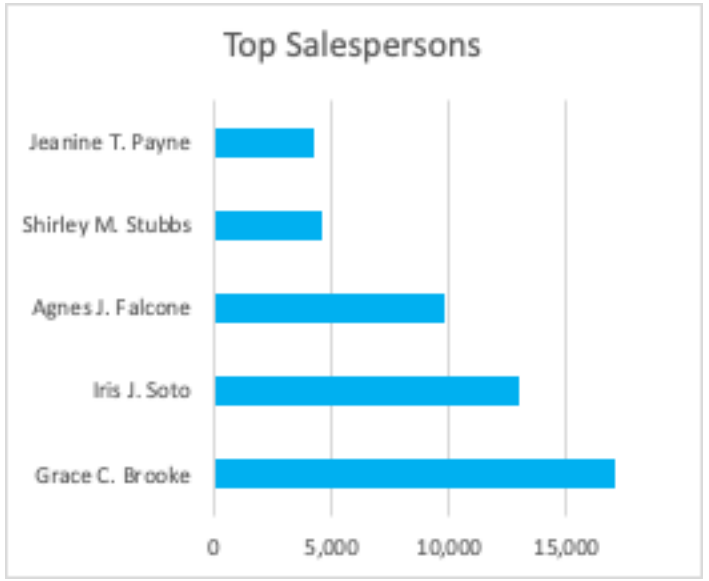


https://trackmaven.com/blog/4-best-graphs-marketing-data

Interchanging column and bar charts

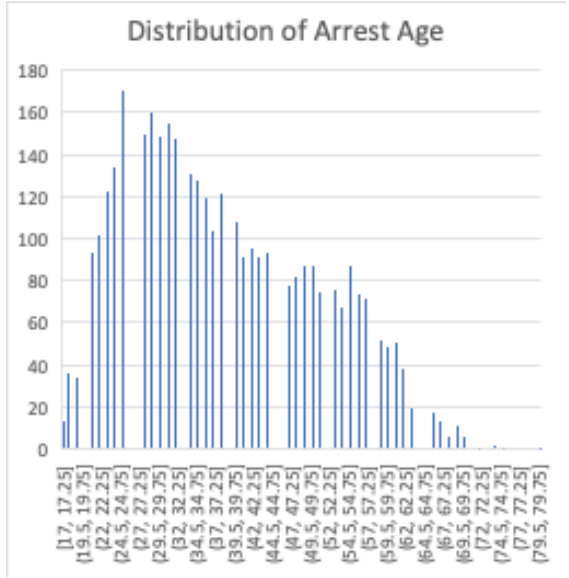
- We may think it doesn't matter if we use a column chart or a bar chart, but sometimes it does.
- Bar charts are better when your data labels are long, to reduce clutter.
- Column charts are better when you have negative values in your dataset.



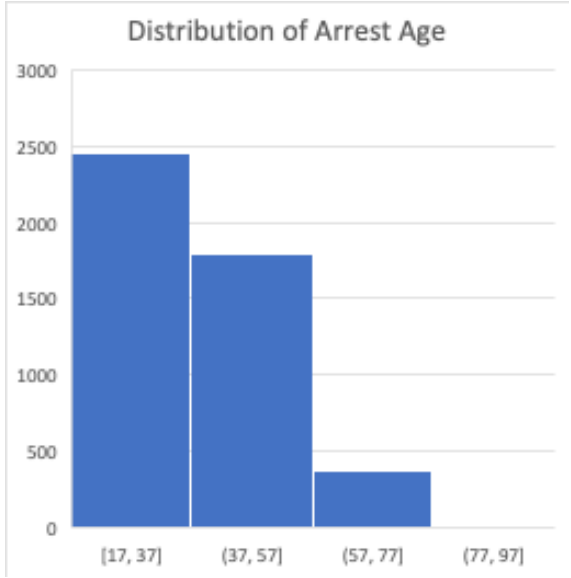


Using the wrong bin size in histograms

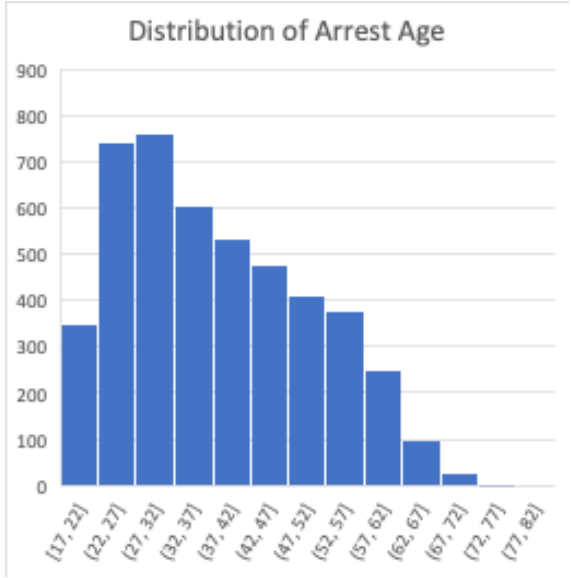
- There is no right or wrong answer as to how wide a bin should be, but you need to make sure that the bins are neither too small nor too large to show patterns.



bin size = 0.25 years



bin size = 20 years



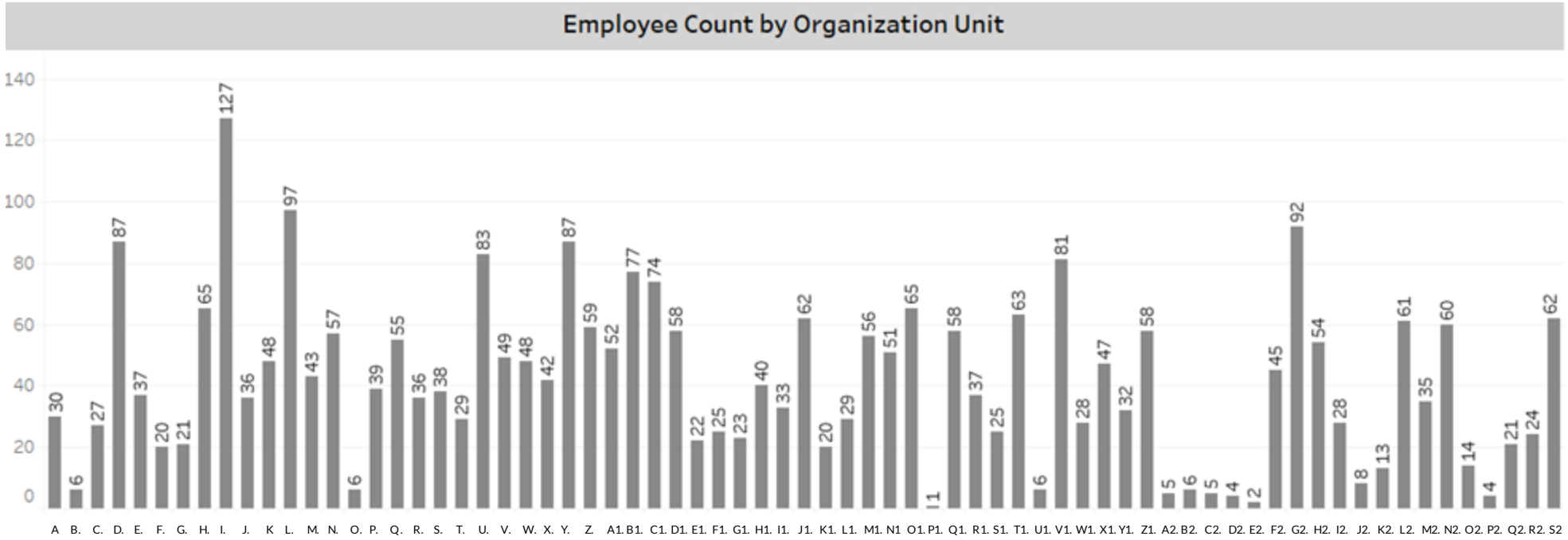
bin size = 5 years

Making difficult comparisons

- Sometimes we like using bright and colorful, graphic visuals to brighten up a mundane report. However, aesthetics should never supersede substance.
- A plain, less creative column chart would be better to represent the information below.

Plotting too much data

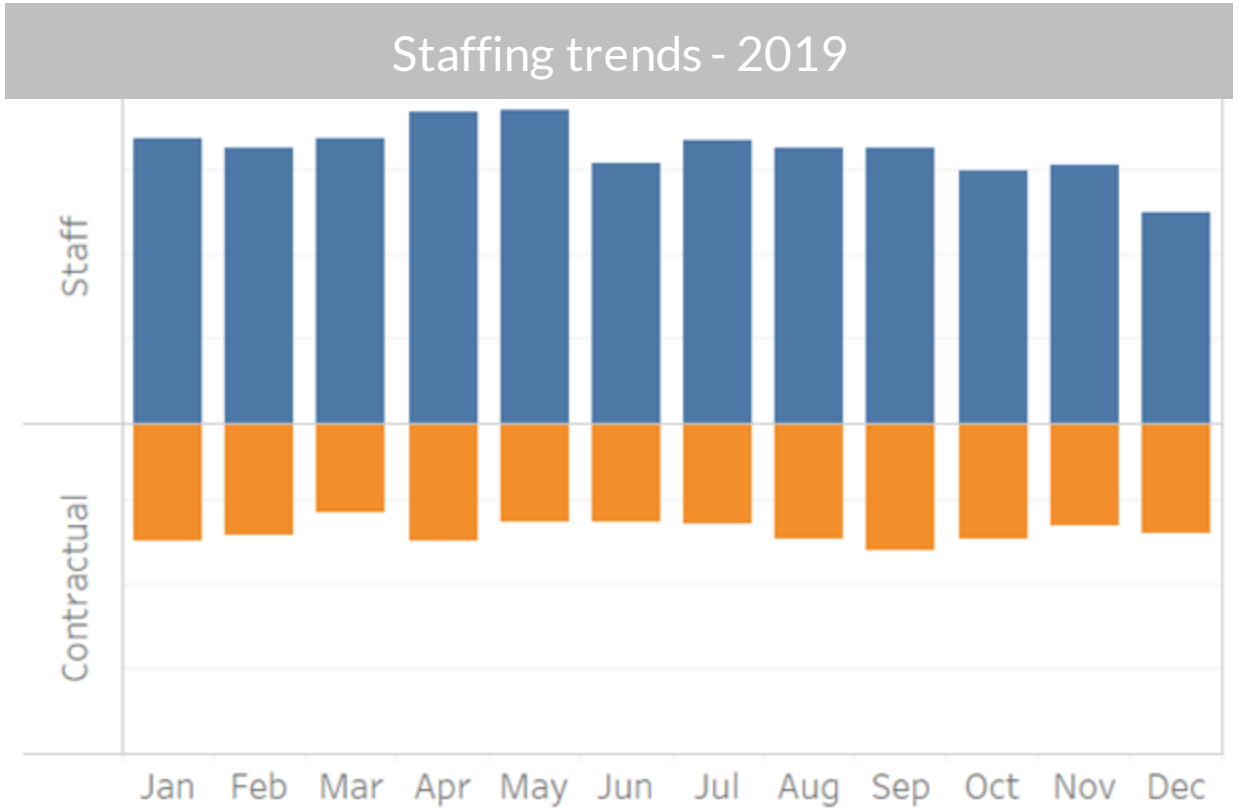
- When there are many data points, visualizing data for the sake of visualization is counterproductive and unhelpful to the viewer.
- The visualization below fails to properly compare employee count by location, and in cramming in too much information, is simply overwhelming.



Not following conventions

- Deviating from convention (such as green is positive and red is negative) can create confusion and misinterpretation of the facts.

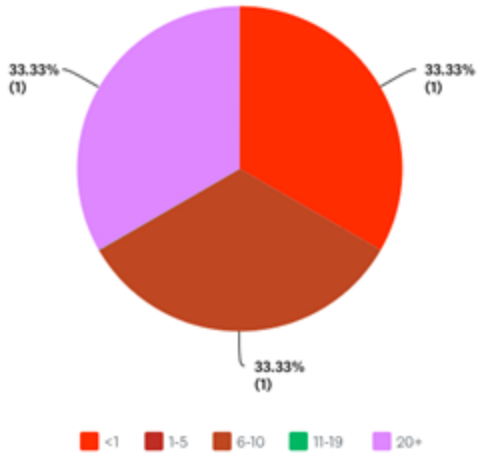
What convention does this chart fail to follow?



Activity: class survey

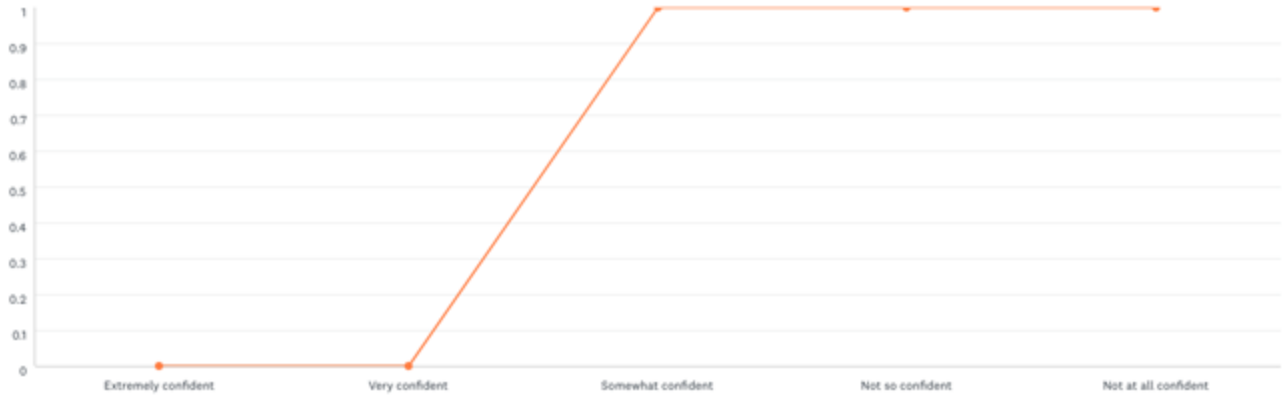
- Turn to page 16 of your participant guide to find the **class survey – part III** activity.
- Review the results of the class survey that you all took earlier at: https://www.surveymonkey.com/stories/SM-WY5Y5RS2/
- Jot down your notes about how you'd improve the visuals provided.

Visual 1



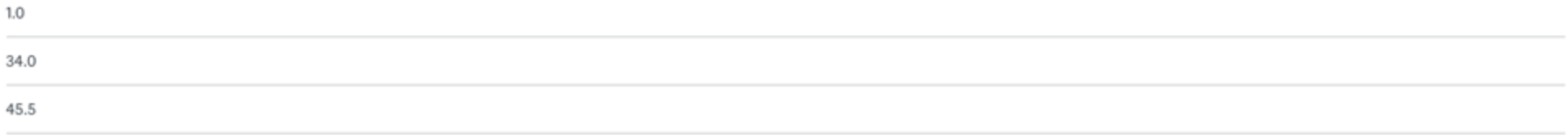
How many years have you worked for your current employer?	
Chart type	• Pie charts generally shouldn't have more than 3 segments; consider a different chart type (if applicable)
Visual design	• Colors are distracting; pick greyscale or branded colors
Clutter	• Consider direct labeling the categories
Other	

Visual 2



How confident are you in your ability to describe the difference between a histogram and a boxplot?	
Chart type	• Line charts are better for continuous data; it doesn't make sense to use a line chart with categorical data
Visual design	• Consider using data markers to reduce cognitive load
Clutter	• Consider if the gridlines are necessary
Other	

Visual 3

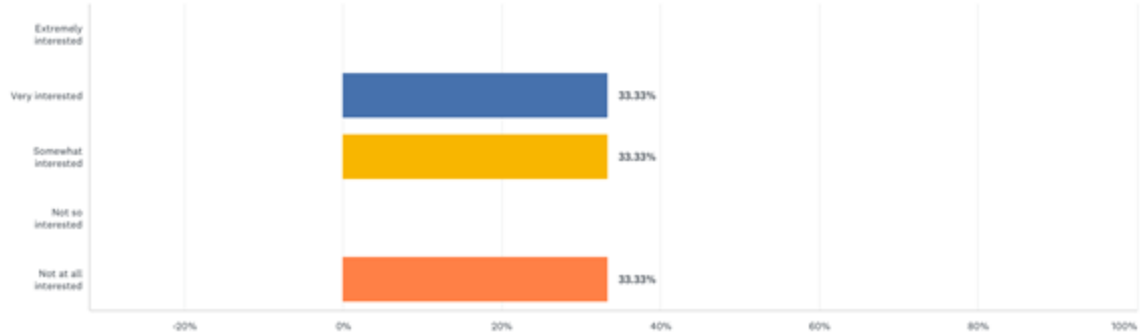


Roughly what percentage of your time at work is spent collecting, analyzing, or visualizing data?	
Chart type	• A table doesn't help you understand the data; should be visualized.
Visual design	
Clutter	
Other	• Data collection method may make visualization difficult

Visual 4

When was the last time you created a chart or graph to present data?	
Chart type	• An area chart doesn't make sense for this type of data
Visual design	• Consider using data markers to reduce cognitive load
Clutter	
Other	

Visual 5

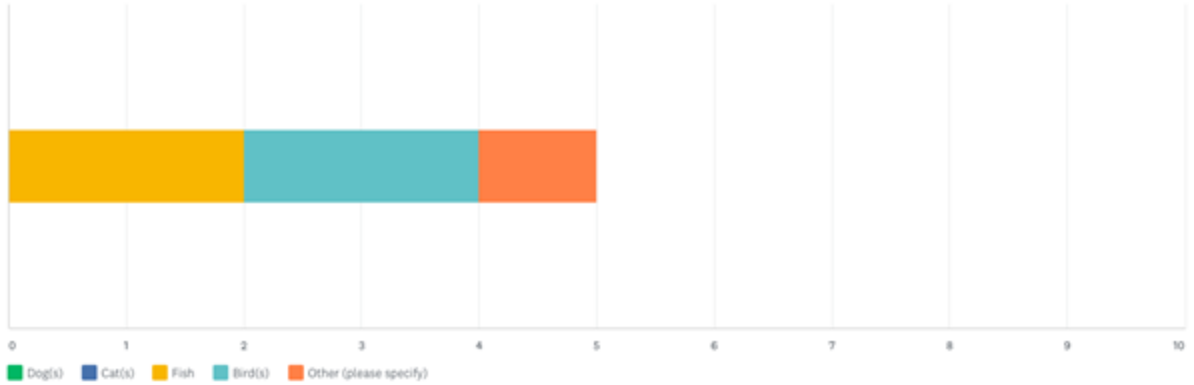


Are you interested in learning the Python programming language?	
Chart type	
Visual design	• The axis shouldn't go below 0 since that's not a possible result
Clutter	• Consider removing the "Y" axis labels since it's direct labeled
Other	

Visual 6

What time did you get out of bed today?	
Chart type	• A table doesn't help you understand the data; should be visualized
Visual design	
Clutter	
Other	• Data collection method may make visualization difficult

Visual 7



What types of pets do you own?	
Chart type	• This is not the best chart type for comparison
Visual design	• Consider direct labeling instead of a legend
Clutter	
Other	• We can't see what's in the "other" category

Visual 8

How many dogs do you own?	
Chart type	• Pie (and doughnut) charts generally shouldn't have more than 3 segments; consider a different chart type (if applicable)
Visual design	• Multiple categories use the same color making it difficult to distinguish them
Clutter	
Other	• Doughnut charts make sense when you want to utilize the middle space, which isn't the case here

Visual 9

How old are you?	
Chart type	
Visual design	• Consider direct labeling instead of a legend
Clutter	
Other	• Consider if, for your purposes, it's better to present the data as a percentage or as a discrete value

Visual 10

Would you ever go sky diving?	
Chart type	• It's hard to "get" this chart within 30 seconds
Visual design	• Consider direct labeling instead of a legend
Clutter	
Other	

Recap

- Data visualization is an art, informed by science.
- Use visual clues to make data visualizations easier for the audience.
- Reduce visual clutter to lower the cognitive load and help transmission of the message.

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

Misleading stats & visual distortions

- Sometimes charts and statistics look presentable but could be misleading.
- Unreliable data comparisons erode credibility and eventually dissuade viewers from using the analysis.

Misleading statistics

Misleading statistics

- "Bill Gates walks into a bar and everyone inside becomes a millionaire…on average."

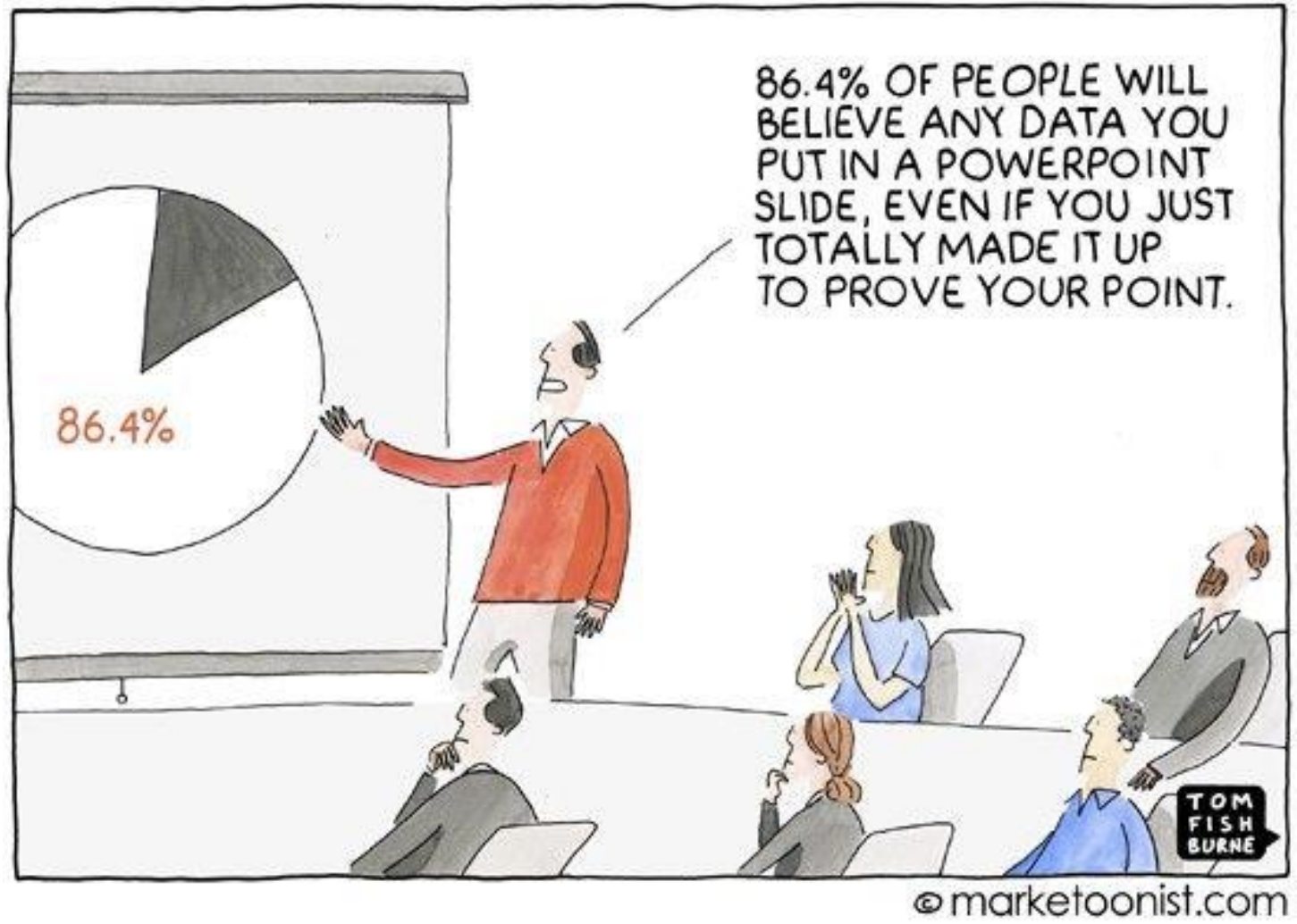
- In 2011, the average income of the 7,878 households in Steubenville, Ohio, was $46,341. But if just two people, **Warren Buffett** and **Oprah Winfrey**, relocated to that city, the average household income in Steubenville would rise 62 percent overnight, to $75,263 per household.

What's wrong with these statements?

https://www.nytimes.com/2013/05/26/opinion/sunday/when-numbers-mislead.html

Misleading statistics

- Numbers don't have to be fabricated to be misleading.
- Misleading statistics are the misusage—purposeful or not—of numerical data.



Misleading statistics

- Misleading statistics can be created through issues with:
  - data collection
  - data processing
  - data presentation

Data collection	• Small sample sizes • Biased sampling • Loaded questions
Data processing	• No/poor data normalization • Ignoring important features
Data presentation	• Hiding context • Omitting certain findings • Visual distortions

Example: Bem's PSI research

"I'm all for rigor, but I prefer other people do it. I see its importance—it's fun for some people—but I don't have the patience for it. If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?" (Daryl J. Bem, in Engber, 2017)

How to avoid being misled?

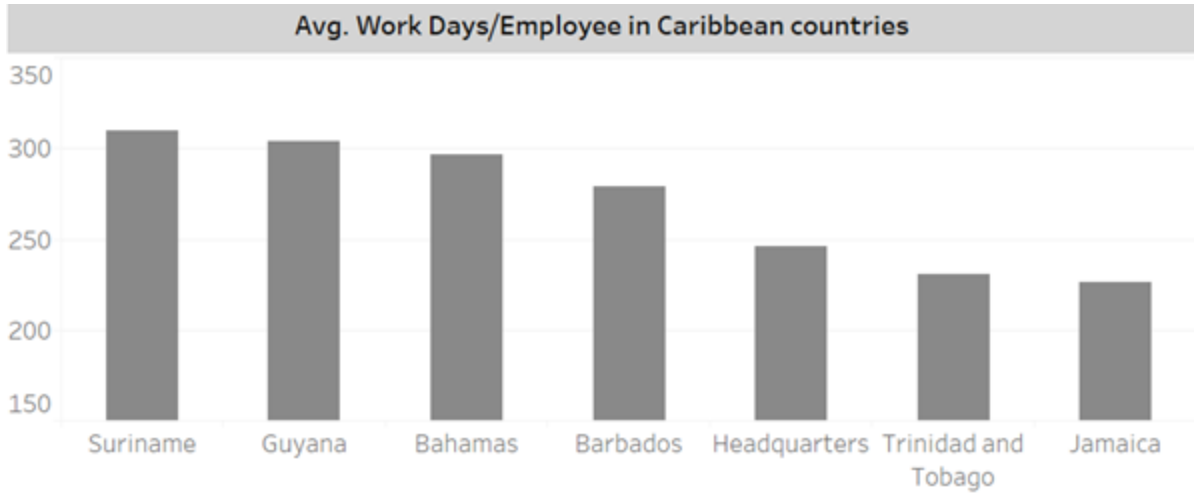
- Do some math. Are there any obvious mistakes?
- Check the source. Is it creditable and current?
- Question the methodology. Is there bias? Is the result statistically significant?
- Conduct research. What does Google tell you?

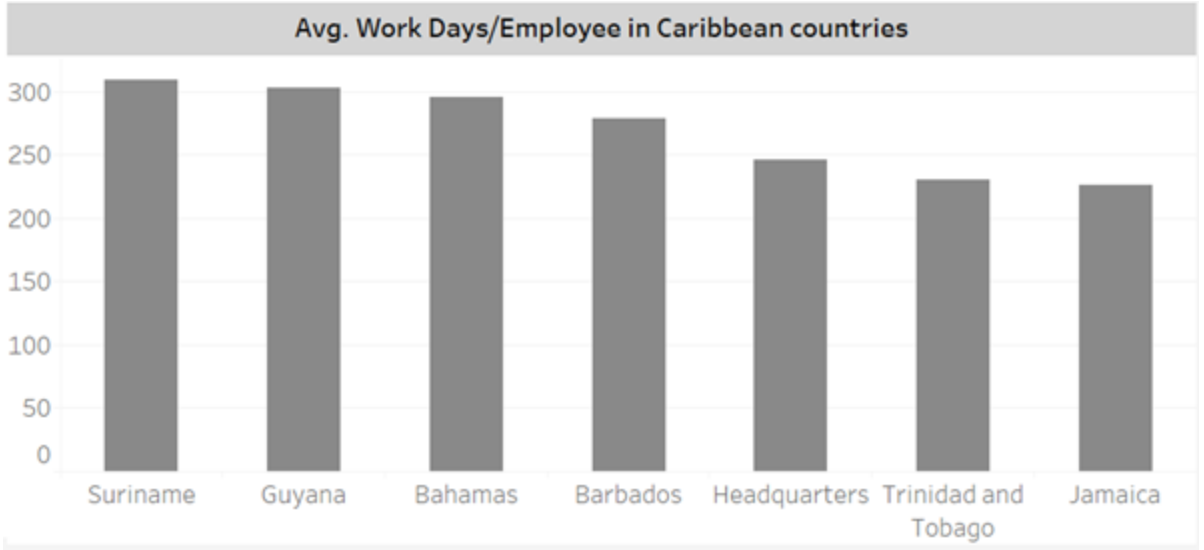
Visual distortions

Visual distortions

- Look at the top graph.
- At first, Jamaica seems to have half the average workdays per employee that Suriname does.
- In reality, the difference is much less.

What's the difference between the two charts?

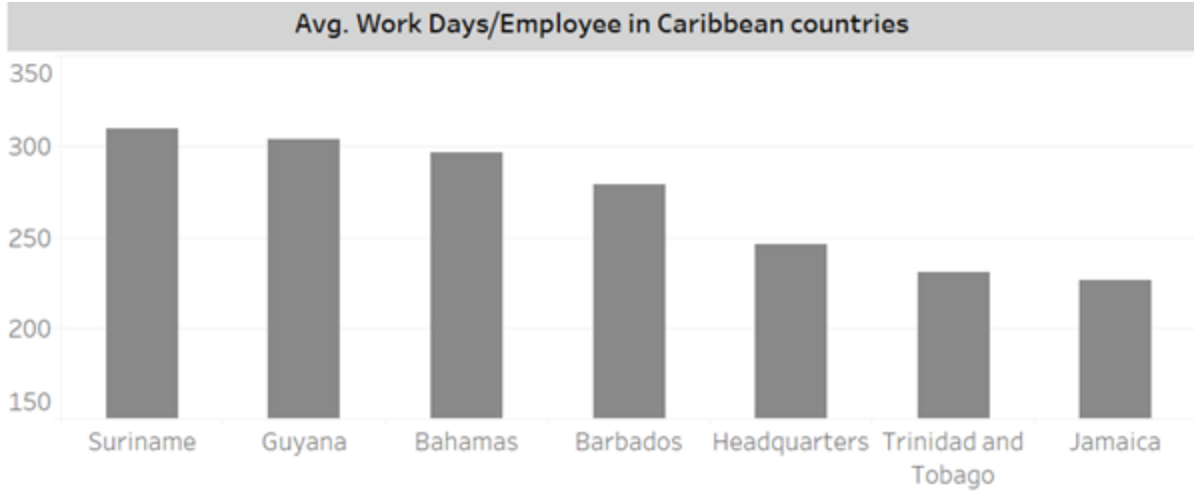


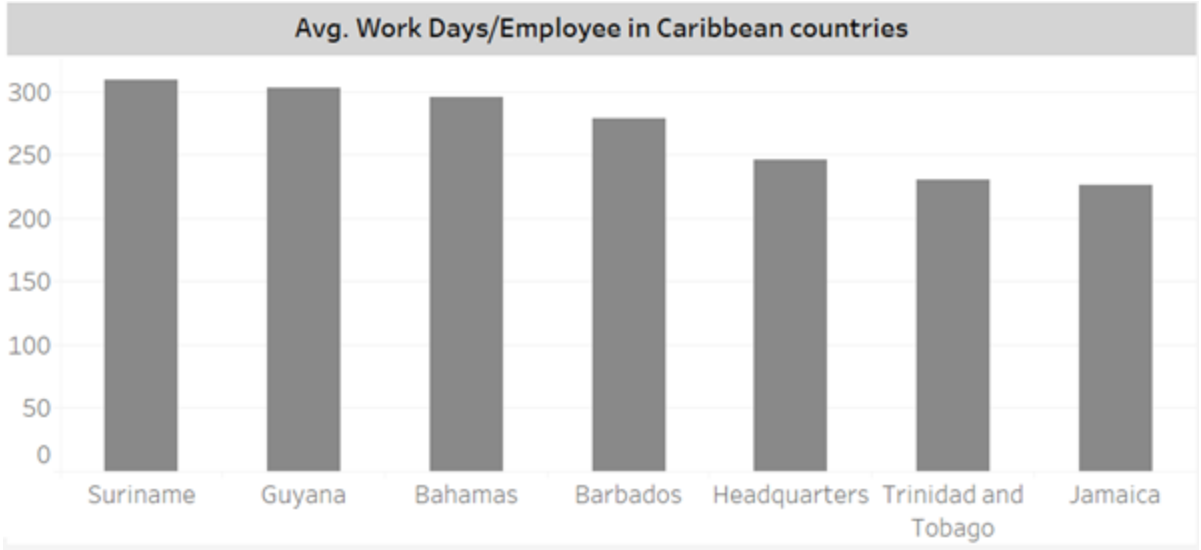


Truncated graphs

- One of the most common manipulations is omitting baselines or beginning the y-axis of a graph at an arbitrary number instead of 0.

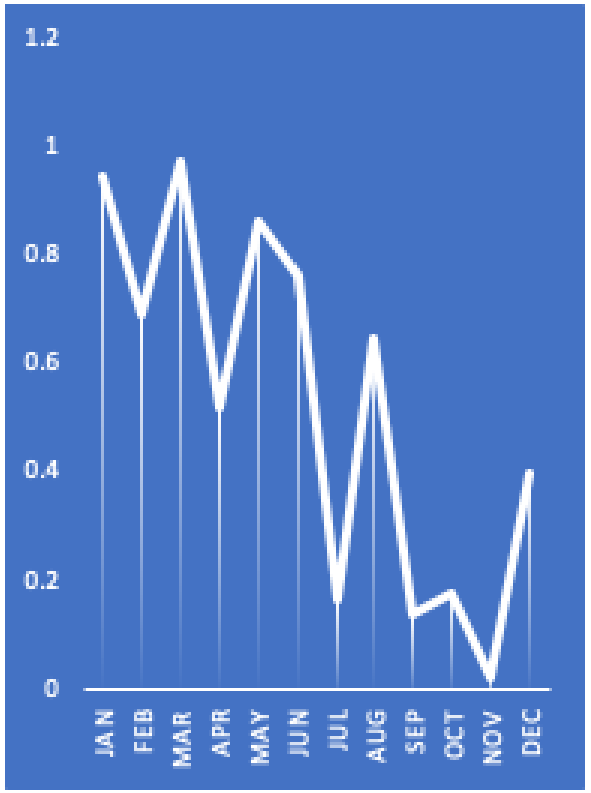
- This creates the impression that there is a significant difference between data points, when in fact, there is relatively little disparity.





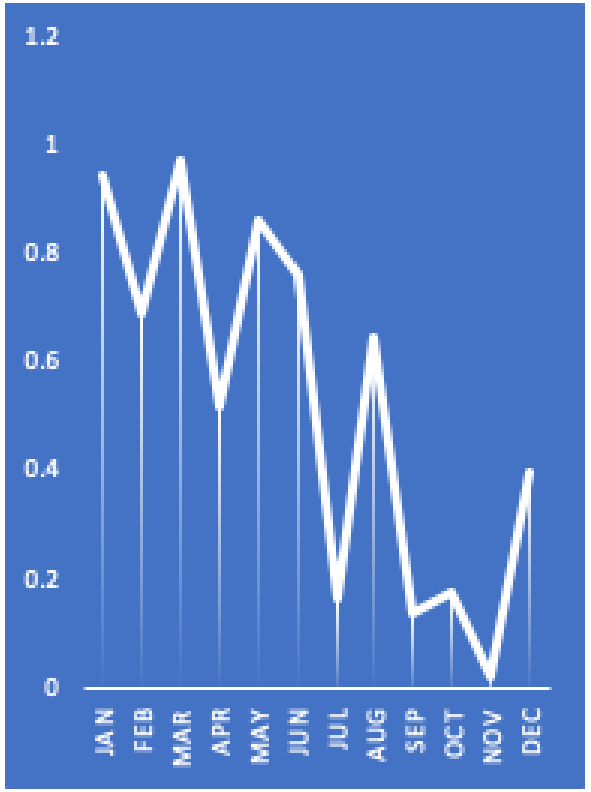
Visual distortions

What distortion has been used in these charts to change how the data appears?



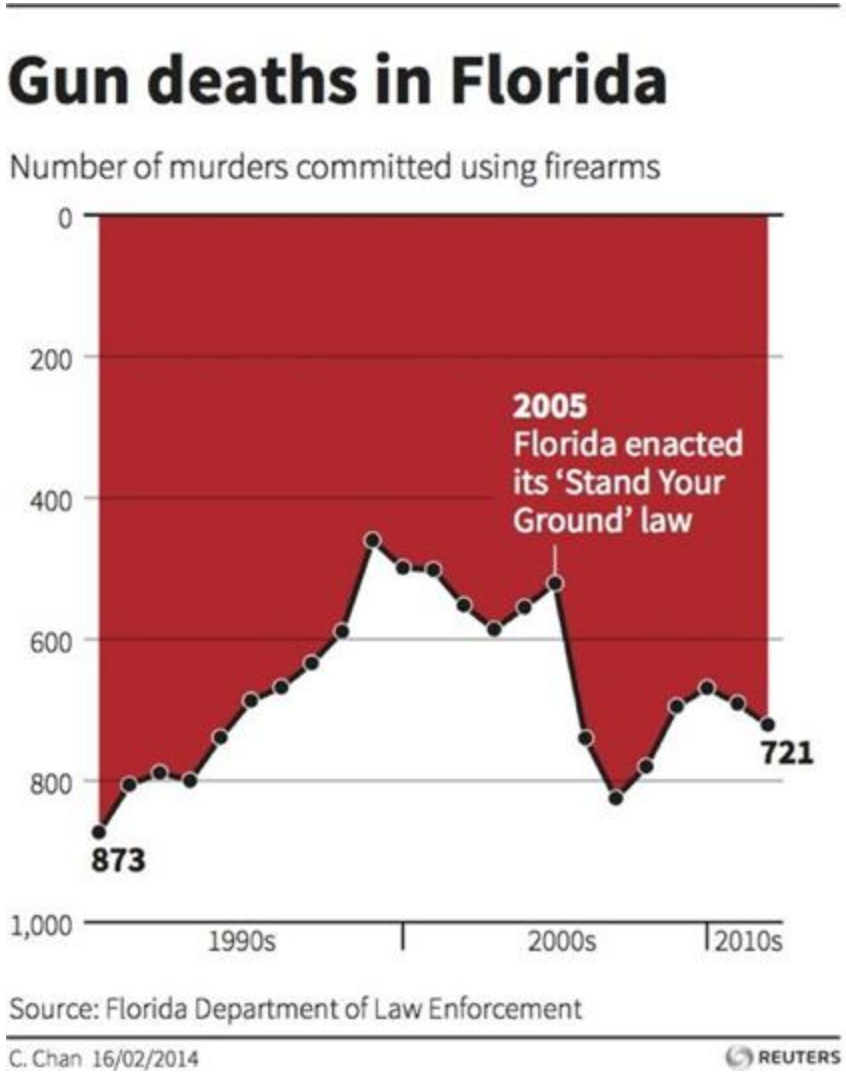
Exaggerated scaling

- Exaggerating the scale of a line graph can easily minimize or maximize the change shown.



Visual distortion

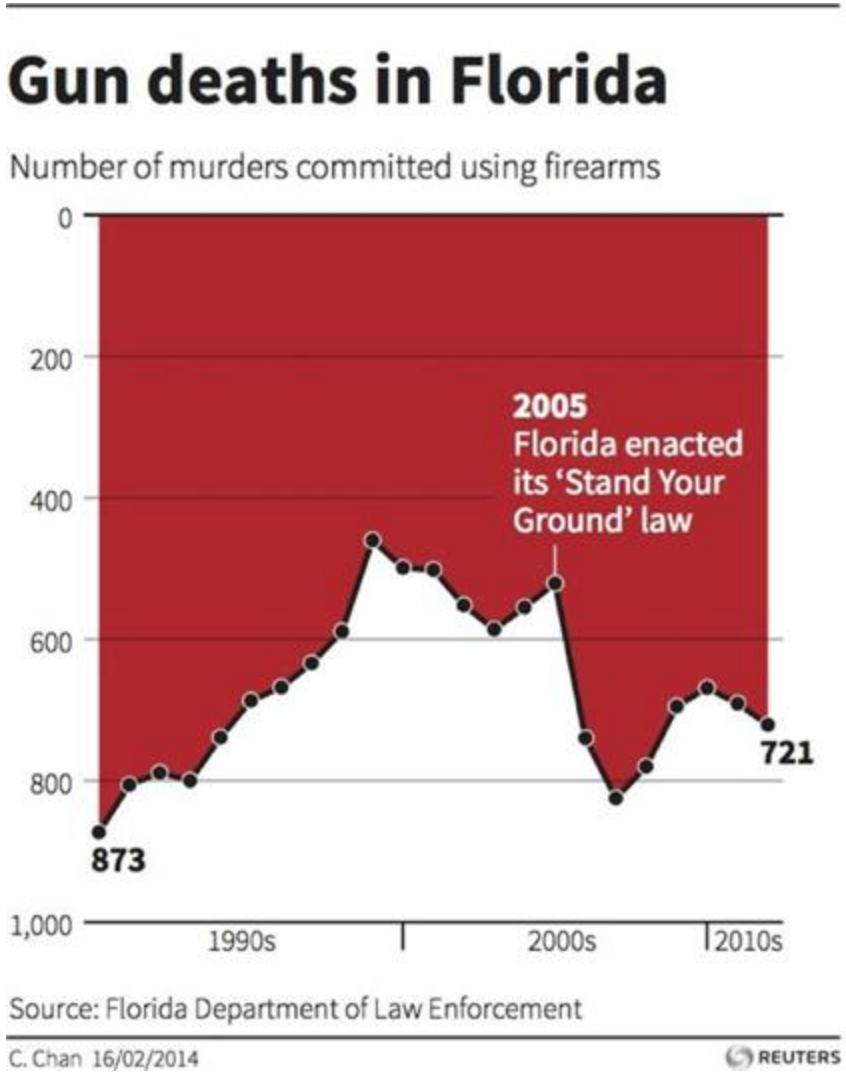
How might this chart be misleading?



Ignoring convention

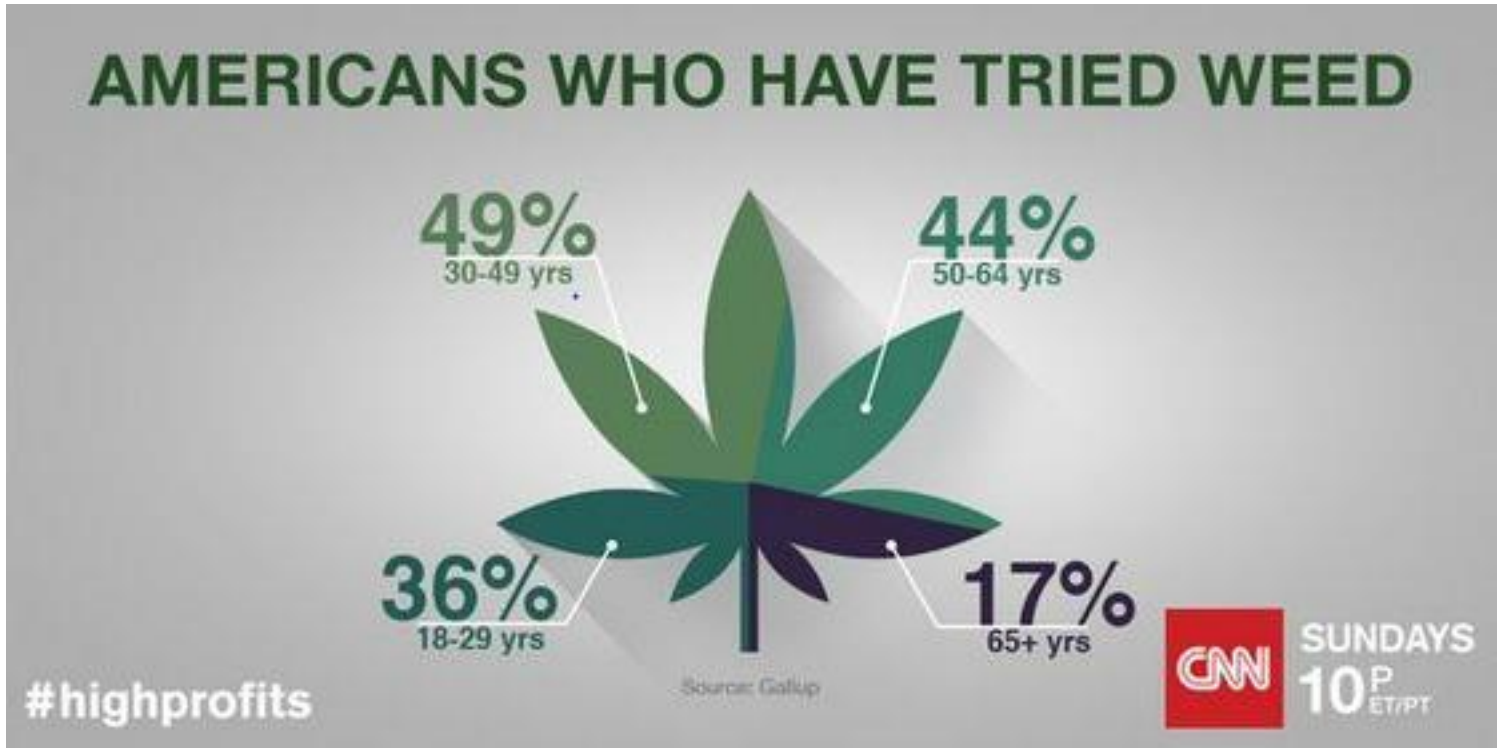
- Deviating from convention (such as green is positive and red is negative) can create confusion and misinterpretation of the facts.

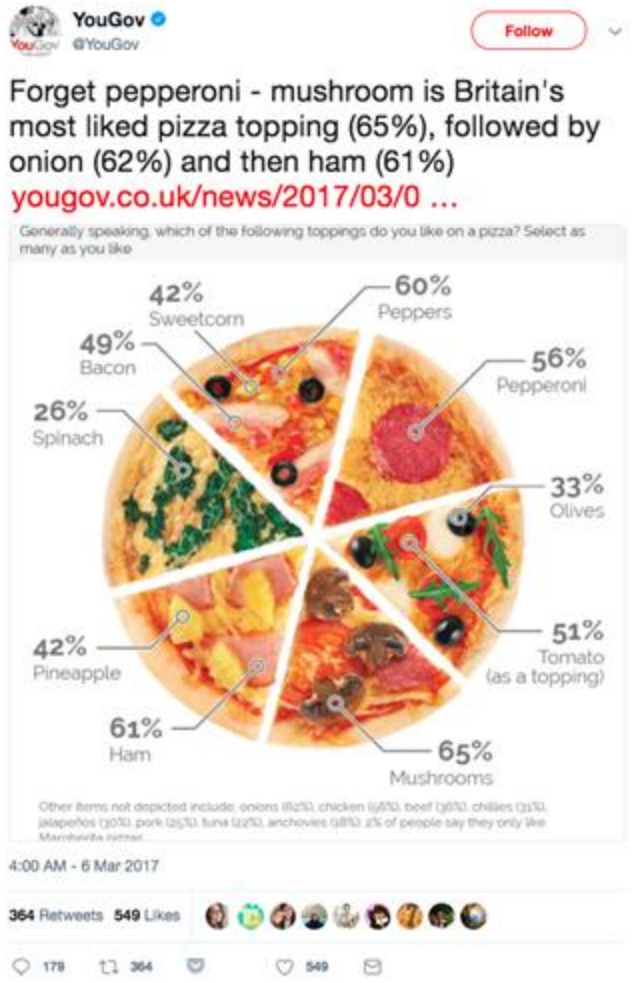
- In this example, the axis also moves downward, making a decrease in murders look like an increase, at a quick glance.



Visual distortion

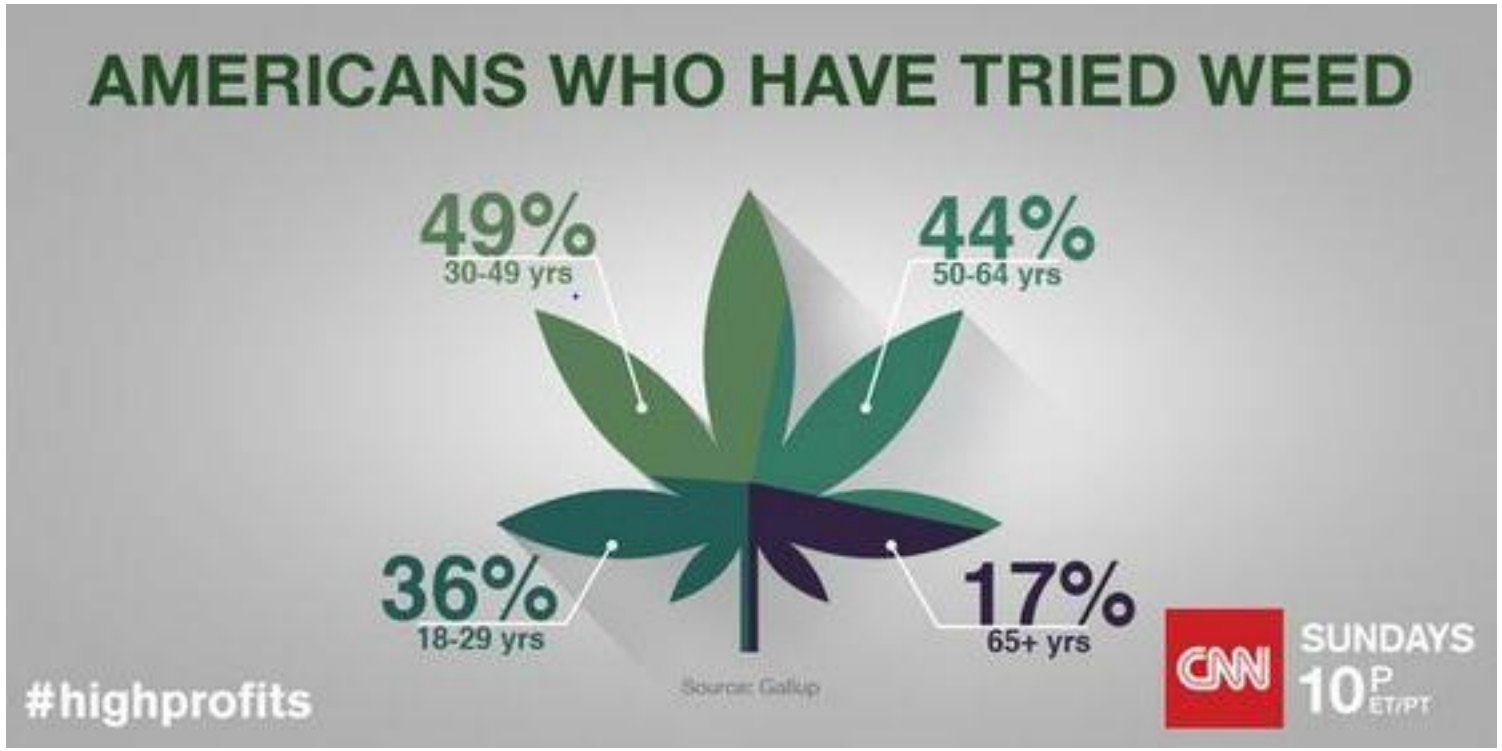
What do you notice about these pie charts?

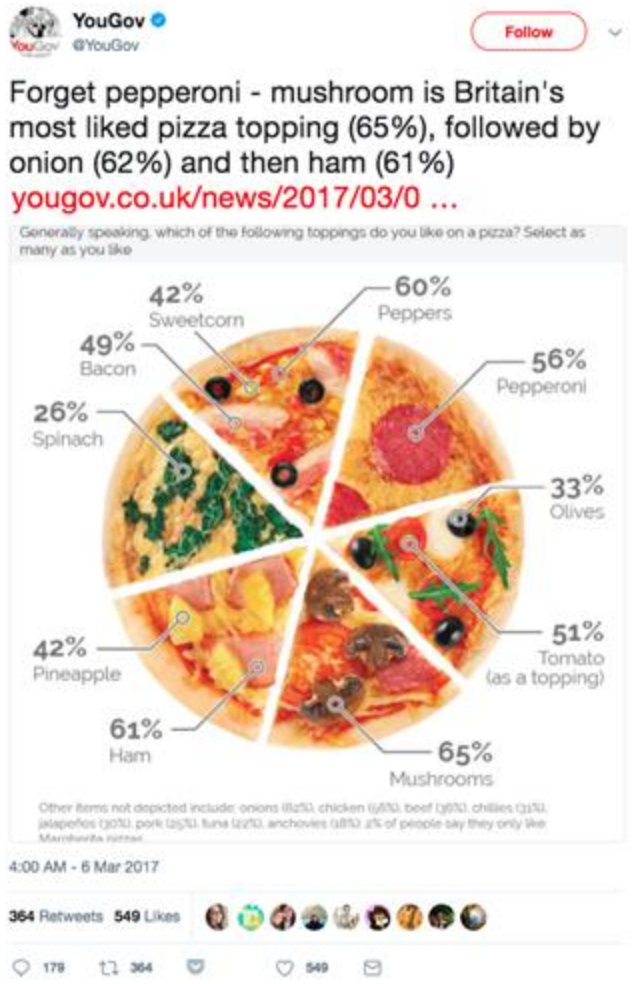




Numbers don’t add up

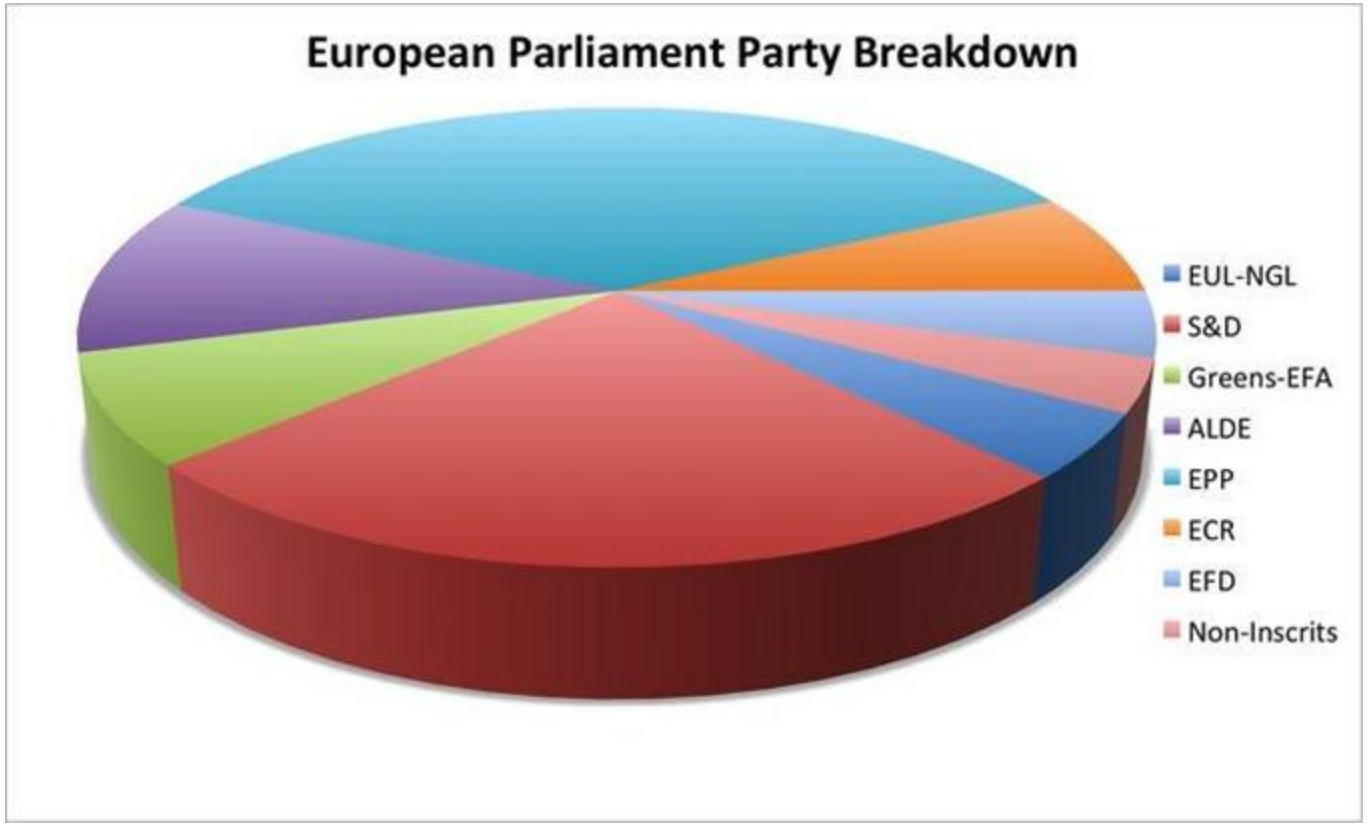
- With pie charts, the sum of each slice must add up to the whole. When the numbers don't add up, you know there's an issue.





Visual distortion

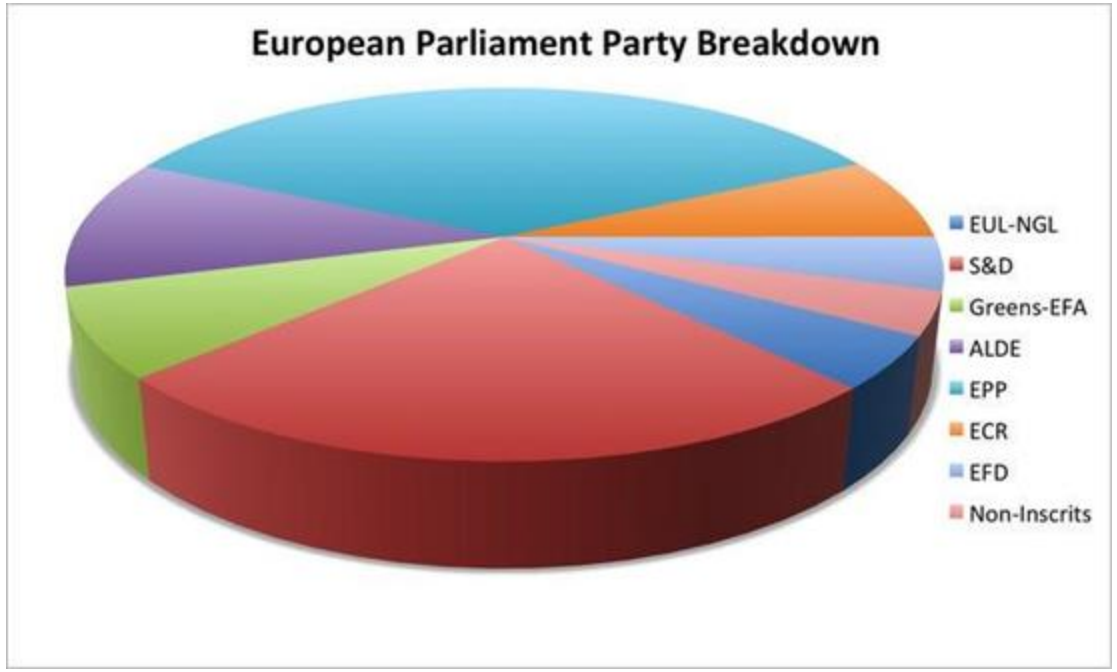
Does the S&D or the EPP party have more representation in parliament?

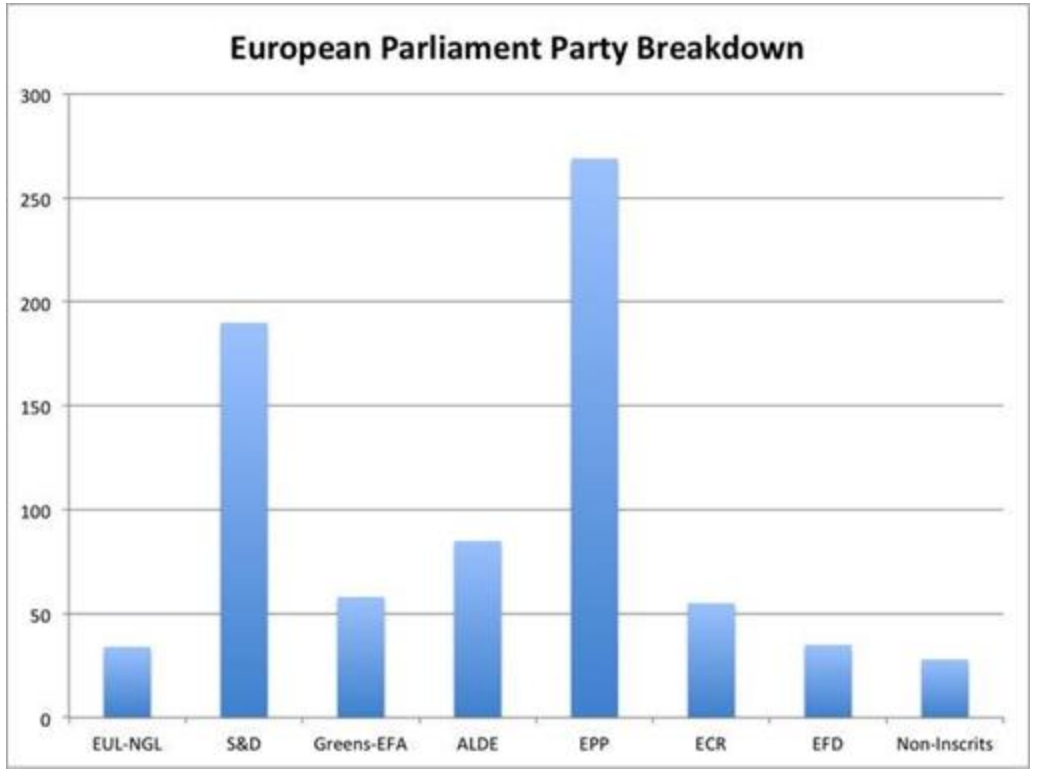


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

3D distortion

- 3D pie charts can be used to distort and cause a misinterpretation of the data.
- The same data is represented in both charts below.

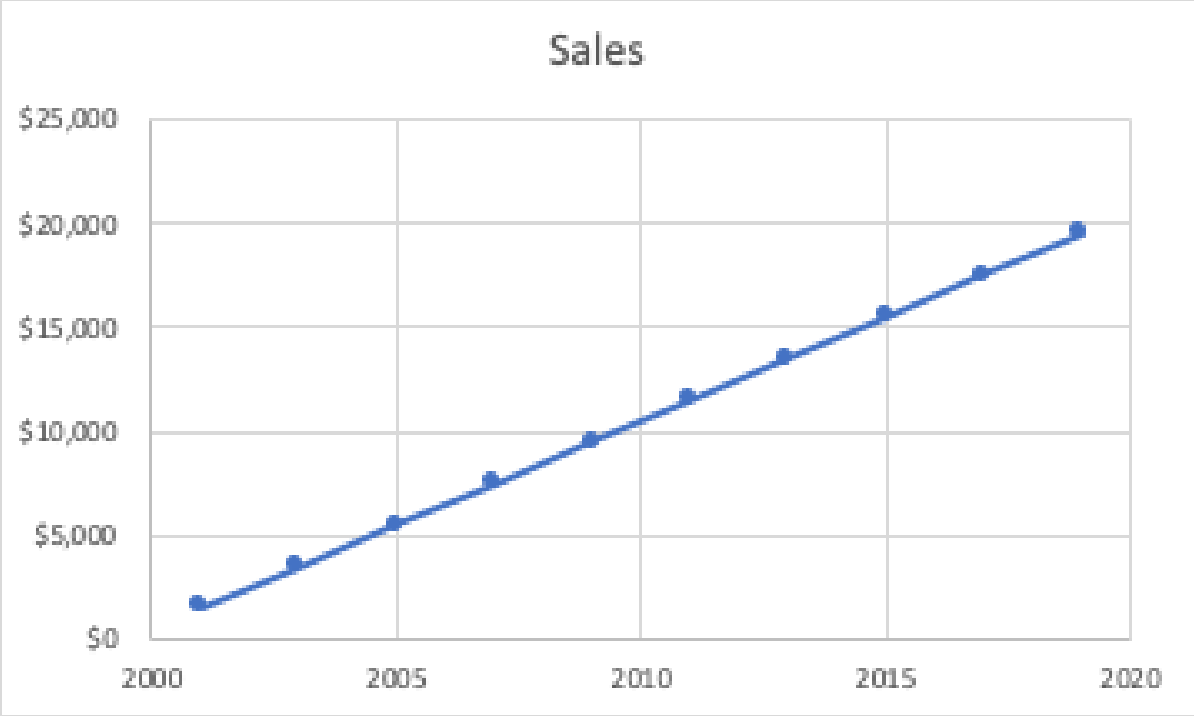


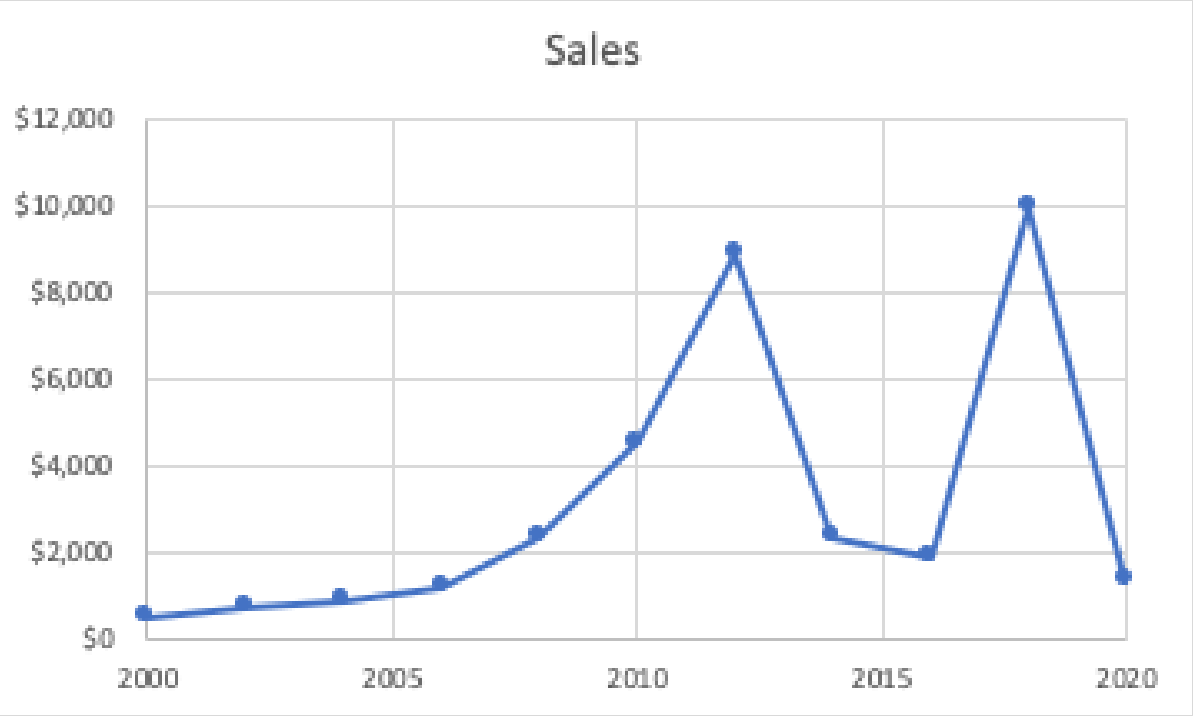


https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

Visual distortion

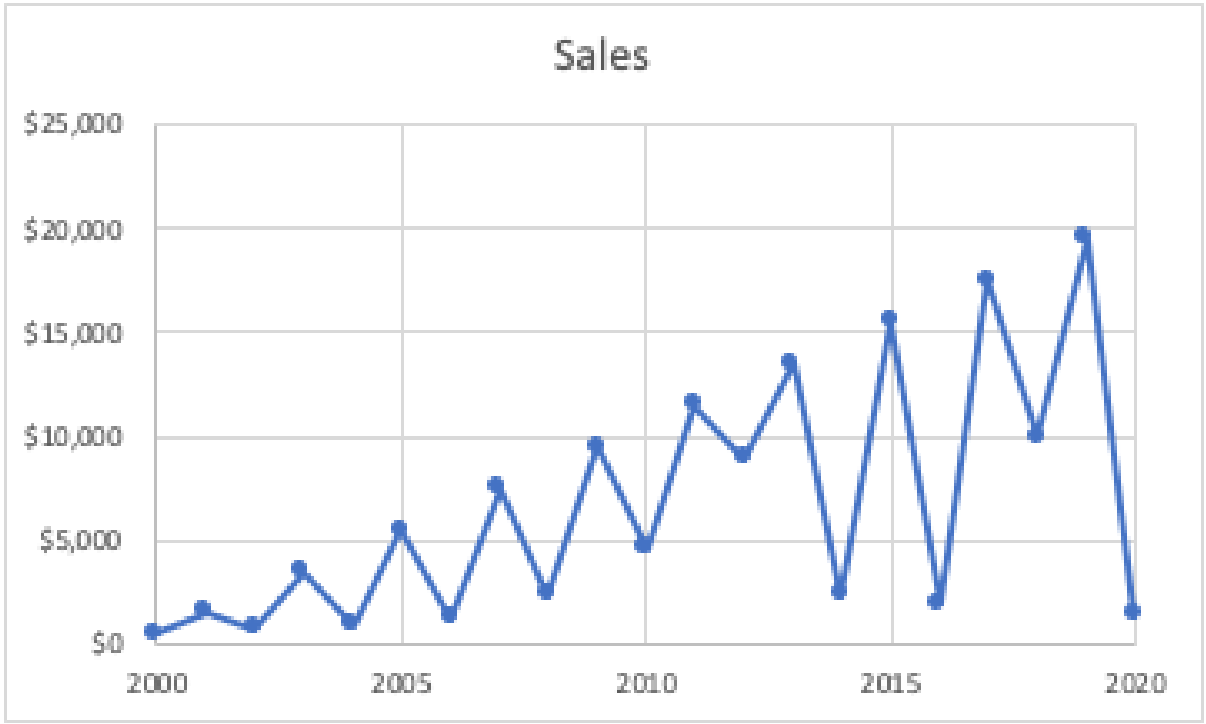
Which company has a better sales trajectory?



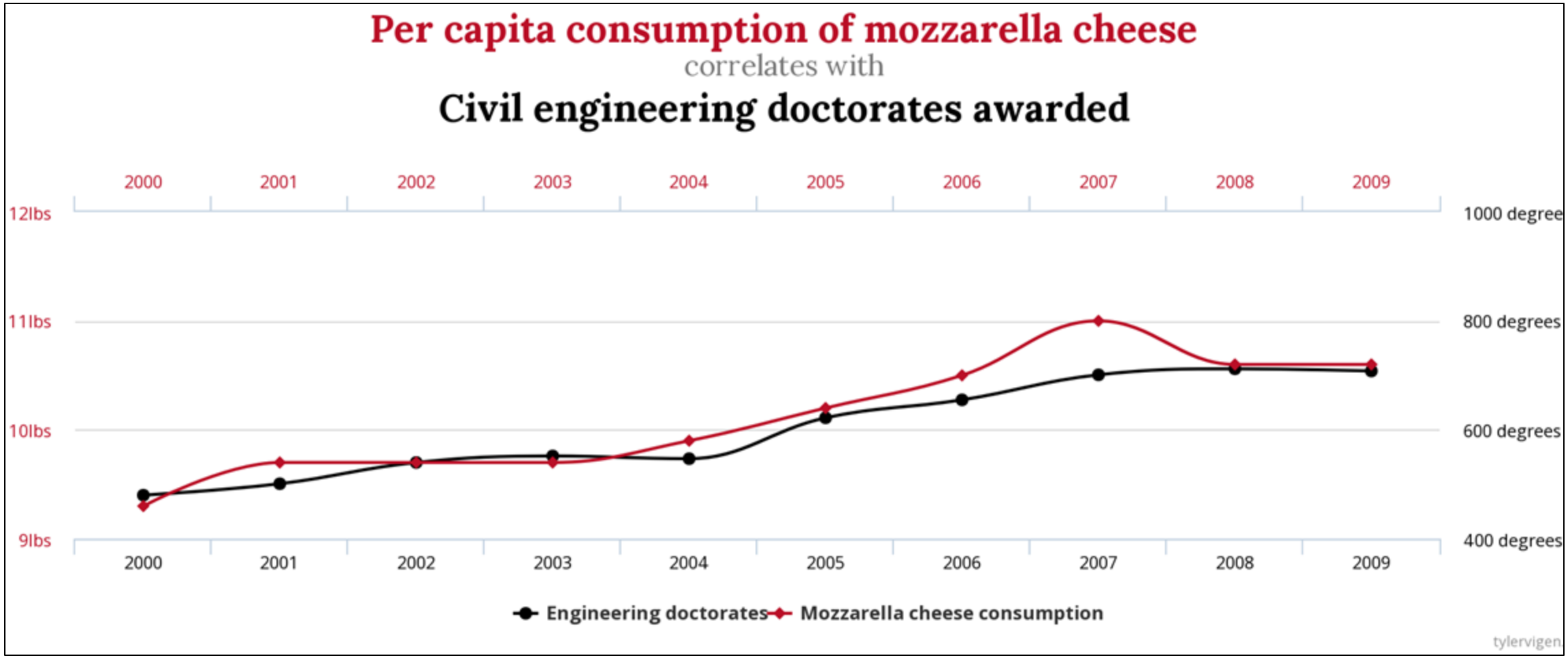


Improper extraction

- Surprise! It's the same company. One graph showed only odd years and the other only even.
- To align to a particular narrative, some may choose to visualize only a portion of the data.
- This is more common in graphs that have time as one of their axes.



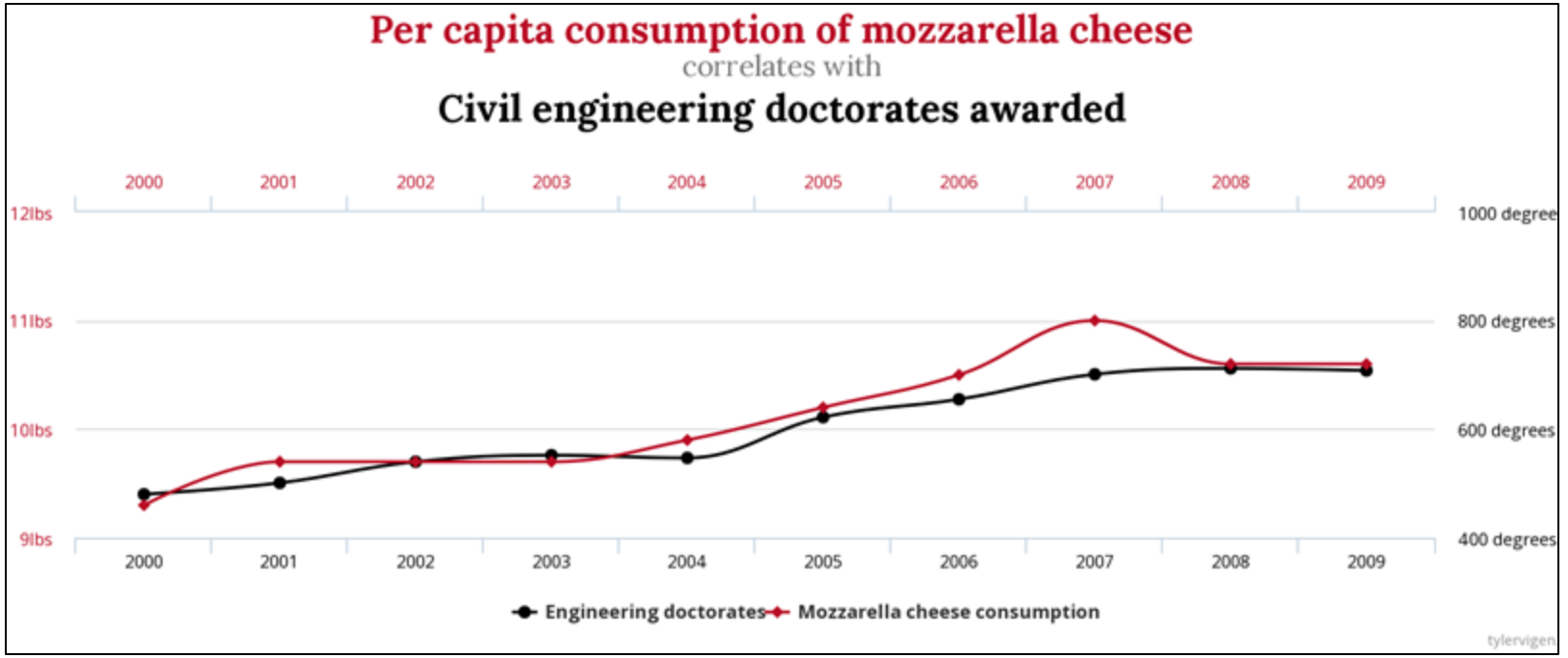
Visual distortion



What story does this visualization tell?

Correlating causation

- Data visualizations can create causal links by the way that data is presented to the viewer.
- However, correlation does not equal causation.



Recap

- To avoid being misled, look for:
  - misleading statistics
  - truncated graphs
  - exaggerated scaling
  - ignored conventions
  - numbers that don’t add up
  - 3D distortion
  - improper extraction
  - correlating causation

Q&A

DATA SOCIETY®

Welcome back
Day 4

Agenda

Day 1

- The basics of data visualization
- Getting started with data visualization
- The anatomy of a chart

Day 2

- Common charts and graphs
- Choosing a visual
- Reducing chart clutter

Day 3

- Visual design theory
- Common mistakes
- Misleading statistics & visual distortions

Day 4

- Data storytelling

A story: the Chip Heath experiment

- Small groups of students are provided with a mass of statistical crime data.
- Each student must give a 1-minute pitch about why nonviolent crime is or isn't a problem.
- The students rate one another's performance, then watch a 10-minute movie, thinking the exercise is over.
- After 10 minutes, students are asked to write down every idea that they remember from their groupmates' speeches.

Stories > statistics

- *"In the average one-minute speech, the typical student uses 2.5 statistics. Only one student in ten tells a story. Those are the speaking statistics. The 'remembering' statistics, on the other hand, are almost a mirror image: when students are asked to recall the speeches, 63 percent remember the stories. Only 5 percent remember any individual statistic."* (*Made to Stick: Why Some Ideas Survive and Others Die*)

- Storytelling has come a long way since the campfire, but narrative remains the dominant cultural framework for conveying experience, connecting emotionally, and changing minds.

What is data storytelling?

1. You focus on an **insight** and
2. persuade an **audience**
3. that the **outcome** of your analysis
4. demands a course of **action**
5. through narrative and visual **communication.**

Data stories and data visualizations

- A **single** data story may make use of **multiple** data visualizations.
- Data stories arrange visualizations into the **linear sequence** of storytelling: a beginning, a middle, and an end.
- Data story formats will likely incorporate other elements to explain and contextualize the visualizations:
  - prose text, either written or spoken
  - annotations, callouts, and labels
  - icons or graphics
  - images or photographs

Can’t I just use a chart?

1. Narratives are super effective, “**sticky**” content delivery mechanisms.

2. Not everyone is a statistician, but they still want to make **evidence-based** decisions.

3. Stories let you overview key findings **quickly.**

4. Stories tap into both the **logical** and the **emotional** aspects of persuasion.

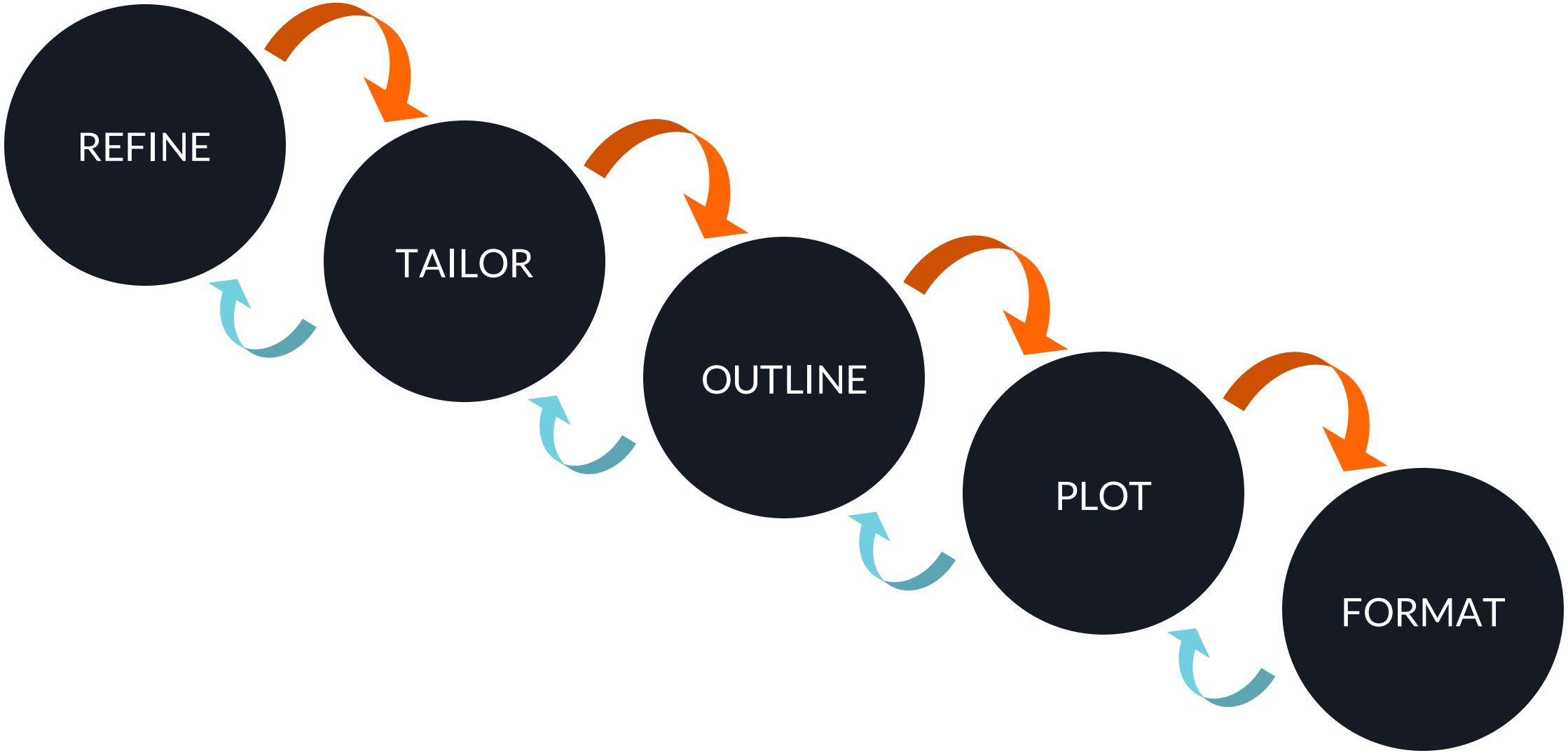
Why choose story?

If your insight is...	A story can...
Unpleasant	Help convince your audience that even unwanted results are actionable.
Disruptive	Encourage your audience to break with tradition, if the upshot is valuable enough.
Unexpected	Explain why a prediction or intuition failed, and offer some analysis and a solution.

Why choose story?, ctd.

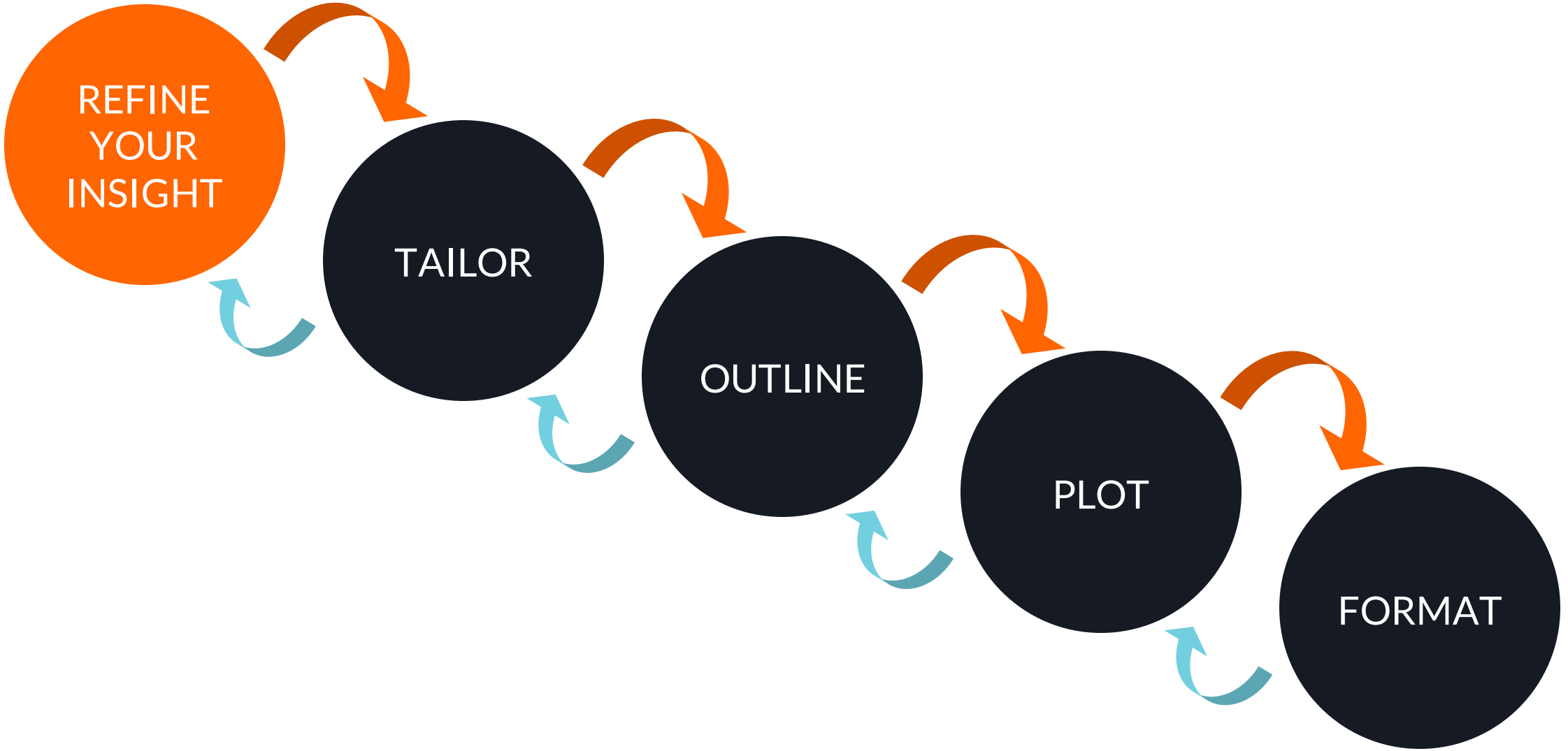
If your insight is...	A story can...
Complex	Guide your audience to a more complete understanding in manageable chunks.
Risky	Embolden your audience to take responsibility for making a tough choice.
Costly	Compel your audience to consider a high-cost solution by underscoring the high value.

Storytelling as process



Refining your insight

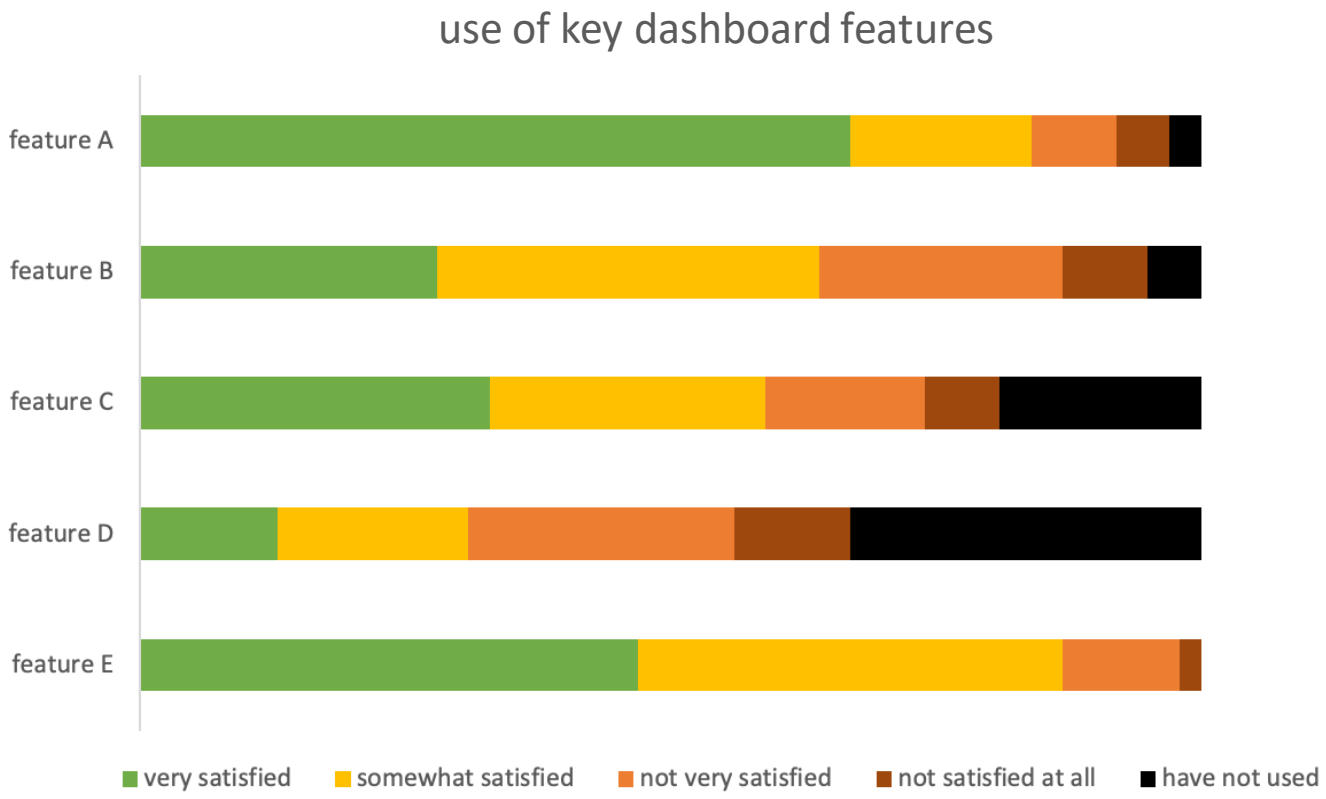
Storytelling step 1



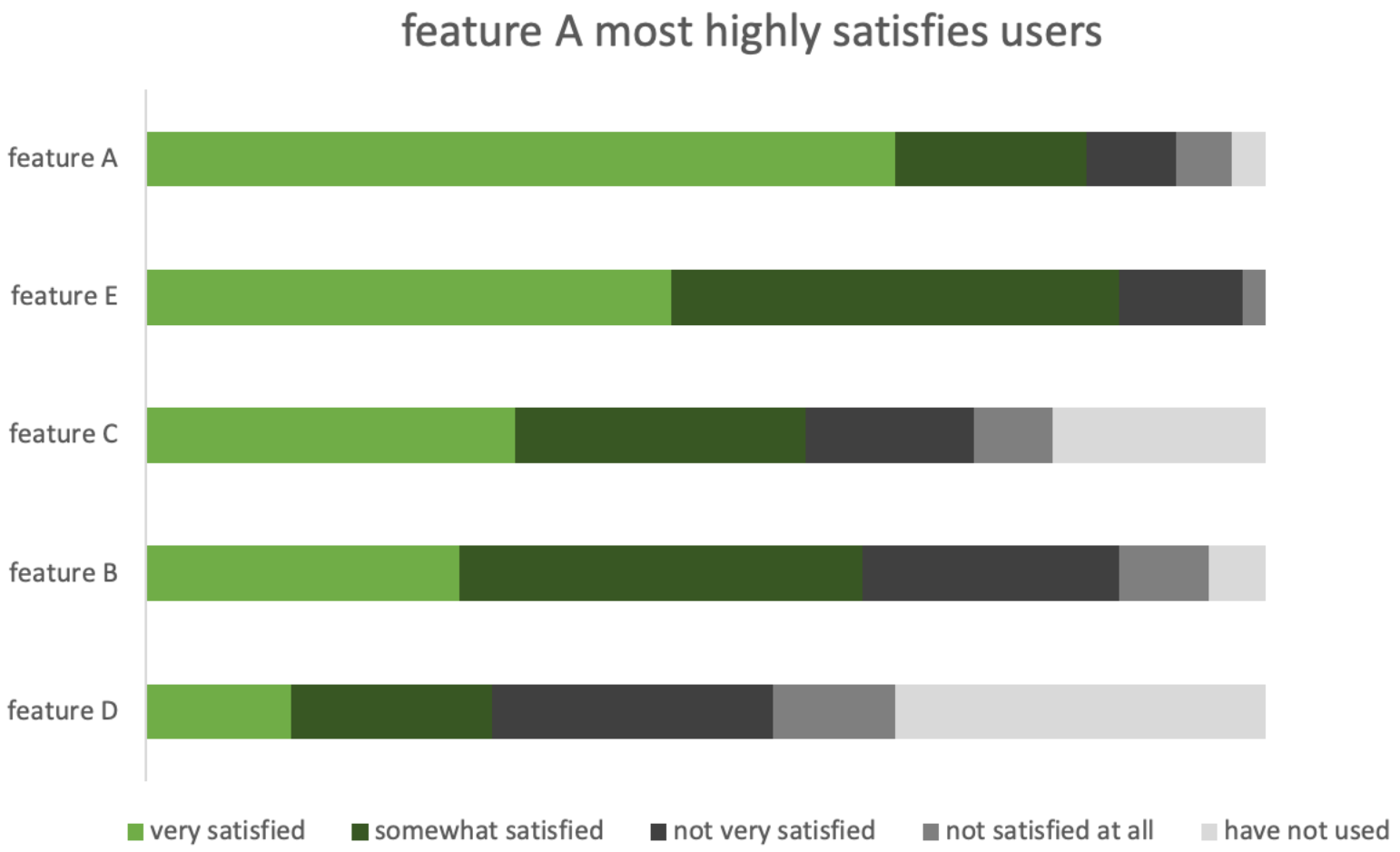
Insight

*Data stories are built around **insights**.*

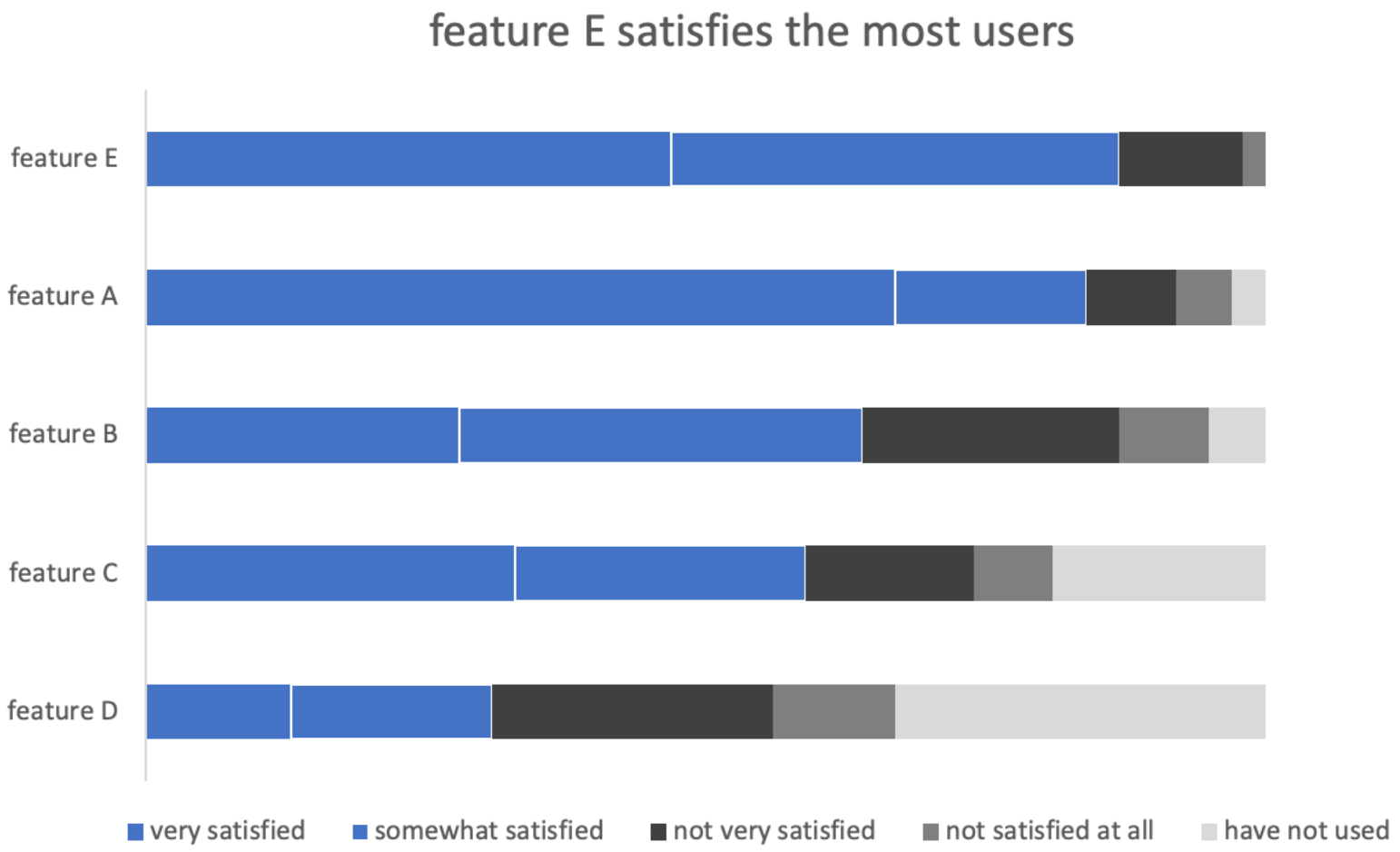
The same data visualization may reveal multiple insights.



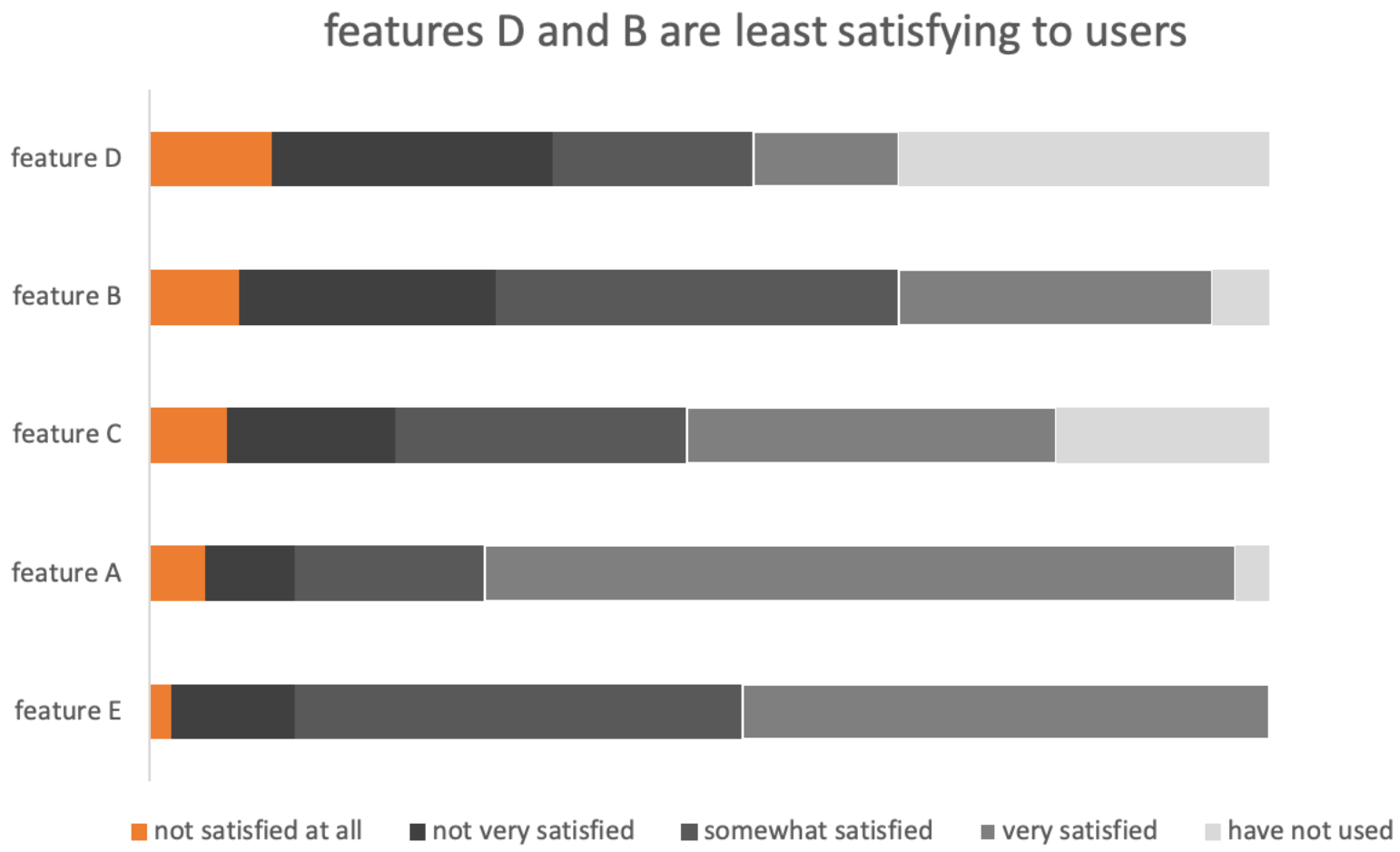
Insight #1



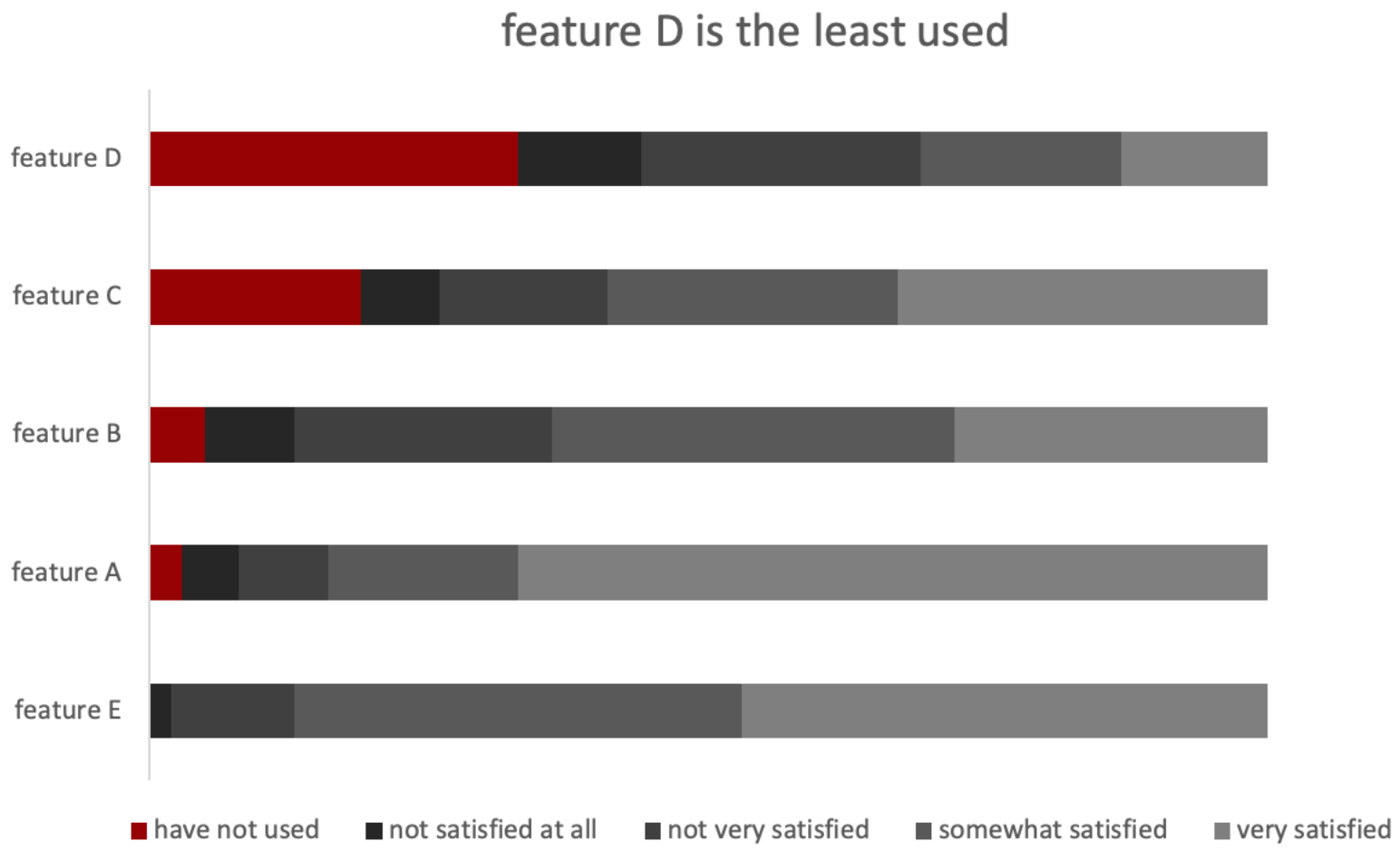
Insight #2



Insight #3


features D and B are least satisfying to users
feature D
feature B
feature C
feature A
feature E
not satisfied at all
not very satisfied
somewhat satisfied
very satisfied
have not used

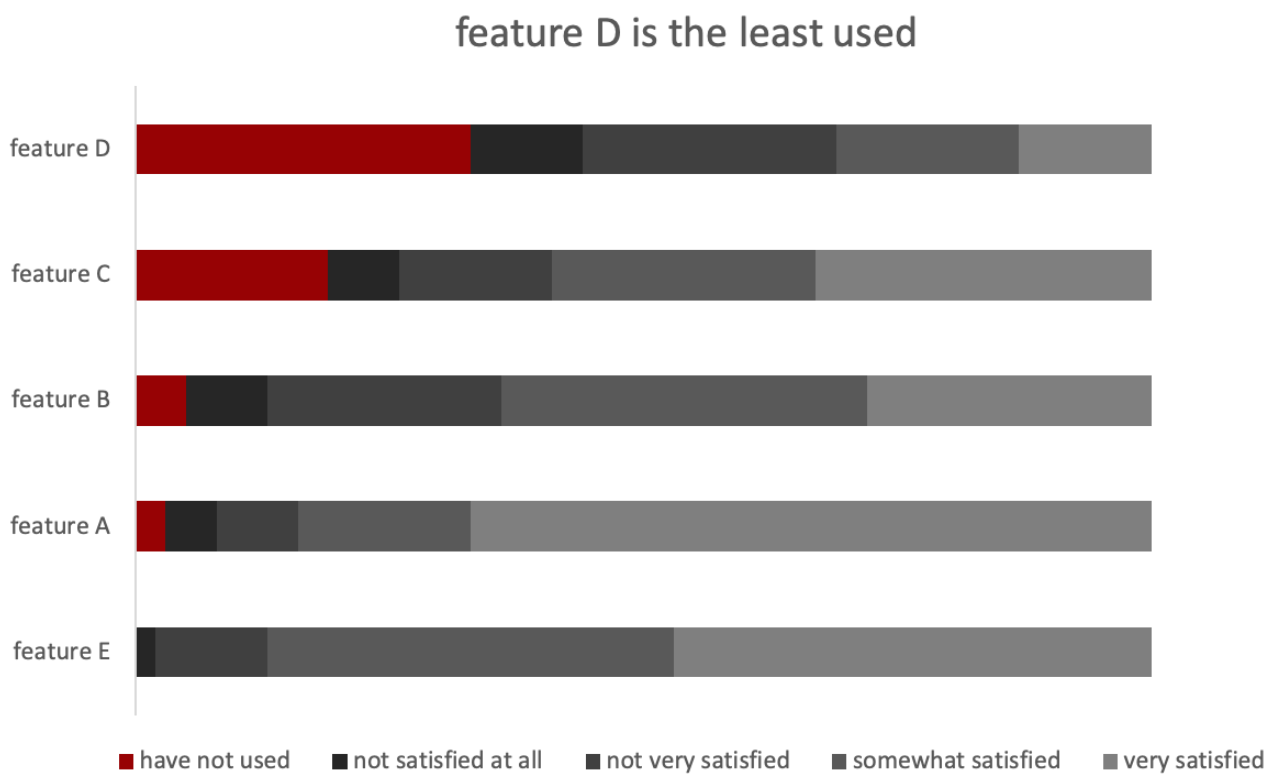
Insight #4



Framing

- In a data story, your **insight** is the most important piece.
- What will make your audience perceive your insight as maximally:
  - **Valuable:** an observation that seems to be rewarding
  - **Relevant:** an observation that seems timely
  - **Practical:** an observation that suggests a realistic and feasible course of action
  - **Specific:** an observation that clearly and completely accounts for a problem
- Make your insight as **concrete** and **contextualized** as possible

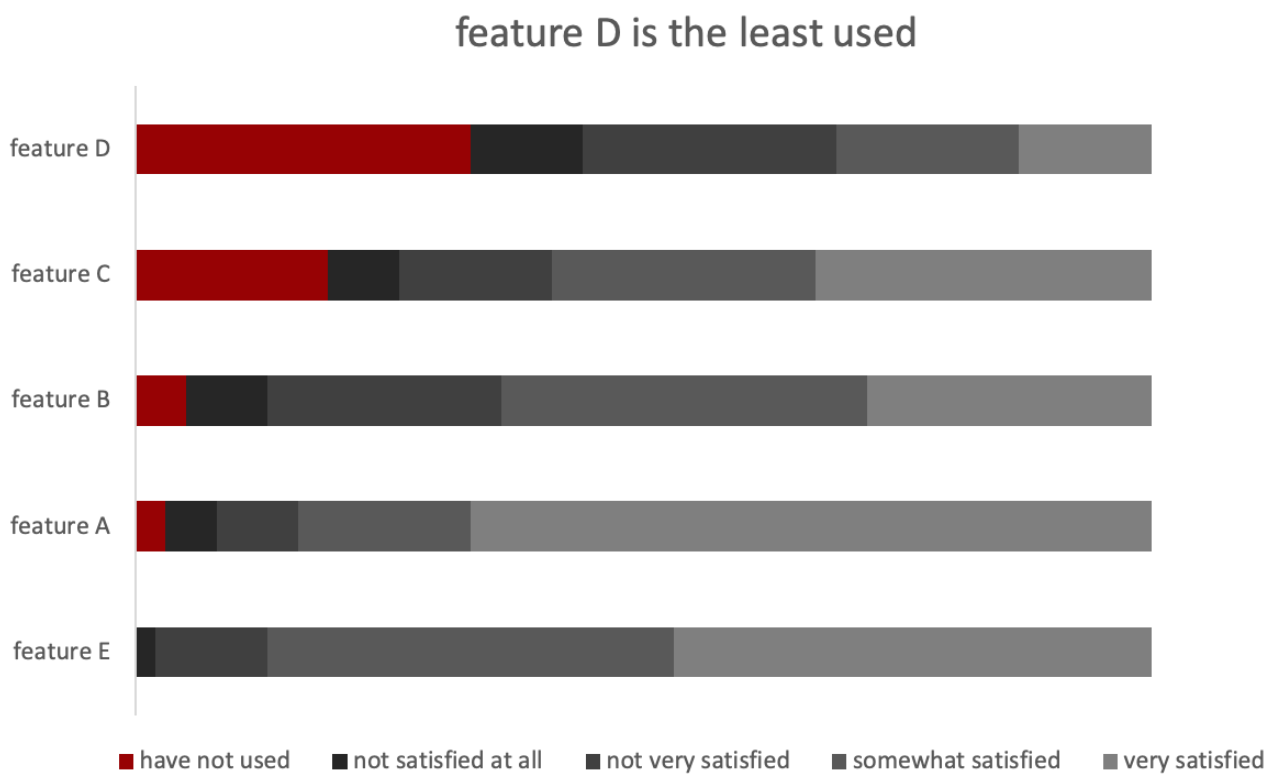
So what?



*Each data story has **one key <u>actionable</u> insight**, the main idea to be communicated.*

- Why should your audience care?
- What should they do about it?
- What is the potential impact?

Actionable insight



From this:

Feature D is the least used. ***So what?***

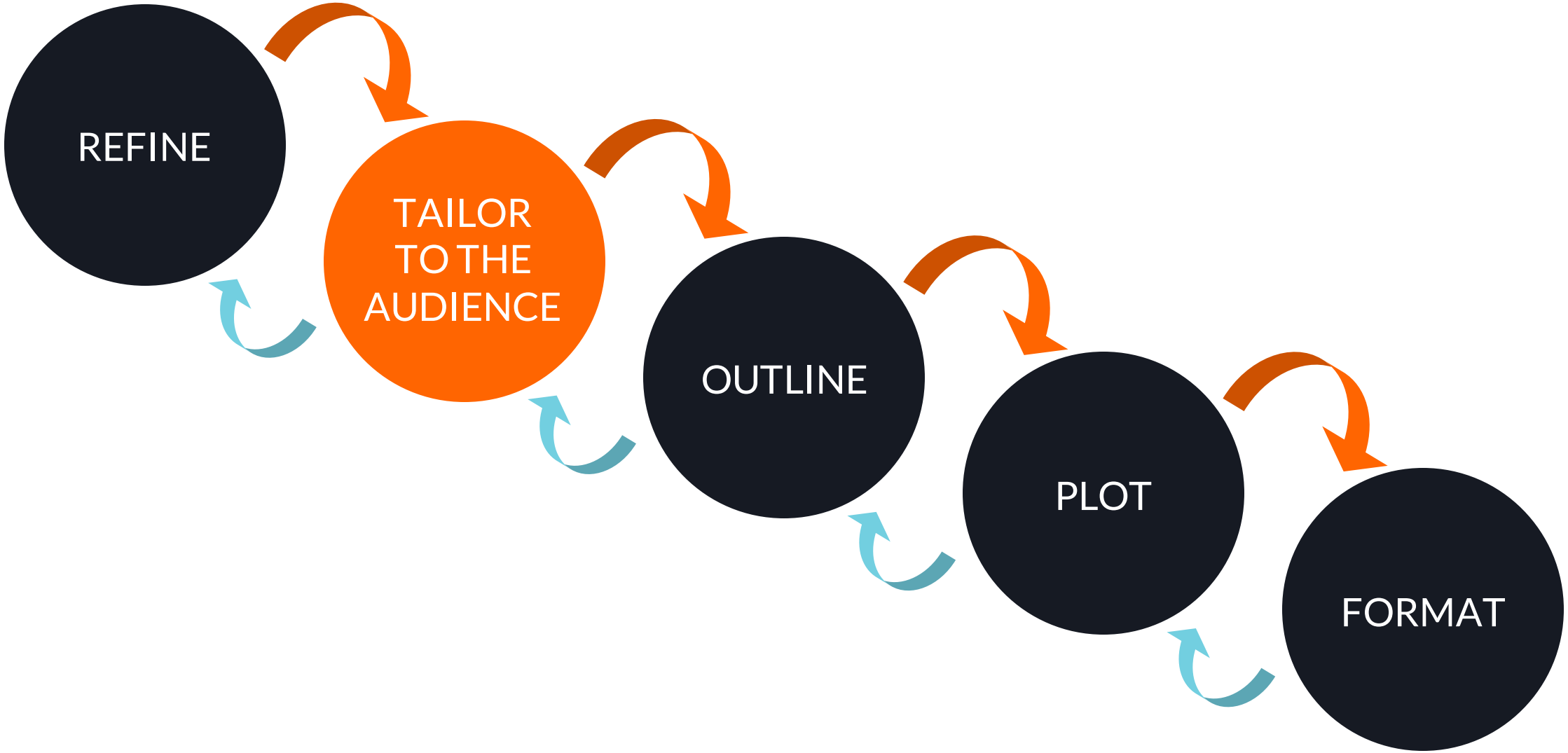
To this:

Features B and D are about equally unsatisfying to users, but feature D is used 5 times less than feature B. We need to figure out why.

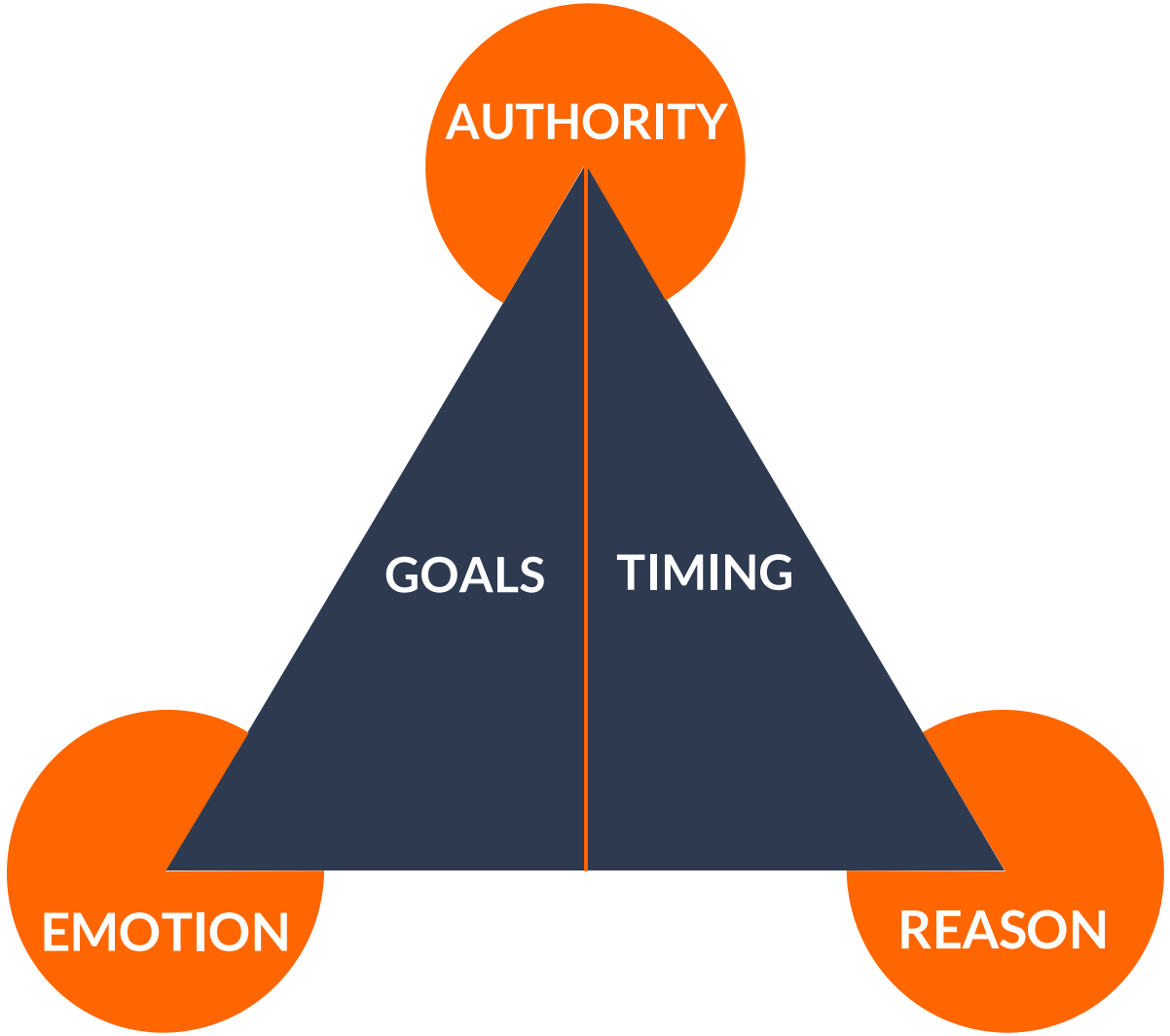
Note: you haven't *decided* on an action yet. You've just increased the urgency to act.

Tailoring to your audience

Storytelling step 2

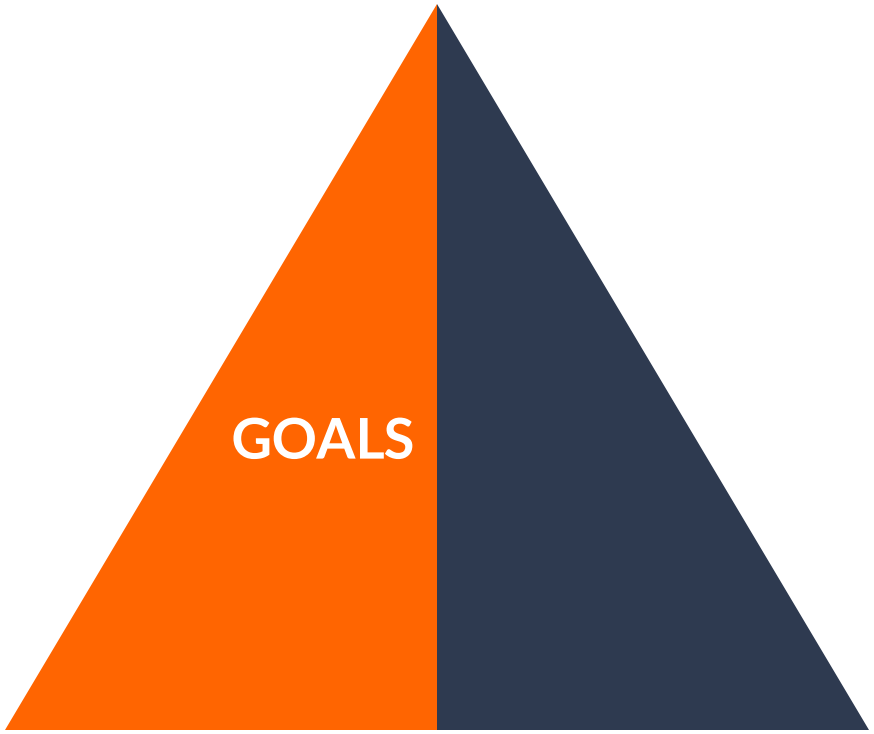


Engaging your audience



Goals and priorities

- What matters most to your audience?
- What key objectives do they most focus on achieving?
- What metrics inform their decision-making strategy?

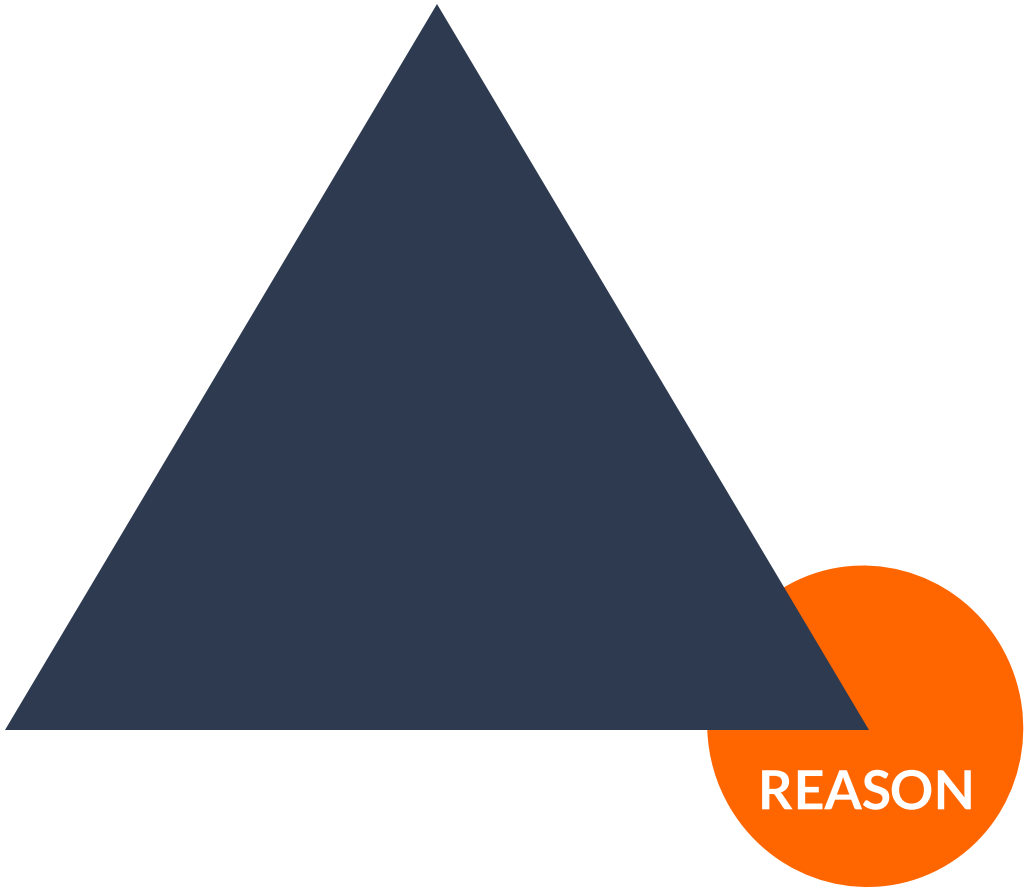


Opportune timing

- What are the most pressing concerns for your audience right now?
- Is your insight best acted on in the short-term or the long-term?
- Is interest in your insight high? Will it be higher later?

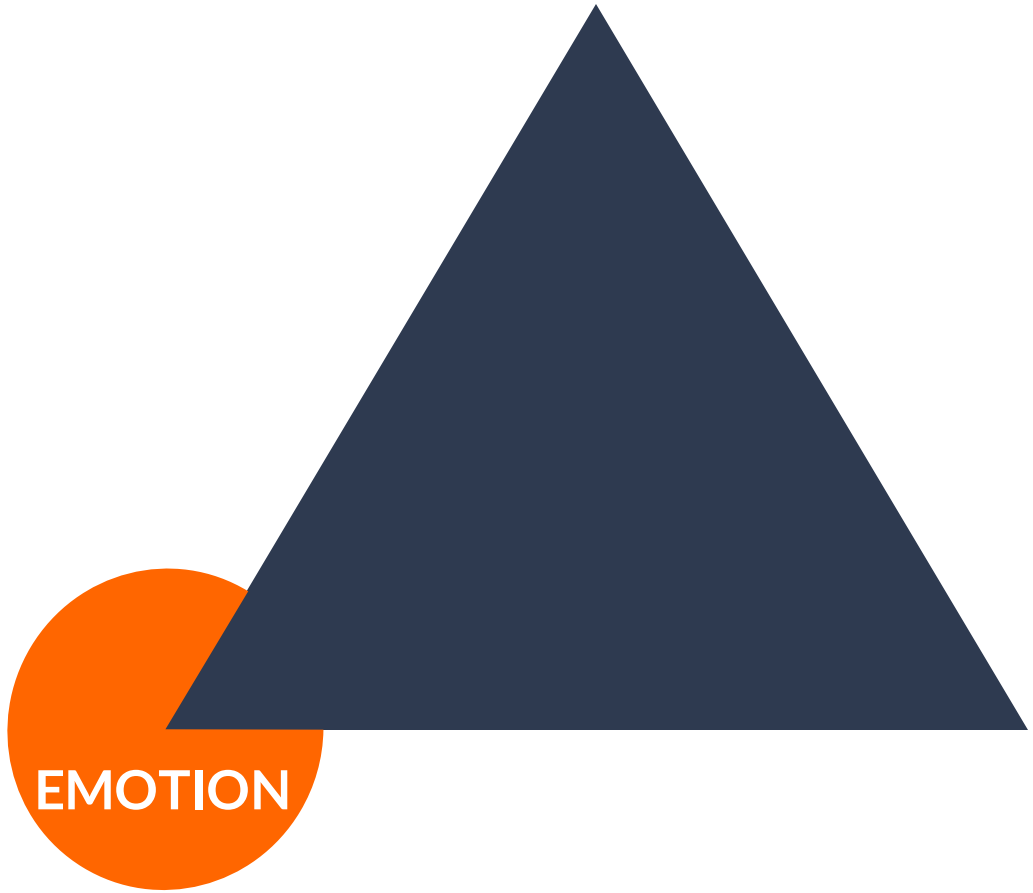
Familiarity and reasoning

- What domain knowledge and expertise does your audience have?
- How would you rate their data literacy and comfort with stats jargon?
- What kind of background will you need to provide?



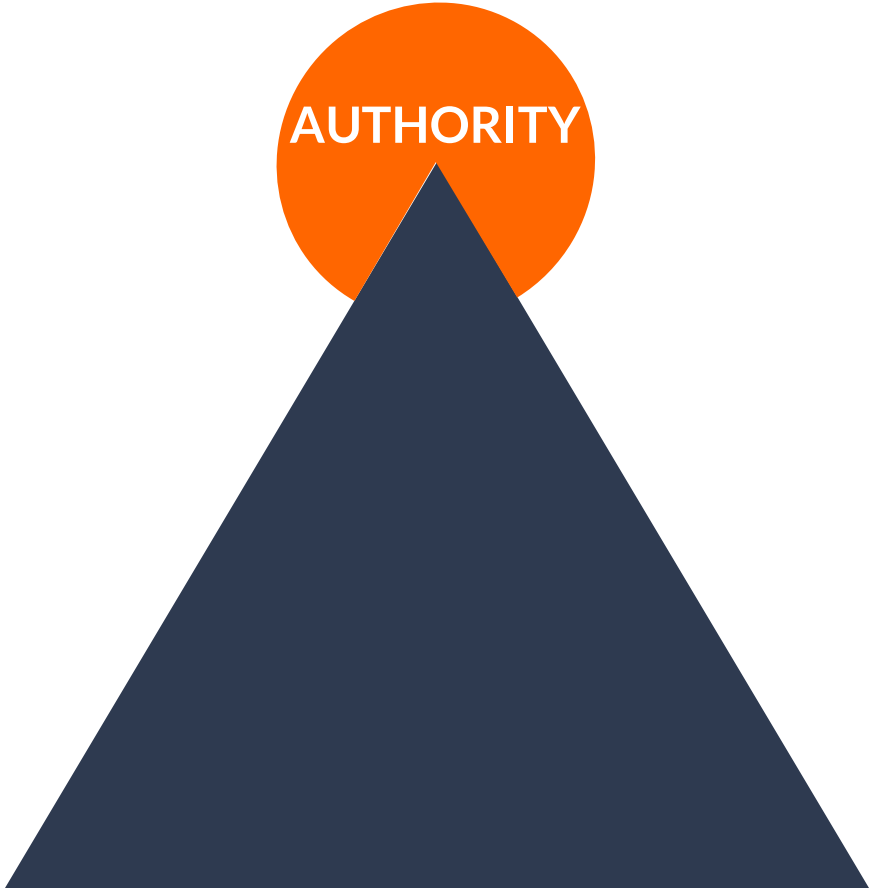
Beliefs and preferences

- What relevant preexisting assumptions and attitudes does your audience have?
- What data format does your audience like most?
- What values and myths will you need to be prepared for?

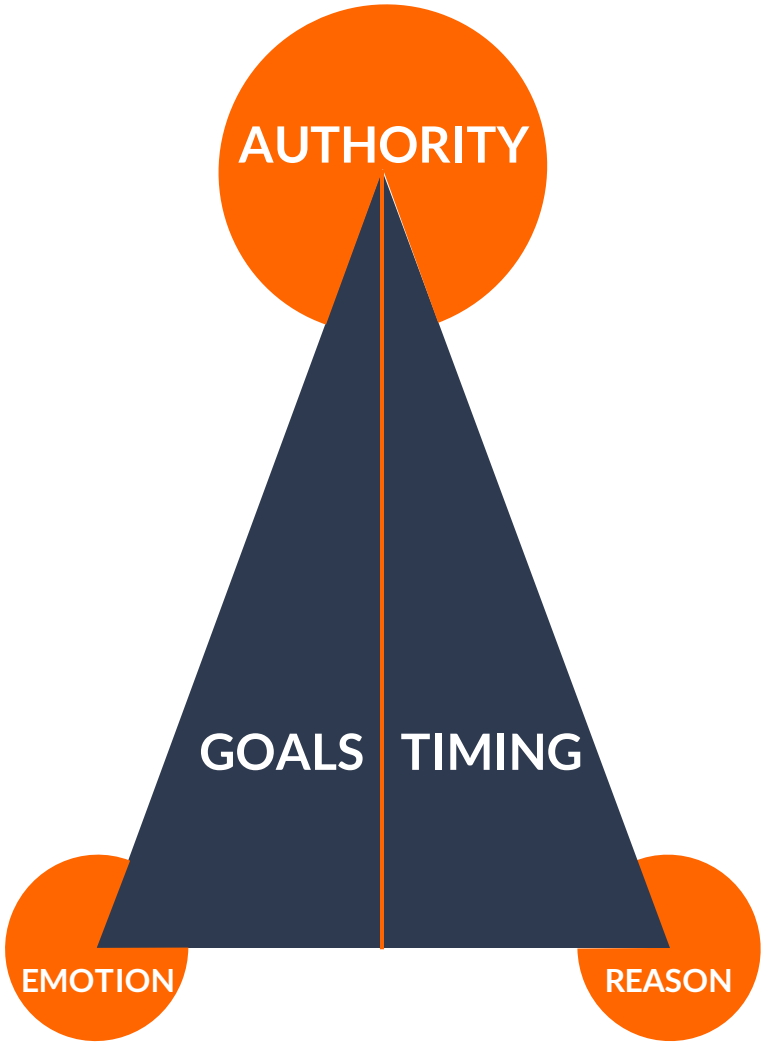


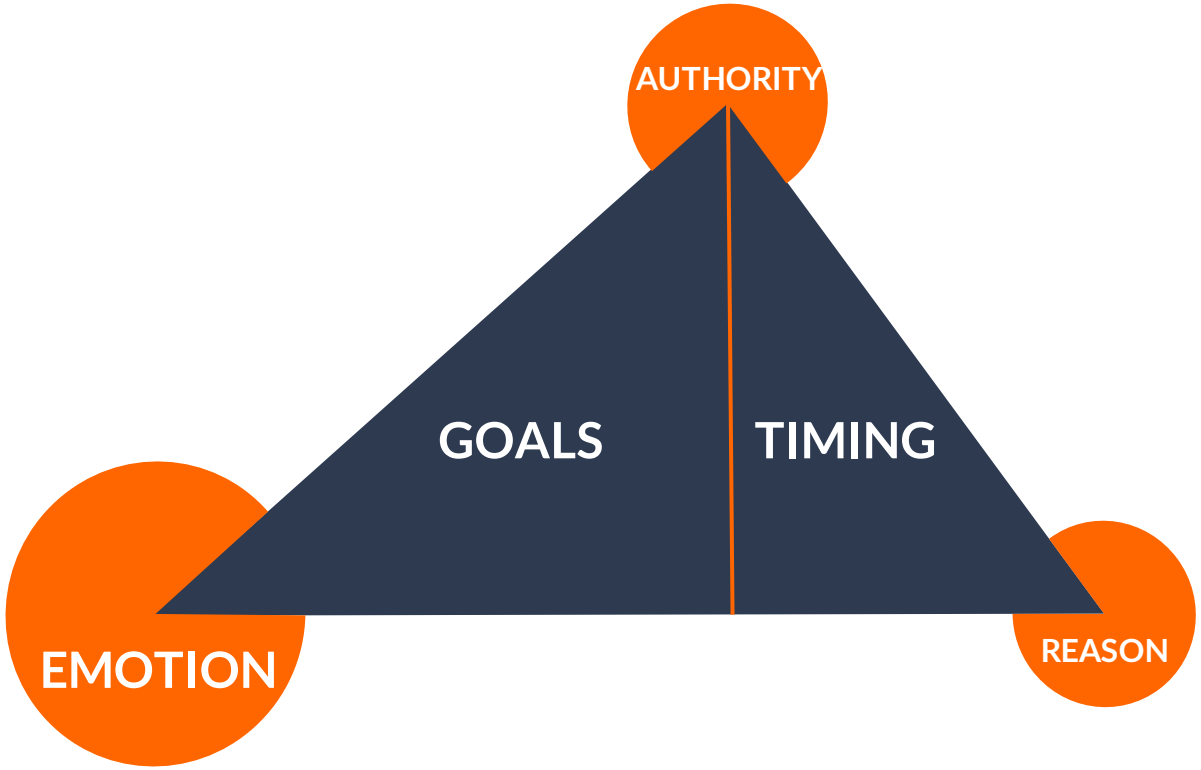
Expectations and authority

- What preconceived notions are your audience expecting you to validate?
- Is there a hierarchy to your audience that you must consider?
- What questions do you anticipate your audience will want answered?



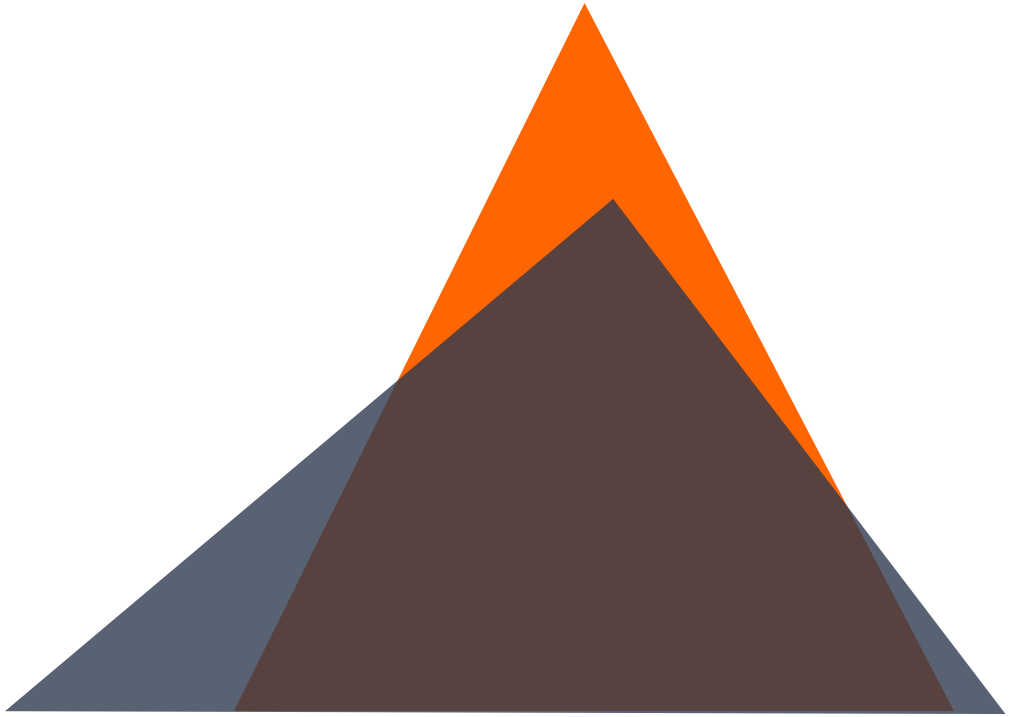
Different audiences, different needs





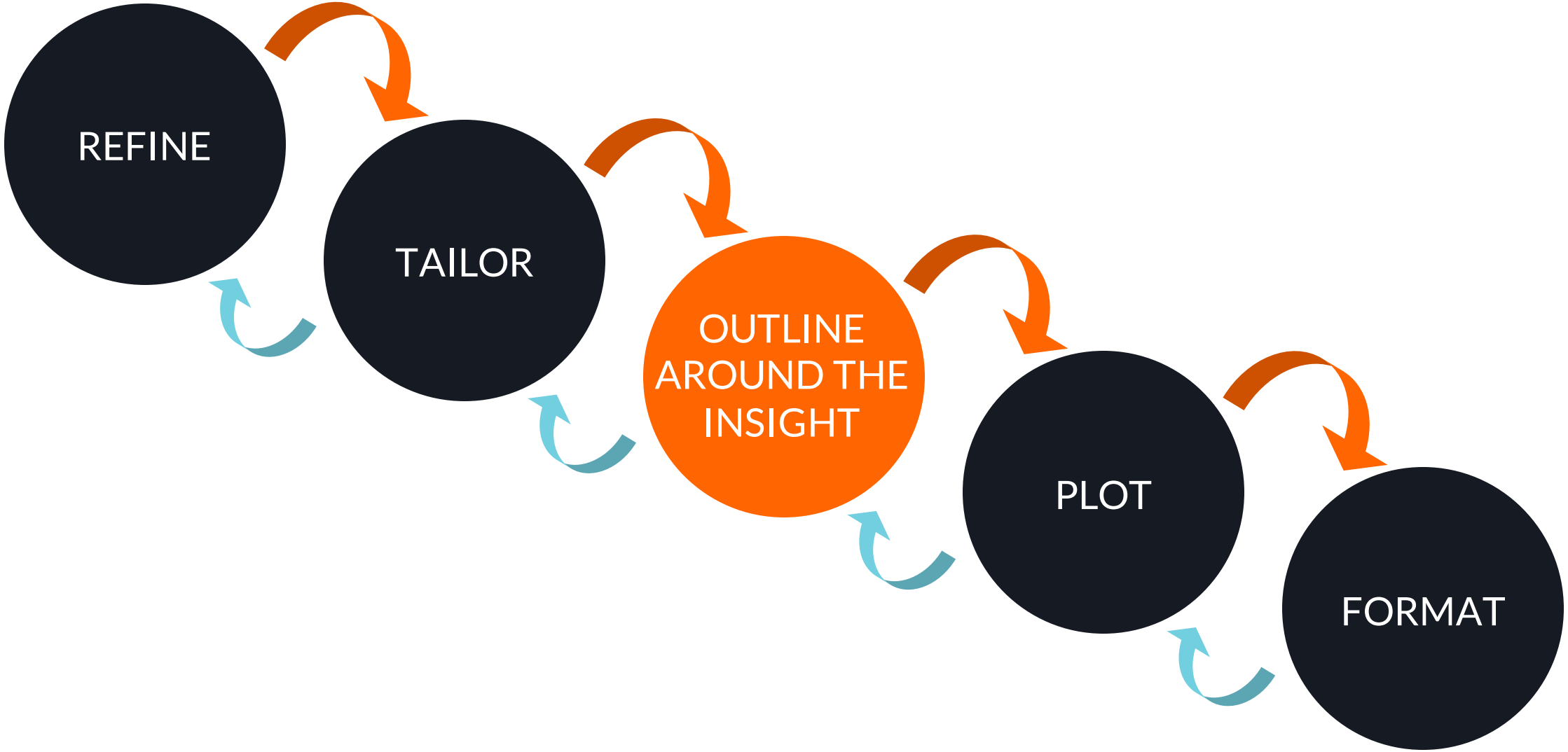
Diversity

- Is your audience comprised of a mix of different people or teams?
- Are their interests in concert, or are they competing?
- Can you accommodate the diversity, or do you need multiple versions of your story?

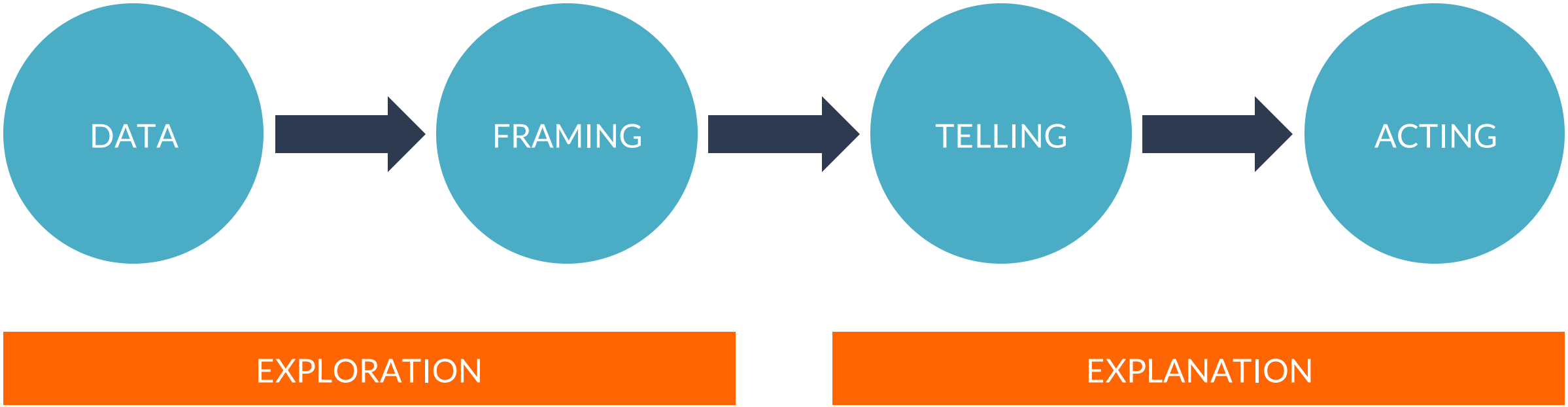


Outlining: from insight to outcome

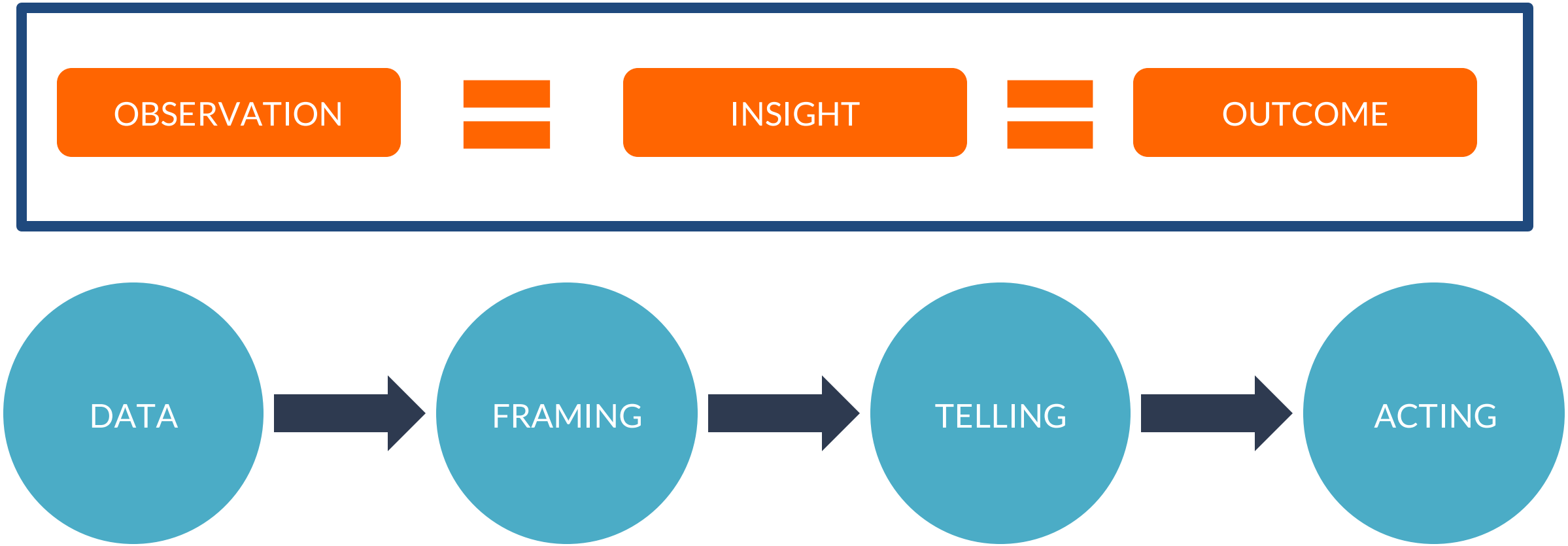
Storytelling step 3



The analytical process



Transforming the key insight



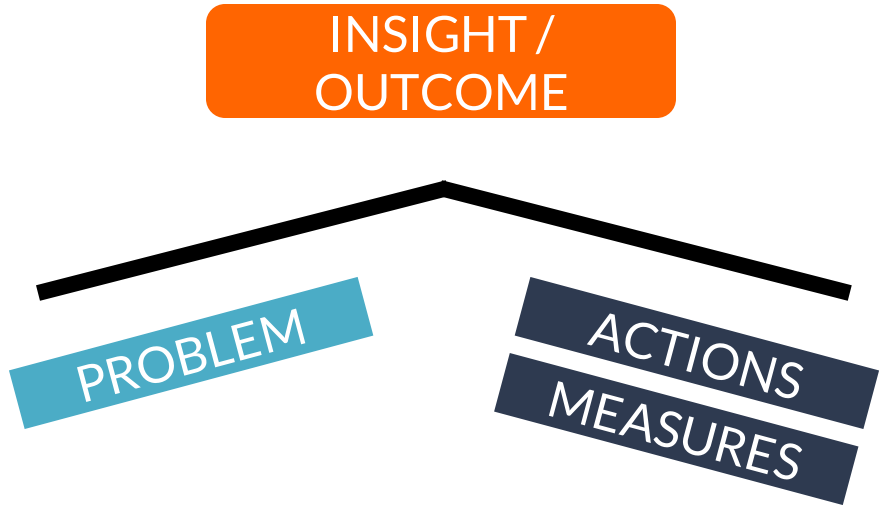
The “aha” moment, or climax

- Treating your key finding as both insight and outcome explains its central importance to your data story.
- An insight is the ***beginning*** of something, a recommended course of **action**.
- An outcome is the ***end*** of something, the diagnosis of a **problem**.

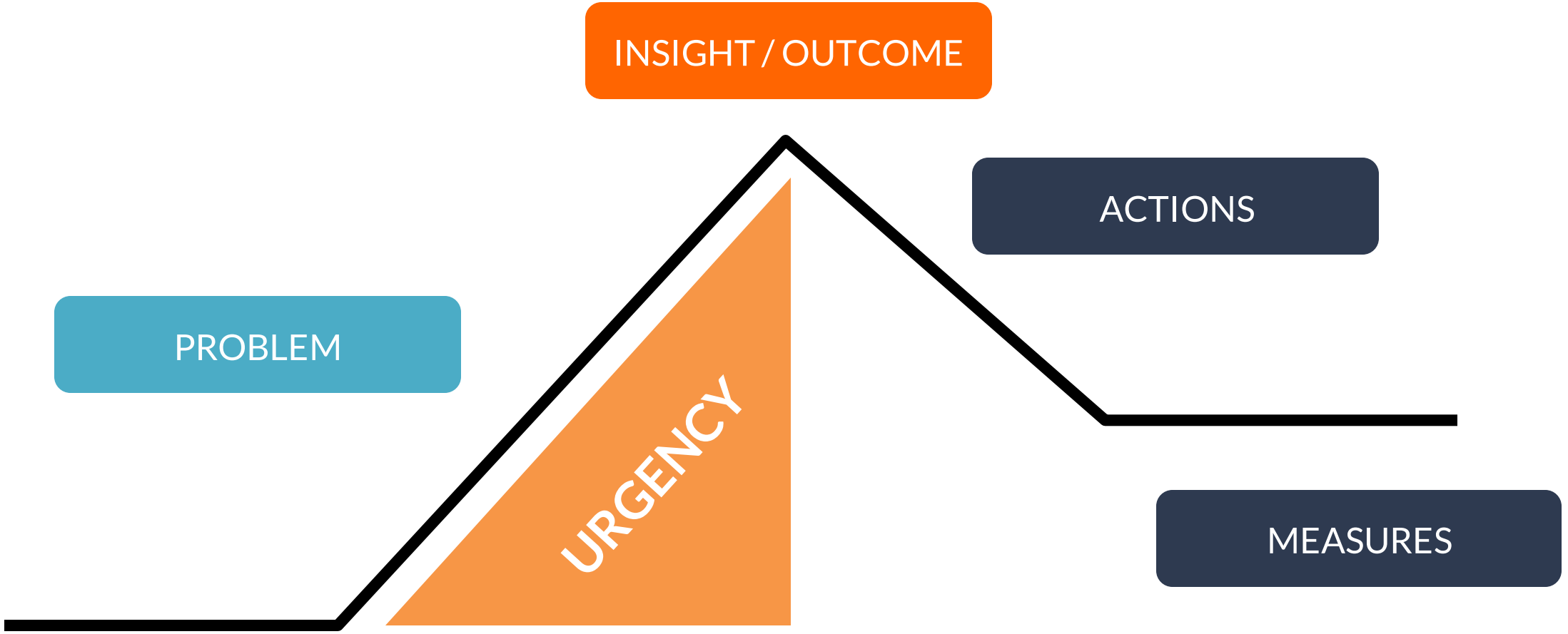
The rest of your story outline

Now that you know your audience, you can start telling them a story based on a few key parts:

- **Problem:** What challenge does your outcome most concretely represent?
- **Actions:** What can your audience do to fix the problem or change the outcome?
- **Measures:** How will your audience know they're achieving the desired outcome?



Putting the outline together



Activity: insight to outcome

- Turn to pages 22-25 of your participant guide to find **from insight to outcome**.
- This activity builds on the results of the user satisfaction survey about the **online banking app**, which we discussed on slides 185-192.
- Read the scenario and then write out how you would present your insights to the described audiences as an outcome. Then make some notes about how you would define the problem, actions, and measures for each audience.

Marketing team

- **Problem:** Even though many users don't actually use "email a representative" (feature D), those who do use it might be frustrated by its removal.
- **Outcome:** Users who do make use of the "email a representative" (feature D), should be directed toward "direct message a representative" (feature B), which performs many similar functions.
- **Actions:** Consider highlighting "direct message a representative" in some marketing materials. Create materials showing the removal of "email a representative" and its replacement. Gather more data on users of other features.
- **Measures:** Surveying user awareness of "email a representative"'s removal, as well as user awareness of other enhancements.

Head of technical support

- **Problem:** Removing "email a representative" (feature D), will cause some users to experience problems with product functionality.
- **Outcome:** Users who make use of "email a representative" (feature D), should be directed toward "direct message a representative" (feature B), which performs many similar functions.
- **Actions:** Update current documentation to redirect users of "email a representative" to "direct message a representative" instead. Create additional support documentation for "direct message a representative", including new enhancements.
- **Measures:** Gathering information for tech support interactions involving "direct message a representative," testing new "direct message a representative" documentation with current users of the "email a representative."

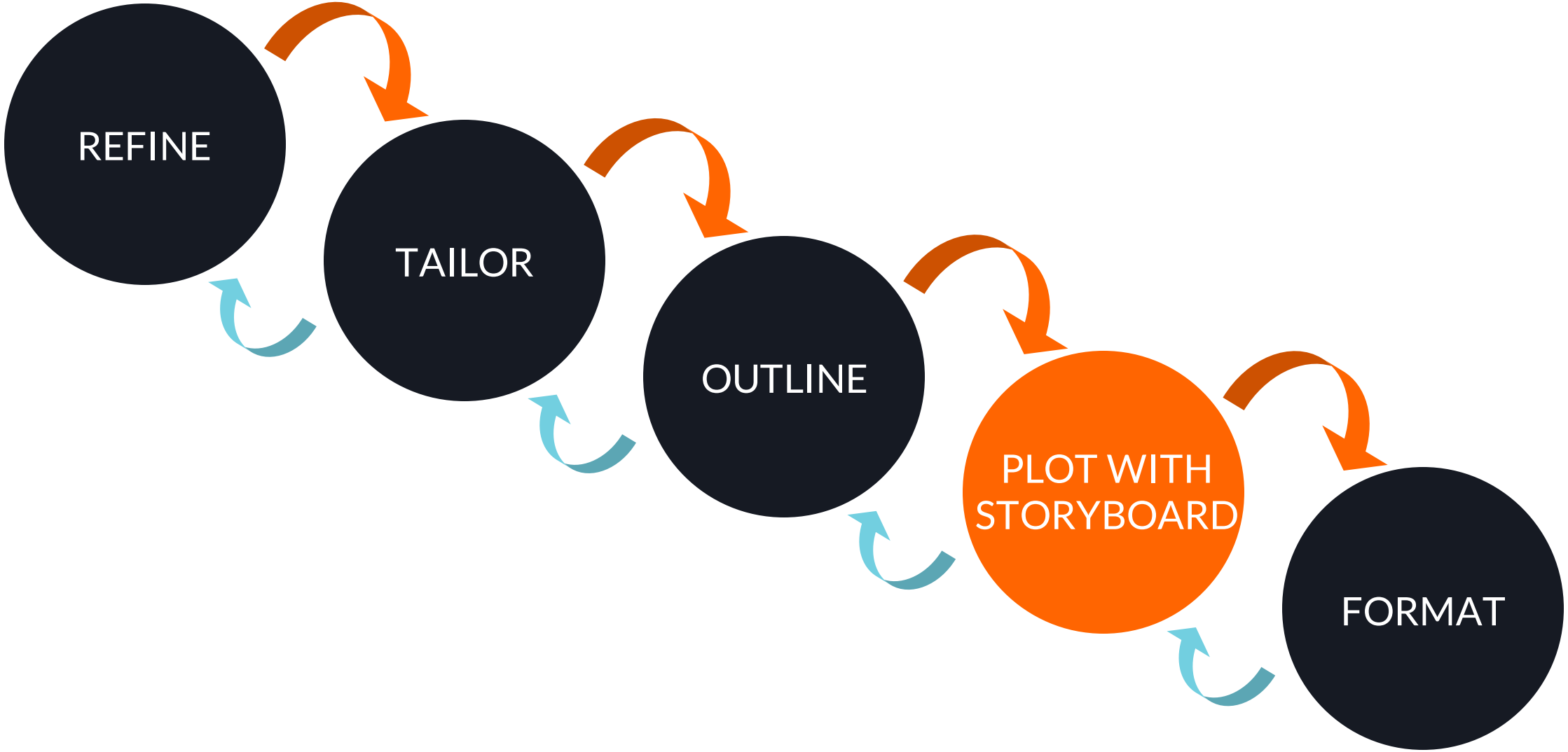
Recap

- Data stories:
  - are more memorable than stats or visualizations alone
  - are built around a clearly defined central insight
  - best compel an audience when they are tailored to its needs
  - set up a problem, reveal an outcome, then suggest actions and measures

Break

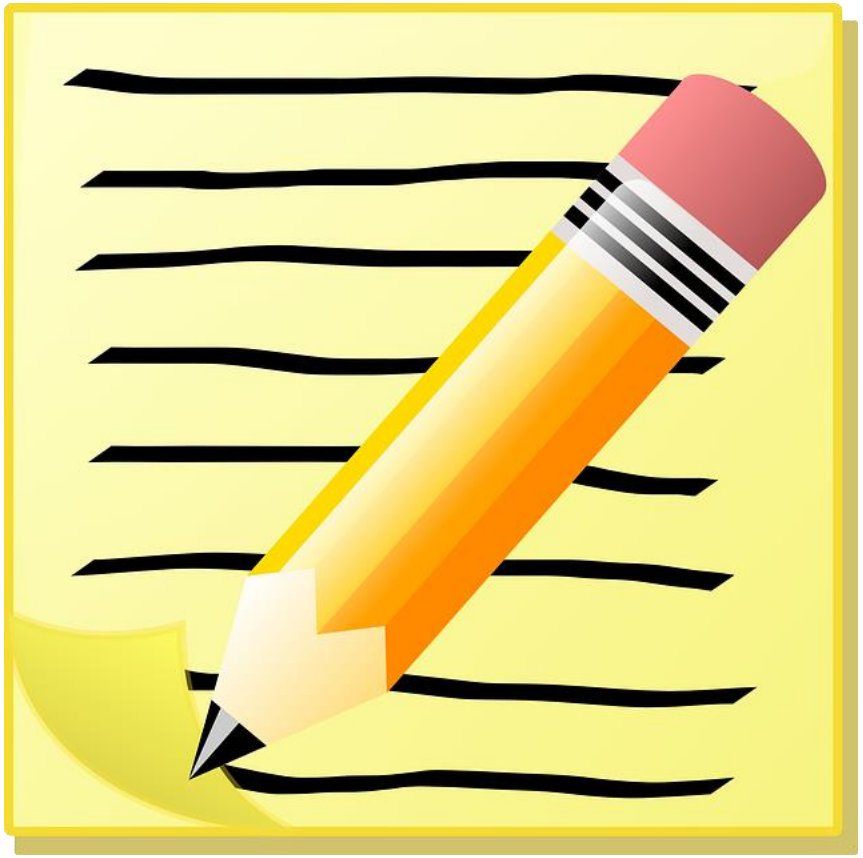
Plotting with a storyboard

Storytelling step 4



Low-tech storytelling

- It's okay for your data story to remain flexible at this early stage
- There are no right answers, only consideration and iteration
- Focus on building the elements of the story first, on paper
- Try out different versions quickly and don't get too attached



Step 1

Refine your insight

- By this point, your key takeaway should be succinct, tailored to the audience, and actionable.

- Position it at the ***climax*** of your data story. It's the most valuable and most relevant insight, and it's the "big reveal" your story is building toward.

Step 2

Define your hook

- Head down to the moment the urgency starts ramping up. This is the ***hook***, your initial problem.

- What concrete observations first pointed you toward the insight? Often, these come in the form of an *anomaly*, *shift*, or *change*.

Step 3

Prepare your hook with setting

- Take one step (but only one step) backward, to the very beginning of the story.

- What background information does your audience need to feel that hook sink in? What is the usual situation that the anomaly, shift, or change deviates from?

Step 4

Select and sequence beats

- Now it's time to get your audience from your hook all the way up to your insight – to make it seem like the appropriate *outcome* of your analysis.

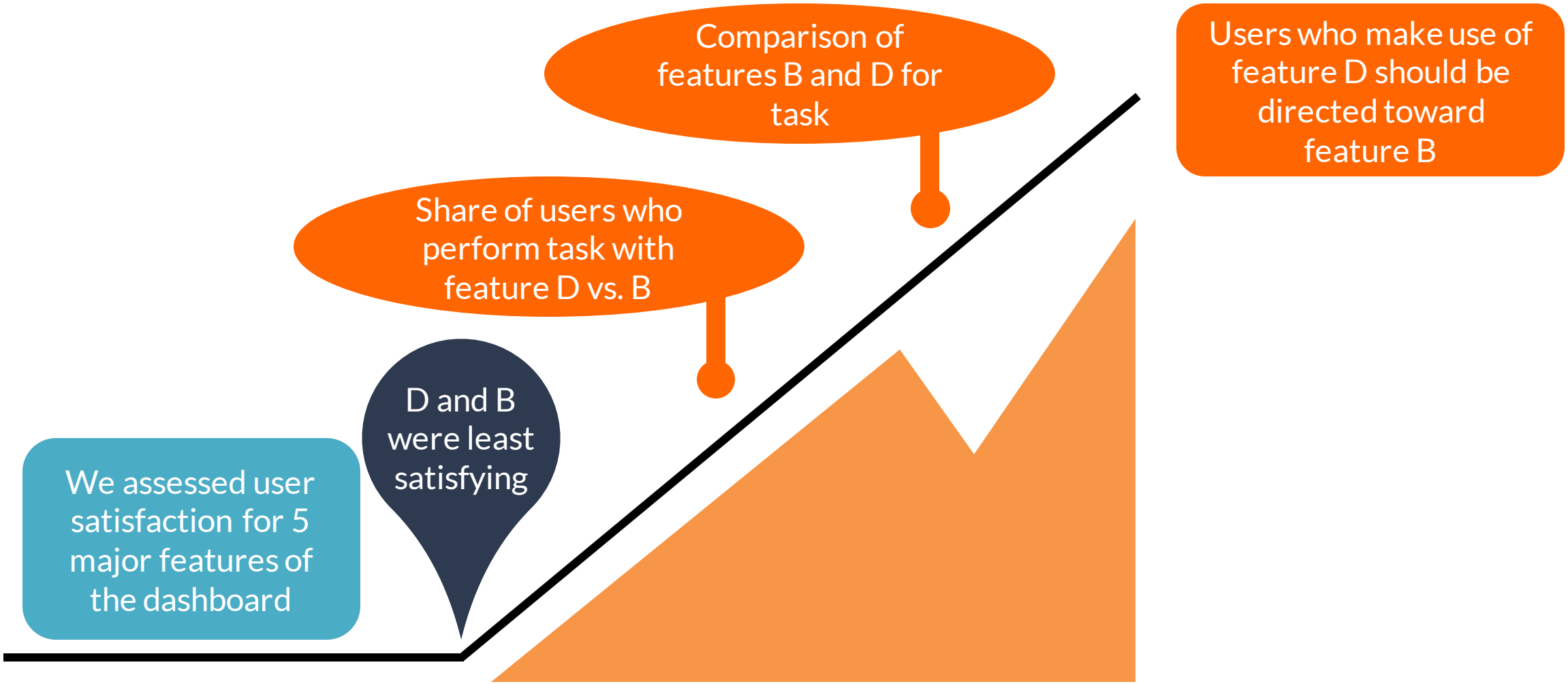
- Which data points will ***best build urgency***? ***In what order*** should you present them?

Beats = data points

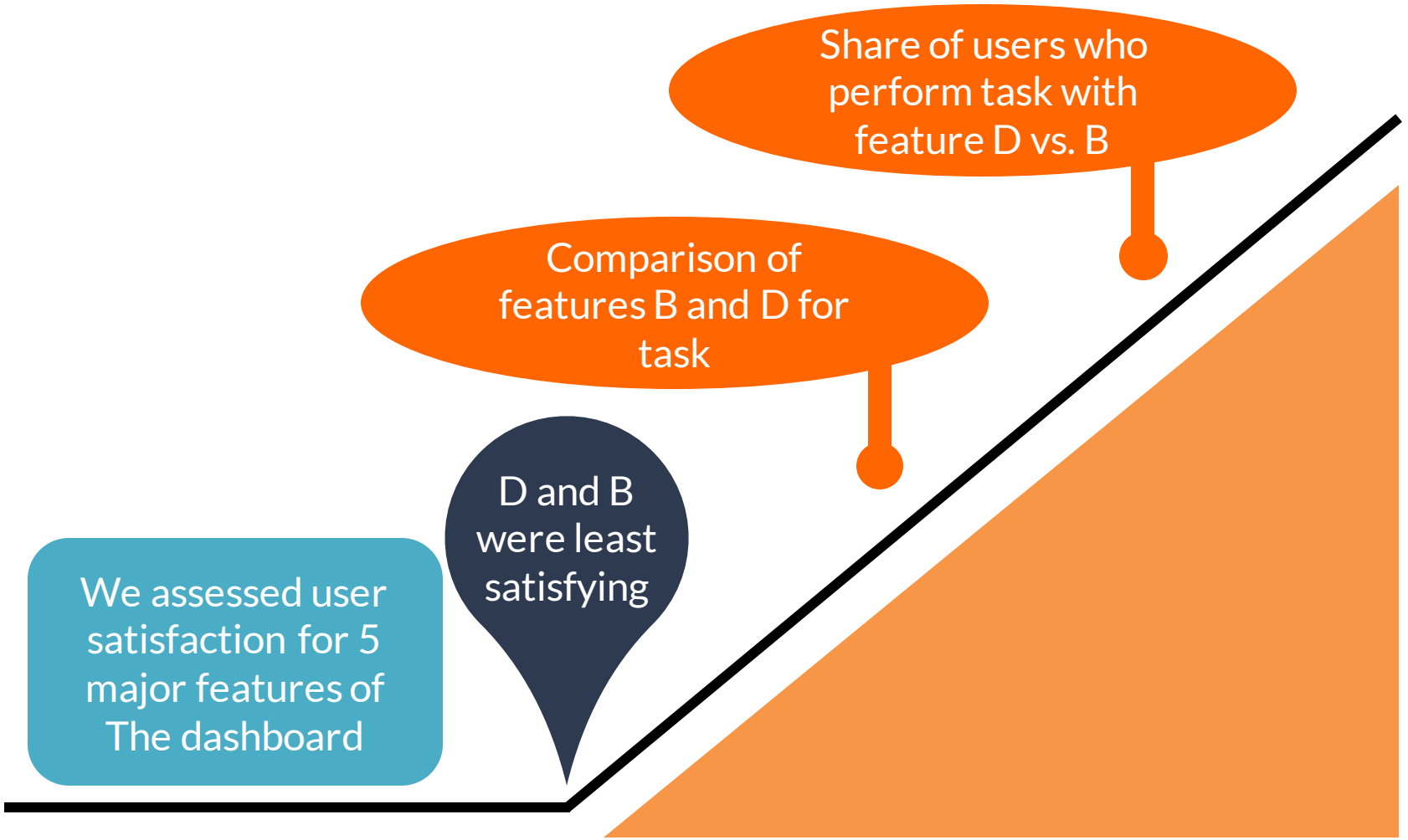
Comparison	Composition	Distribution	Relationship
Change over time Projection	Zoom out Drill down	Cluster Outlier	Intersection Projection

- When storyboarding, ***describe or sketch*** the visualizations and beats.
- Focus on selecting and sequencing the most relevant data points.

Ordering your beats



Reordering your beats



Users who make use of feature D should be directed toward feature B

It's more compelling to discuss the features first, then users.

The insight/outcome is focused on the users, not the features.

Step 5

Offer your course of action

- Good stories don't end on a climax. They resolve some tension by proposing a solution.

- What ***steps*** would you recommend taking? What ***options*** should be considered? What ***preference*** stands out? What ***measures*** seem relevant?

Step 6

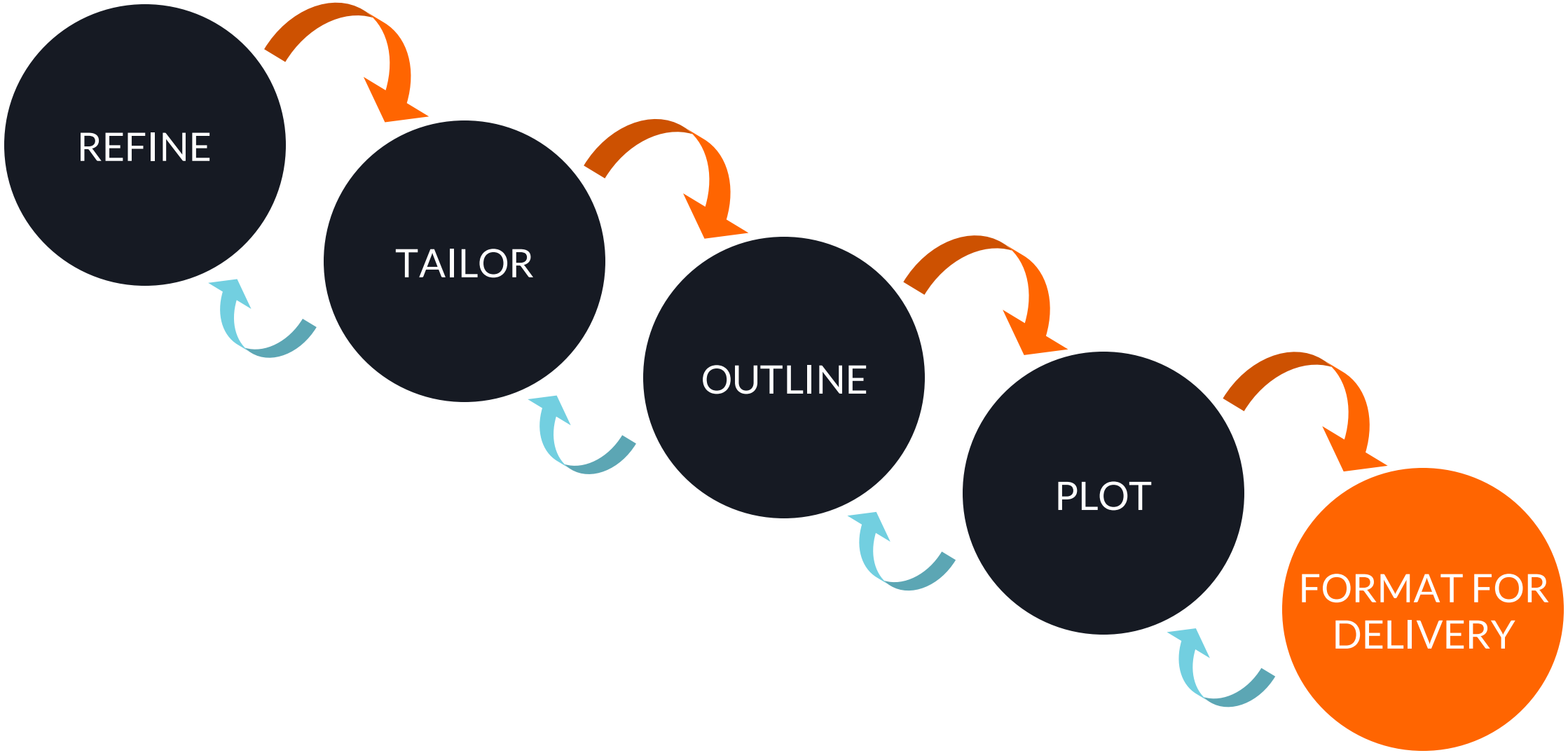
Format your data story

- You now have a complete data story! But it still doesn't have any visuals. **That's okay.**

- Before choosing visuals, you need to think about the overall communicative situation.

Formatting for delivery

Storytelling step 5



Direct vs. indirect communication

Being able to interact with your audience changes how you tell a data story.

- **Direct communication** affords you a high degree of flexibility and control over the story
  - adjust pace, drill down, pull back, respond to audience feedback in real time
- **Indirect communication** affords you a high degree of detail and text
  - annotate, explicate, provide as much context as your reader / viewer / listener needs
- These affordances are **inversely relational**

Data storytelling formats

Slide Deck	Document	Interactive	Hybrid
Sequence of slides intended for real-time presentation	Illustrated text (report, infographic) to be read anytime	Digital object intended to align function with user experience	Blend / compromise of at least two formats

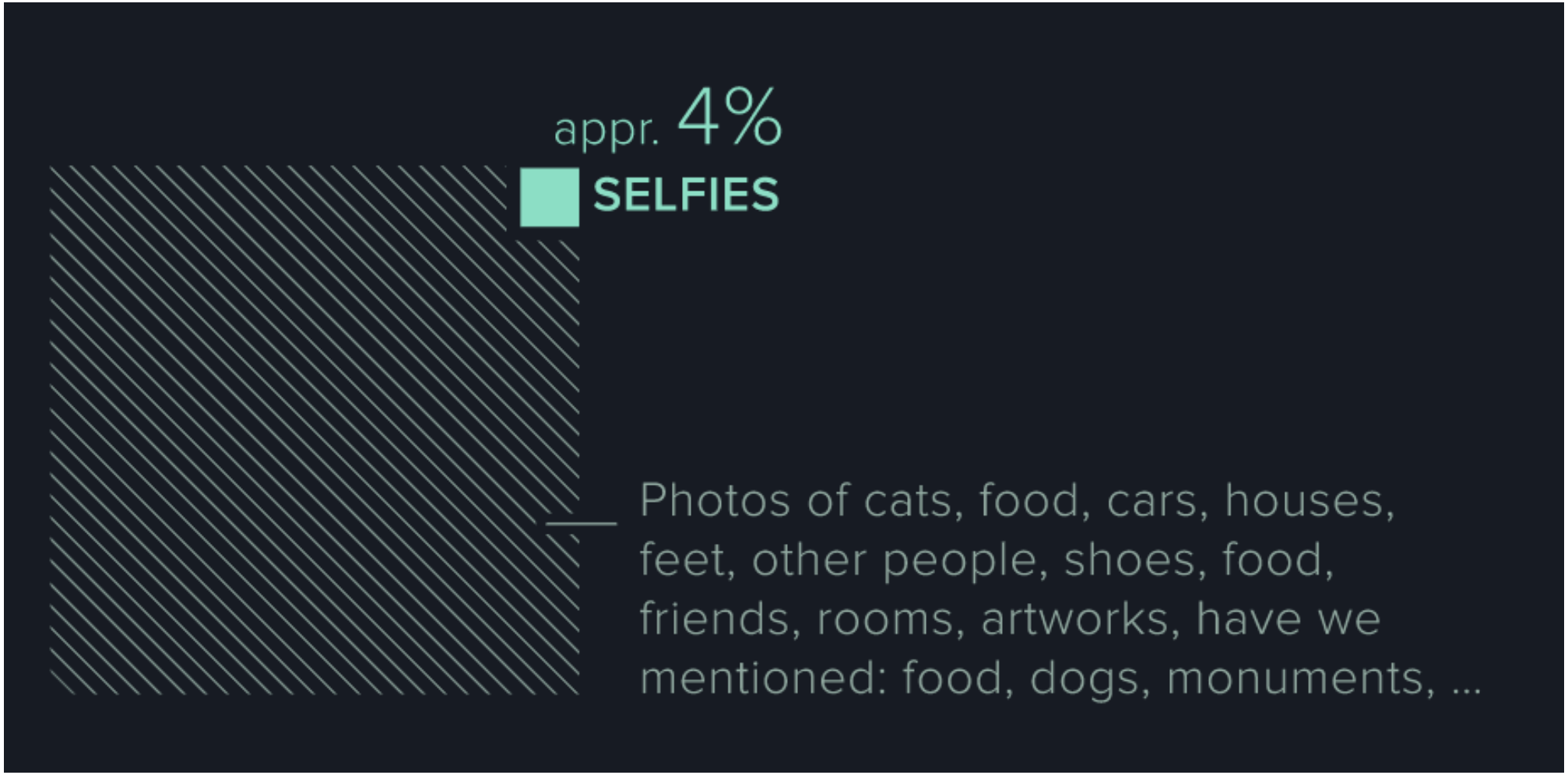
Formatting matters

- You may find yourself needing to alter the way you tell your data story based on the affordances of the format.
- Sometimes the format is a given, but other times, it will depend upon your input and the use case.
- The same kinds of considerations about choosing a visualization tool apply to selecting a format for your data story.
- As with visualizations, the simplest storytelling format is often the best.

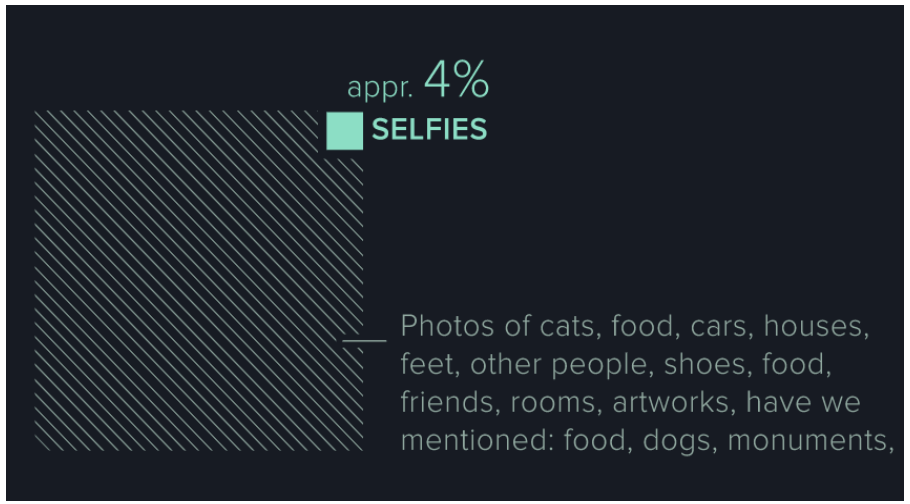
Same story, different formats

- For explanatory purposes, sample story formats drawn from a single source
- By comparing how similar data points are represented in different formats, we can better understand how format affects storytelling
- More information about Selfiecity can be found here

A sample slide

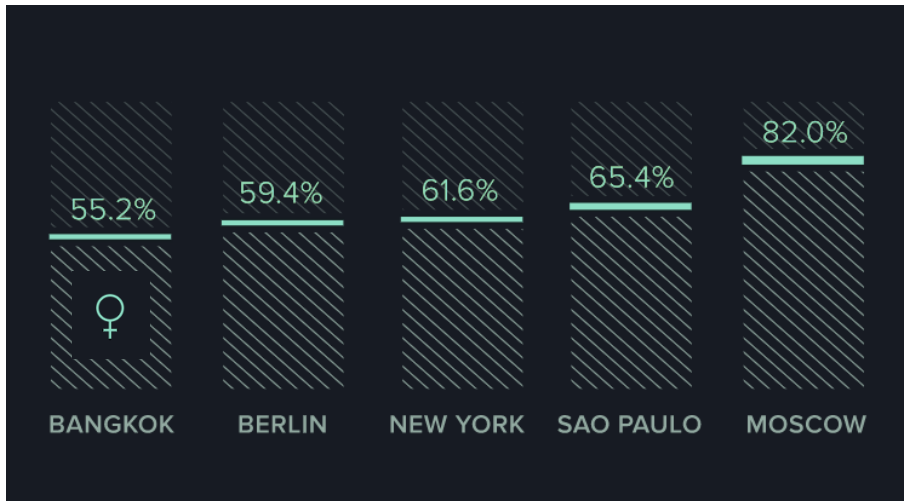


Sample slide 1, with script



- **Slide 1:** "Our first observation was surprising. We found that, despite stereotypes in the media, people take fewer selfies than is often assumed. Depending on the city, only 3-5% of the images we analyzed were actually selfies. The rest were photos of other objects: cats, food, cars, houses, feet, other people… You name it, we saw pictures of it, whether we wanted to or not."

Sample slide 2, with script



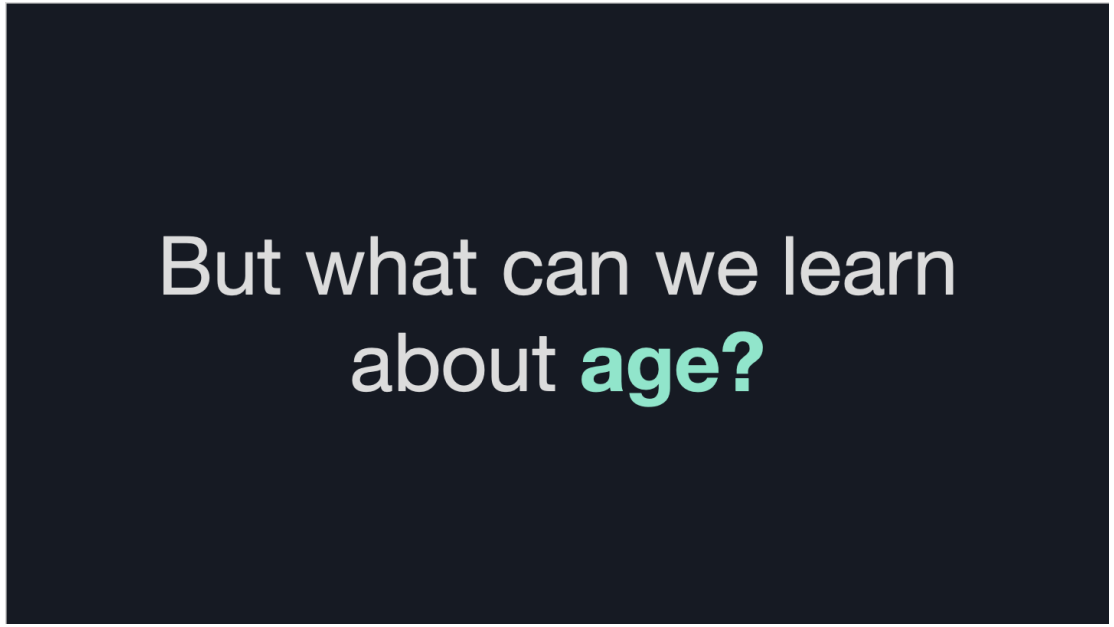
- **Slide 2:** "But in each of the five cities we analyzed, we found significantly more women taking selfies than men. There was quite a range here. You can see that women in Bangkok were 1.3 times as likely as men to take selfies, they were 1.9 times more likely in Berlin… all the way up to 4.6 times more likely in Moscow. That's about 5 selfies of women for every one of a man."

Slide deck

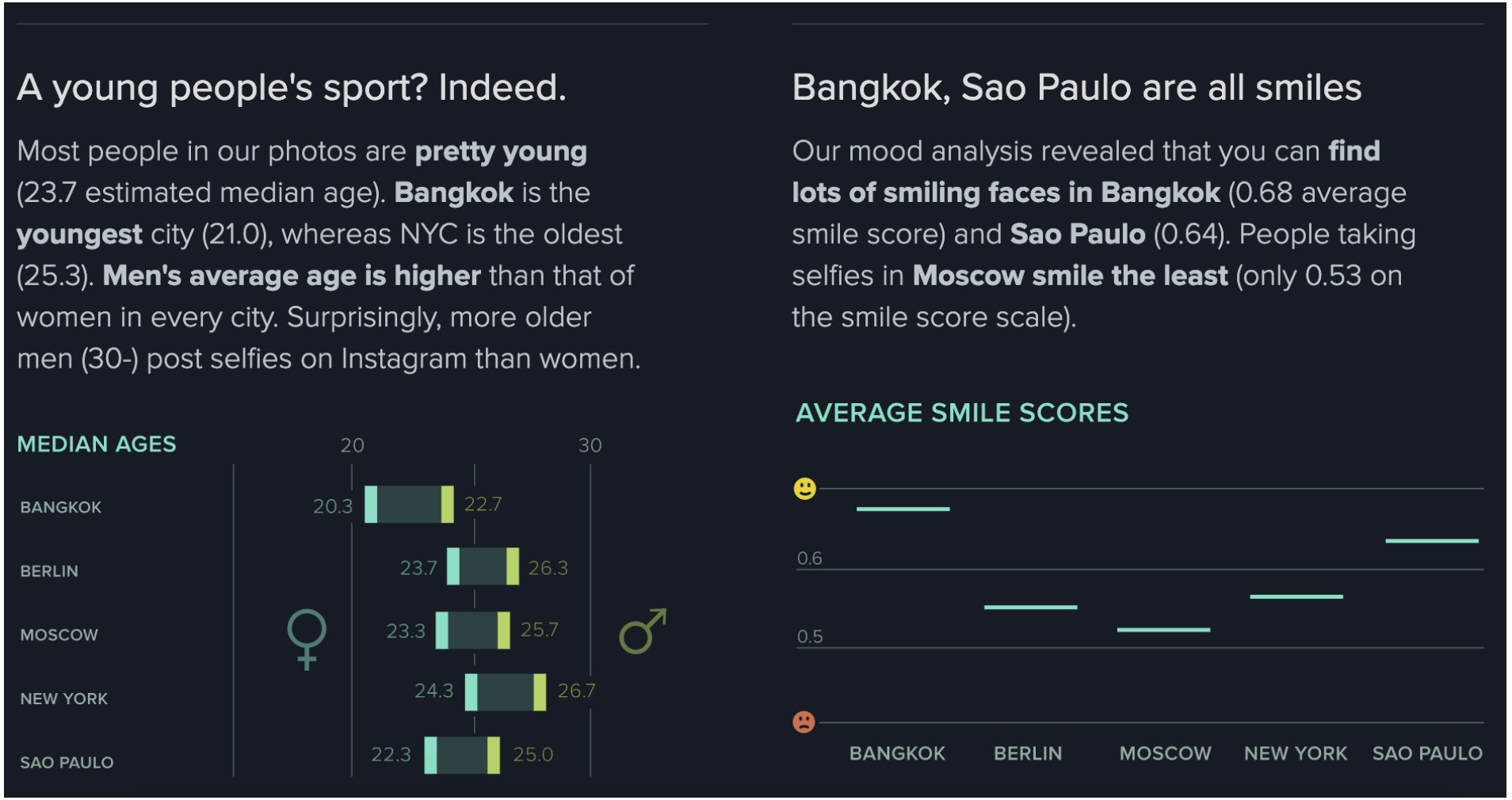
- Primarily visual medium, <50 words per slide
- Accompanying script provides explanation
- Presence of audience means negotiating pace, interruptions, digressions, and questions
- Build suspense across slides

Slide deck, ctd.

- Slide decks use **headers** and **section slides** to focus audience attention.
- Some slides may **only** contain a single line of text, or a single image, photo, or icon.

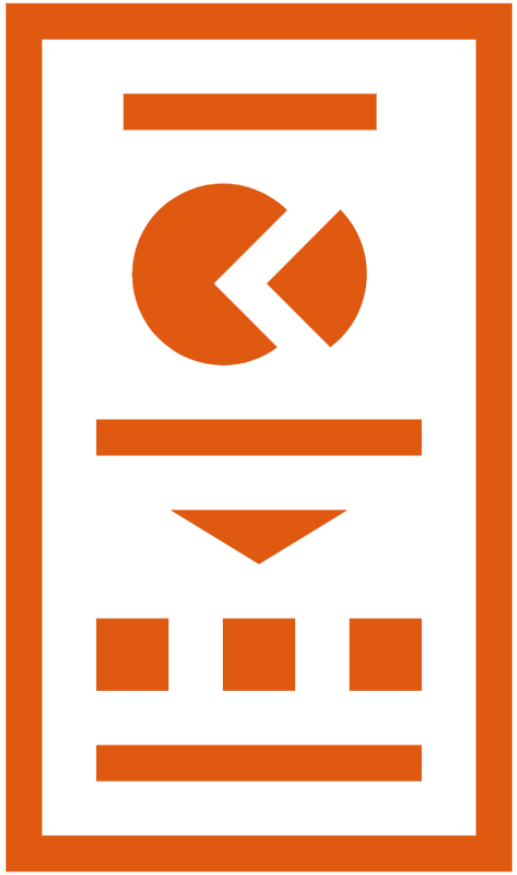


A sample infographic



Document - infographic

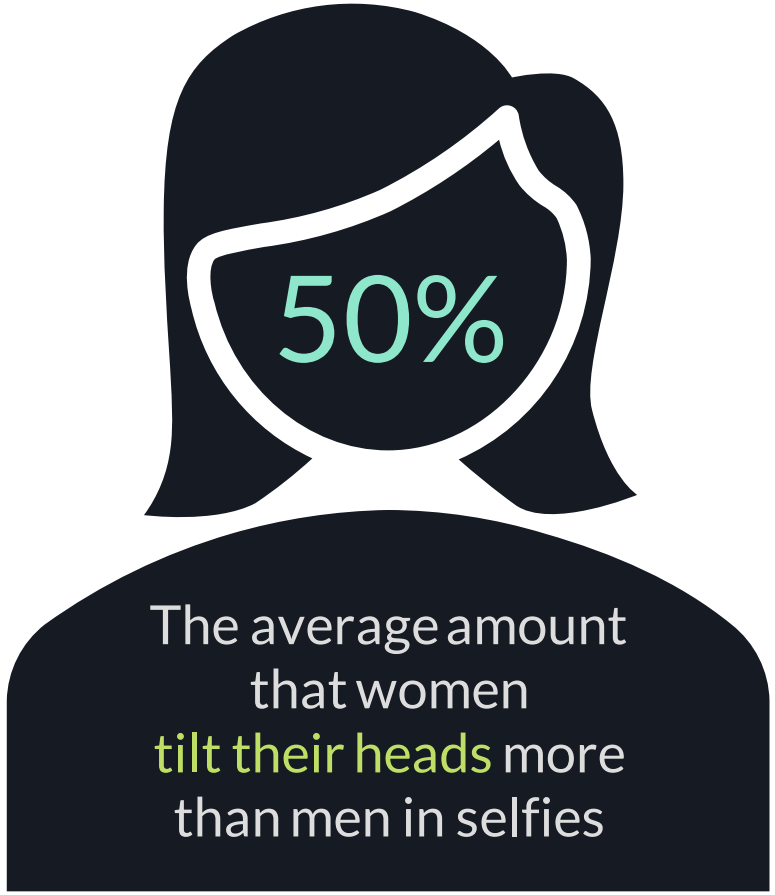
- Primarily visual medium, <50 words of explanation per visual
- Visuals feature easily interpretable callouts, icons, titles, labels, and annotations
- Text itself is treated as a visual
- Key takeaways at top, but build urgency vertically / downward



Document – infographic, ctd.

- Infographics use **section headers, size, shape, and color scheme** to focus attention

- Some beats may **only** consist of text, an icon, or an image, rather than a chart or graph.

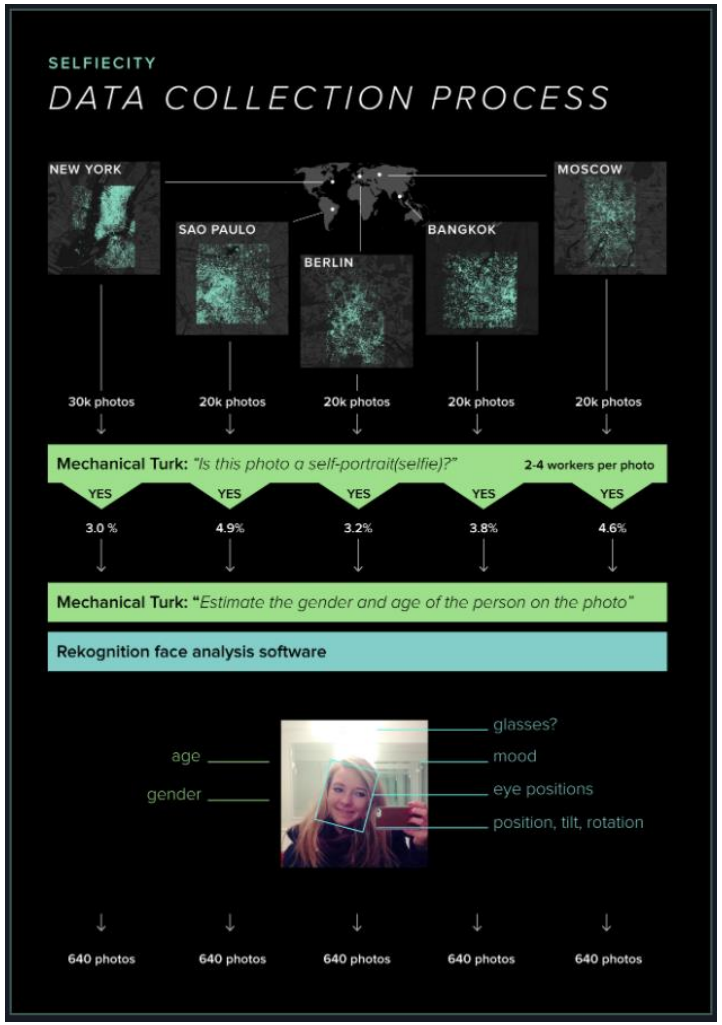


A sample report

HOW WE COLLECTED AND FILTERED THE DATA

To locate selfies photos, we randomly selected 120,000 photos (20,000-30,000 photos per city) from a total of 656'000 images we collected on Instagram. 2-4 Amazon's Mechanical Turk workers tagged each photo. For these, we asked Mechanical Turk workers the simple question "Does this photo shows a single selfie"?

We then selected top 1000 photos for each city (i.e., photos which at least 2 workers tagged as a single person selfie).



We submitted these photos to Mechanical Turk again, asking three "master workers" (i.e. more skilled workers) not only to verify that a photo shows a single selfie, but also to guess the age and gender of the person.

On the resulting set of selfie images, we ran automatic face analysis, supplying us with algorithmic estimations of eye, nose and mouth positions, the degrees of different emotional expressions, etc.

As the final step, one or two members of the project team examined all these photos manually. While most photos were tagged correctly, we found some mistakes. We wanted to keep the data size the same (to make visualizations comparable), so our final set contains 640 selfie photos for every city.

Document – report

- Primarily textual medium, high word count blocked out as paragraphs
- Visuals can have greater informational density – if they're thoroughly explained
- Visuals function as illustrations of the surrounding explanatory text
- Begin with summary, then sequence of sections, then conclusion

Document – report, ctd.

- Reports use **a table of contents, titles, section names, and headers** to focus attention

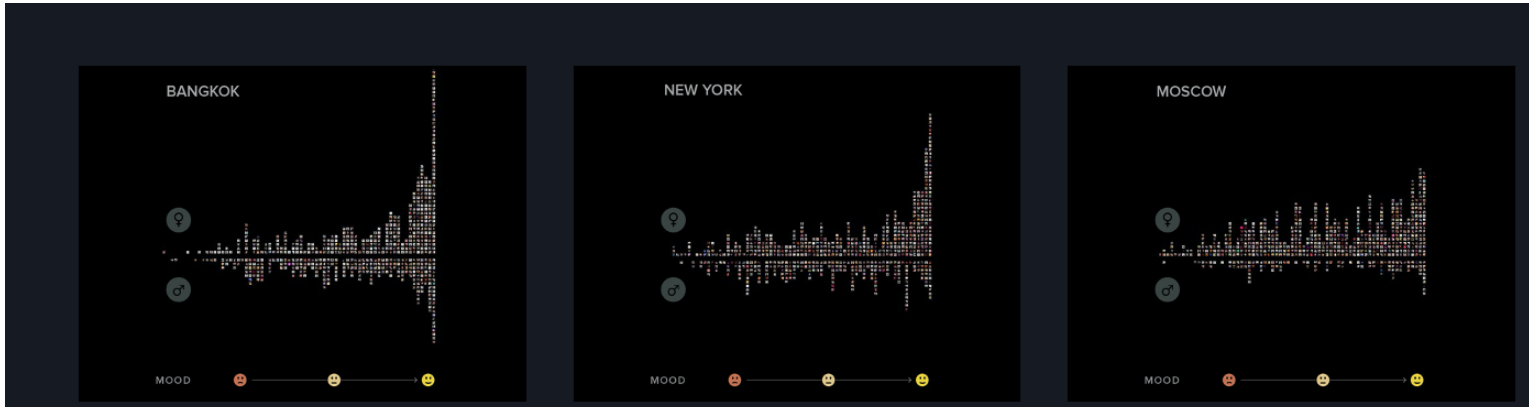
- Individual images may also be given **captions** to more succinctly restate their purpose than the surrounding text

WHAT CAN SELFIES SHOW US?

1. How many photos are selfies?
2. Do men or women take more selfies?
3. How old are selfie-takers?
4. How many selfies display smiles?
5. What methodology did we use?

Hybrid

a.k.a. slides without a show, or a visual report

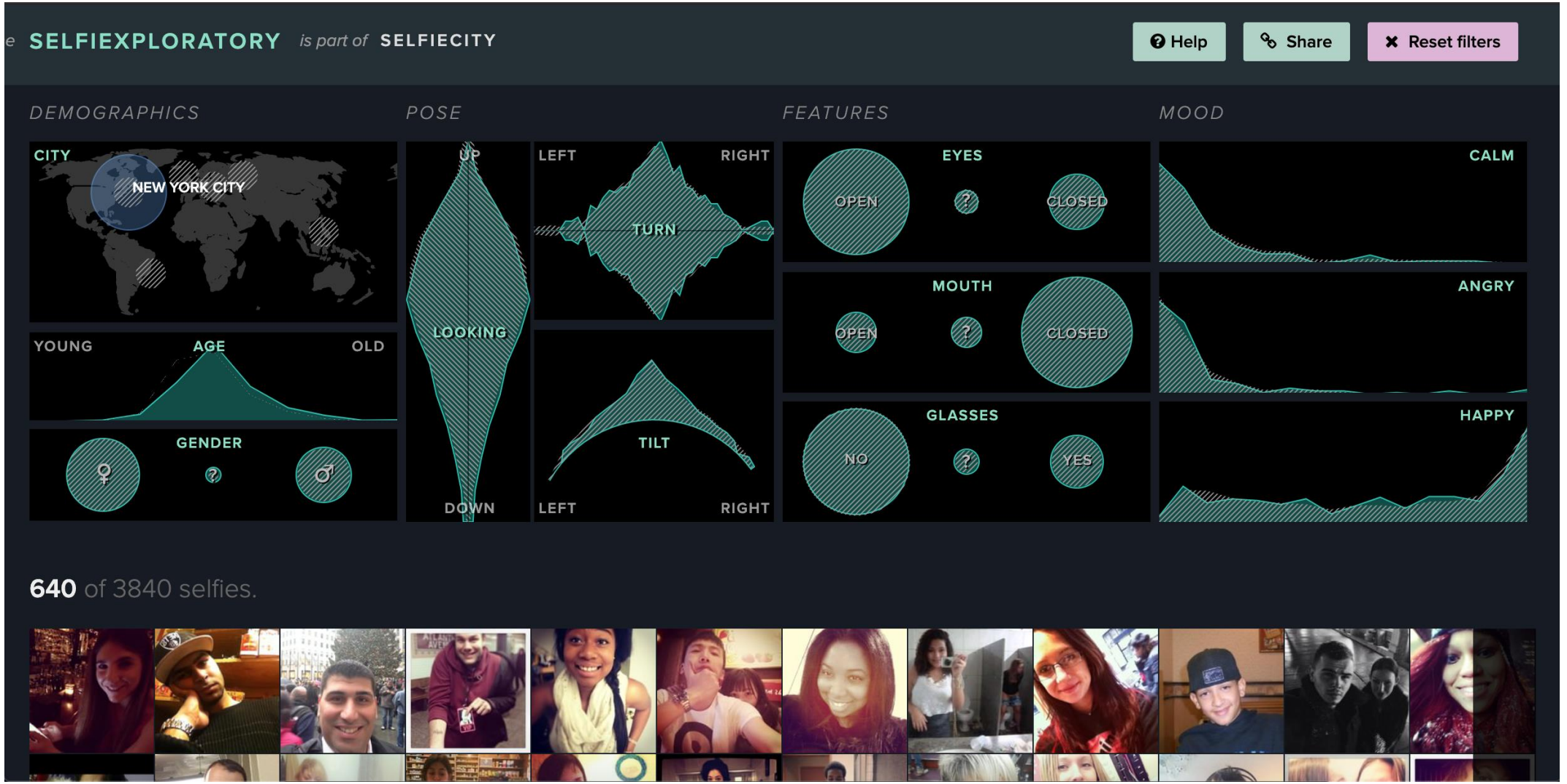


One of the other elements we researched was the **mood** of each selfie subject, sub-classified according to gender.

Out of the five cities whose selfies we studied, **Bangkok residents appeared to be the happiest**, or at least to smile most frequently in photos irrespective of gender.

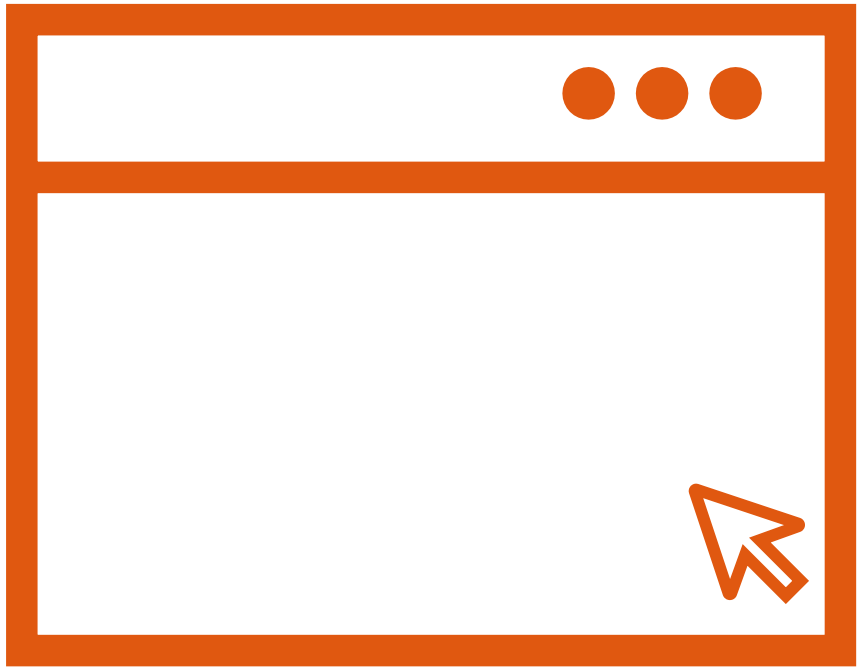
Residents of Moscow appeared to be the least happy, not because that many more unhappy people appeared in photos, but because of a much wider distribution of neutral expressions throughout the sample.

A sample interactive



Interactive

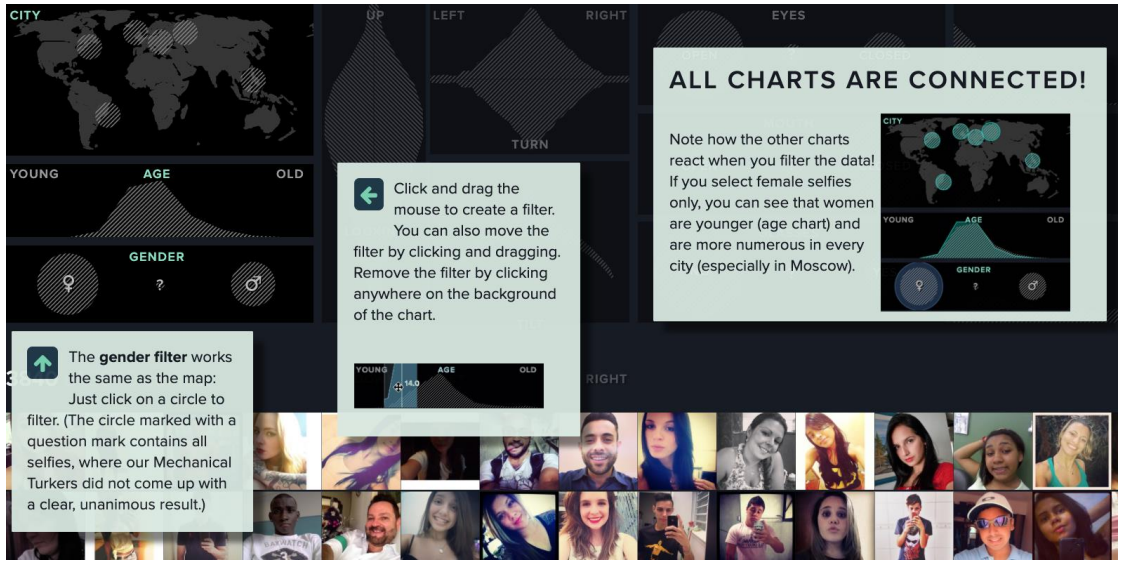
- Visual and textual medium
- Visuals and text may be dynamic rather than static
- Highly variable experience based on user choice, rather than a pre-defined narrative
- Takeaways are revealed as discoveries through manipulation



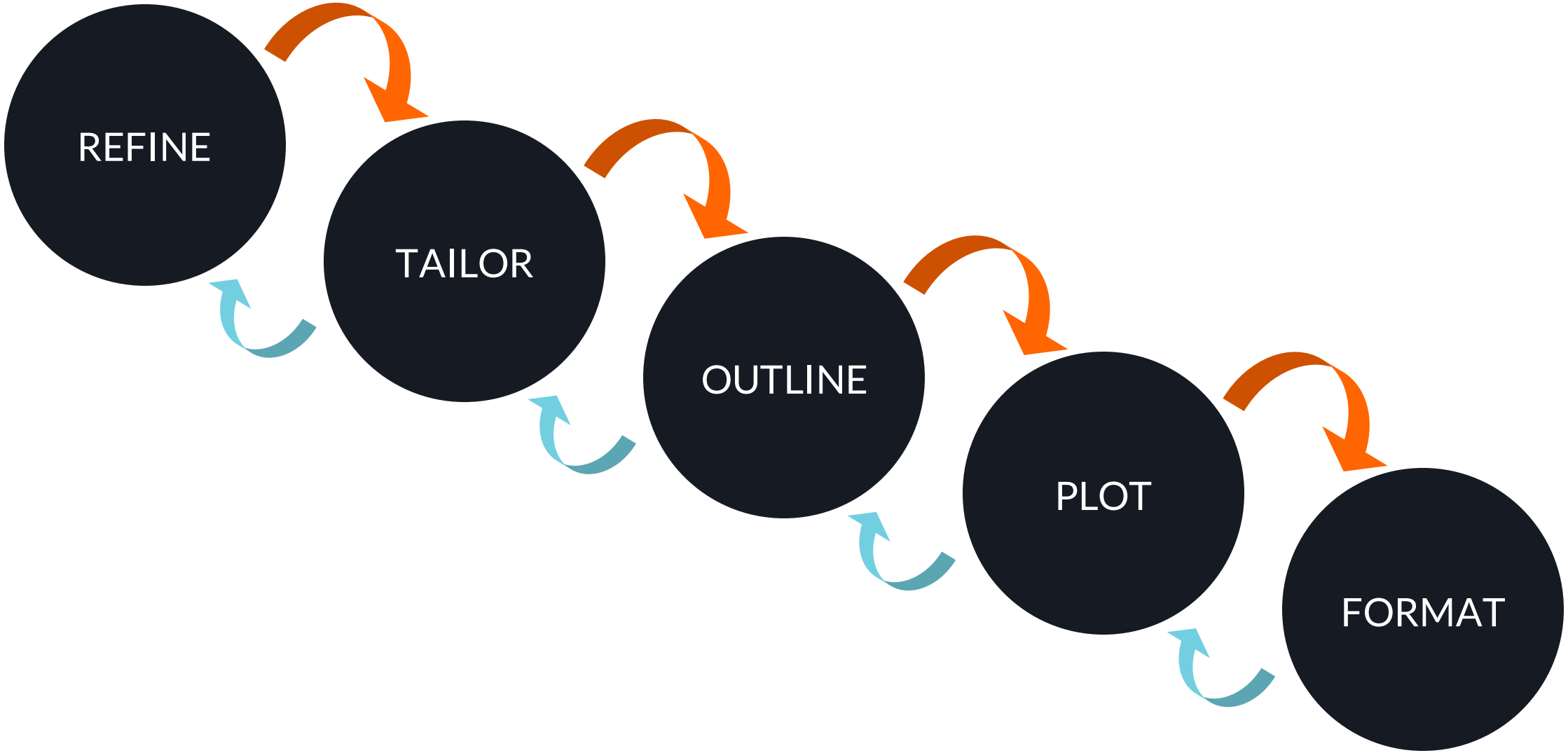
Interactive, ctd.

- Interactives use **filters, clickable objects, scalable objects, and movement / responsive design** to focus attention

- Building a digital tool demands attention to navigation, graphic design and layout, informative text, security, and accessibility



The complete data story



Recap

- Low-tech storyboarding lets you shape your story before sinking time into visualizations
- Data stories have an orderly sequence of parts that help increase tension
- Different formats will affect the shape of the story and the relationship of image and text

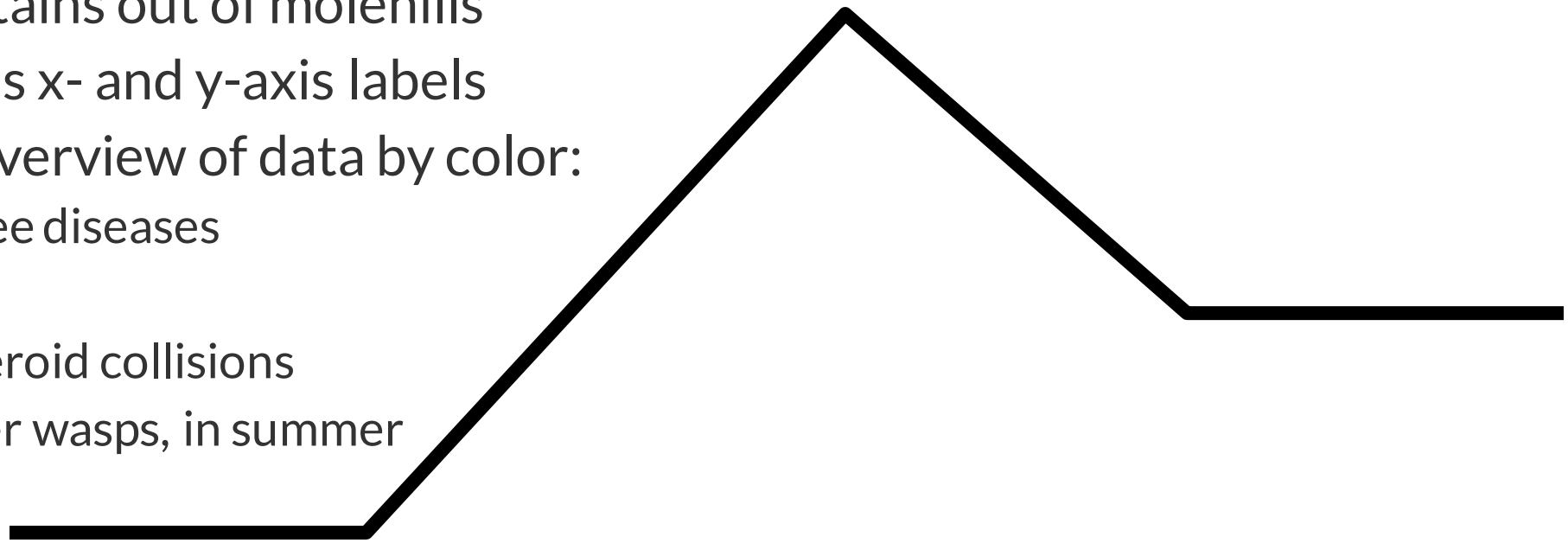
Activity: reverse storyboarding

- Turn to pages 26-27 of your participant guide to find **reverse storyboarding**.
- **To complete this activity, you will need to watch a clip of this video from 2'46" to 6'16"** (a clickable link is also available in your participant guide).
- Begin this activity by watching the video clip through. Be aware of the different appeals to reason, emotion, and authority that the presenter uses. Note how the audience reacts to different parts of the story.
- Then, watch the video a second time. While watching, fill in the blanks on the diagram in your participant guide.

Activity: reverse storyboarding

SETTING

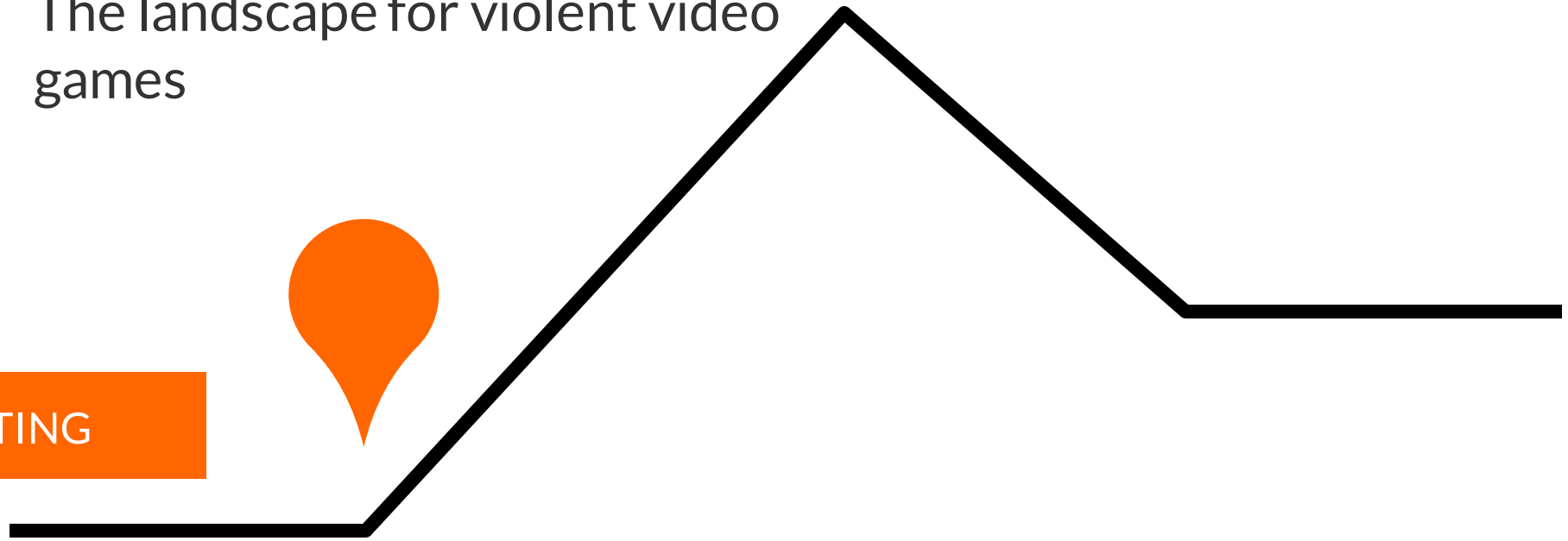
- Landscape of world's fears
- Title already suggests problem: "Mountains out of molehills"
- Explains x- and y-axis labels
- Basic overview of data by color:
  - Three diseases
  - Y2K
  - Asteroid collisions
  - Killer wasps, in summer



Activity: reverse storyboarding

HOOK

- “I love being a data detective”
- “An interesting and odd pattern”
- The landscape for violent video games

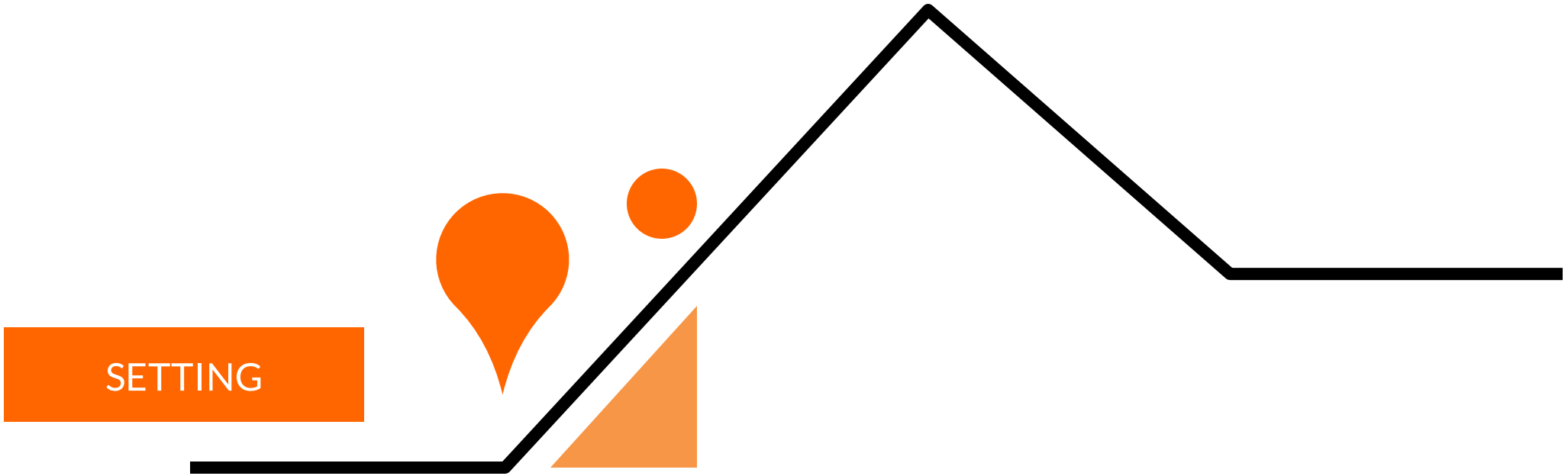


SETTING

Activity: reverse storyboarding

BEAT 1

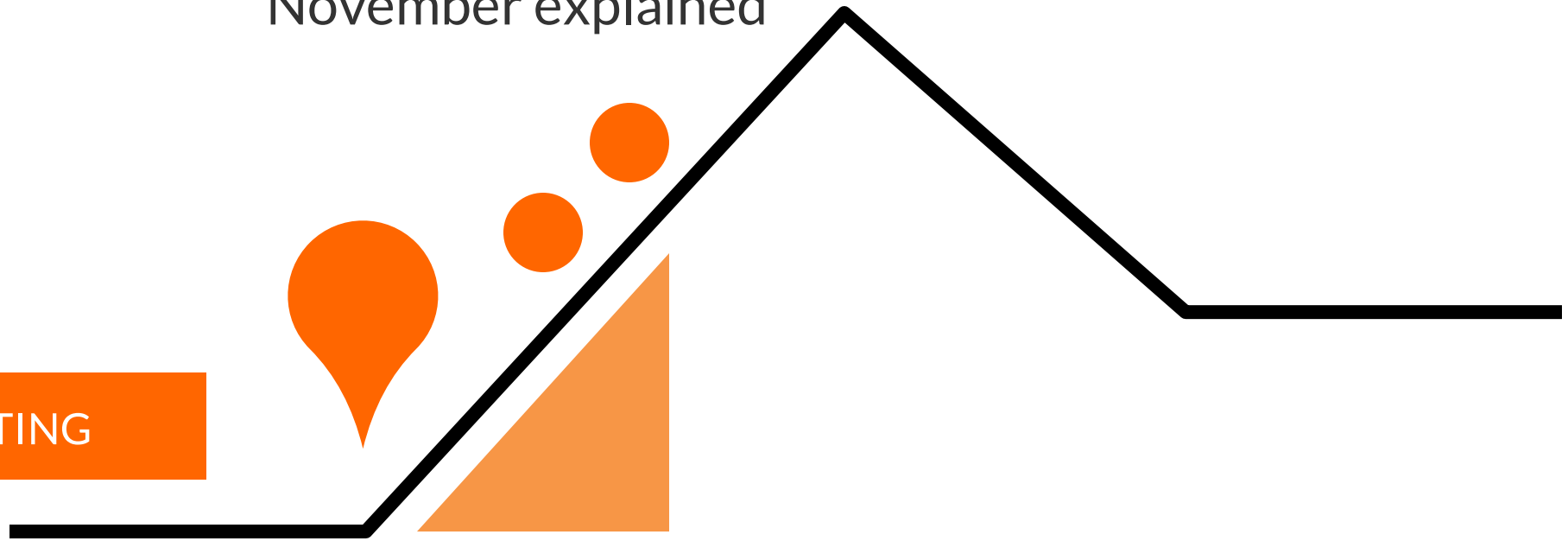
- “odd, regular pattern in the data, twin peaks every year”



Activity: reverse storyboarding

BEAT 2

- "same month, every year"
- April / November labels, plus November explained

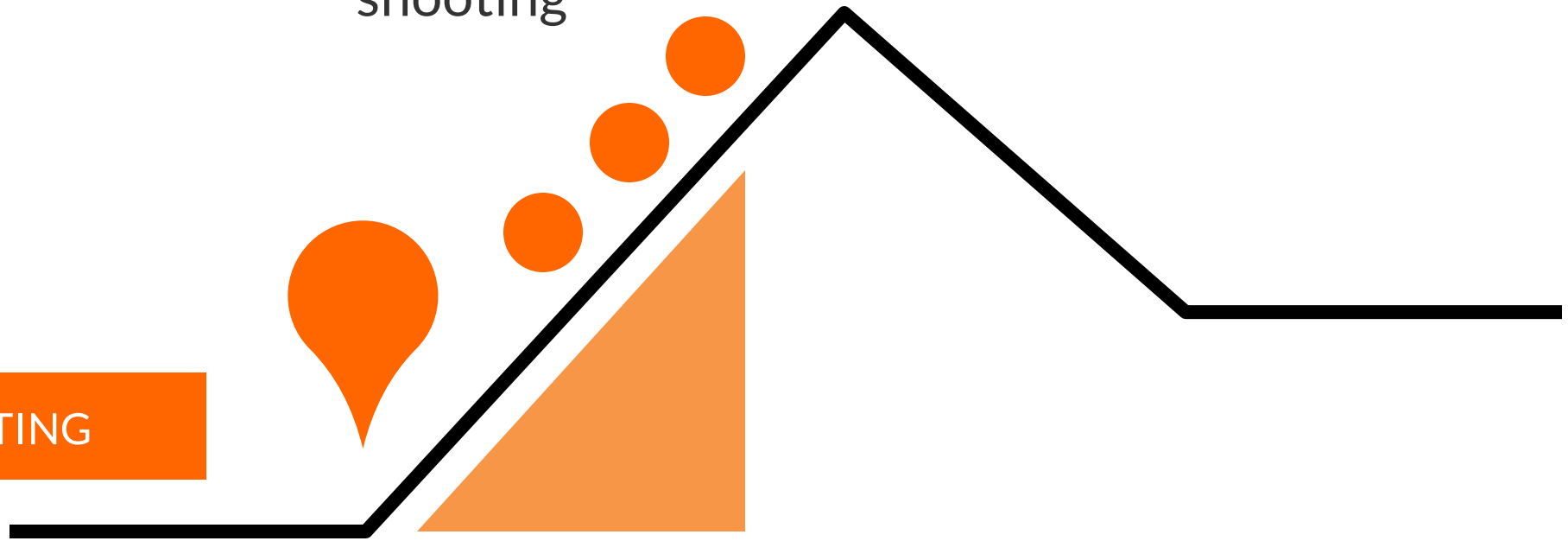


SETTING

Activity: reverse storyboarding

BEAT 3

- Context: regular April anniversary of Columbine shooting

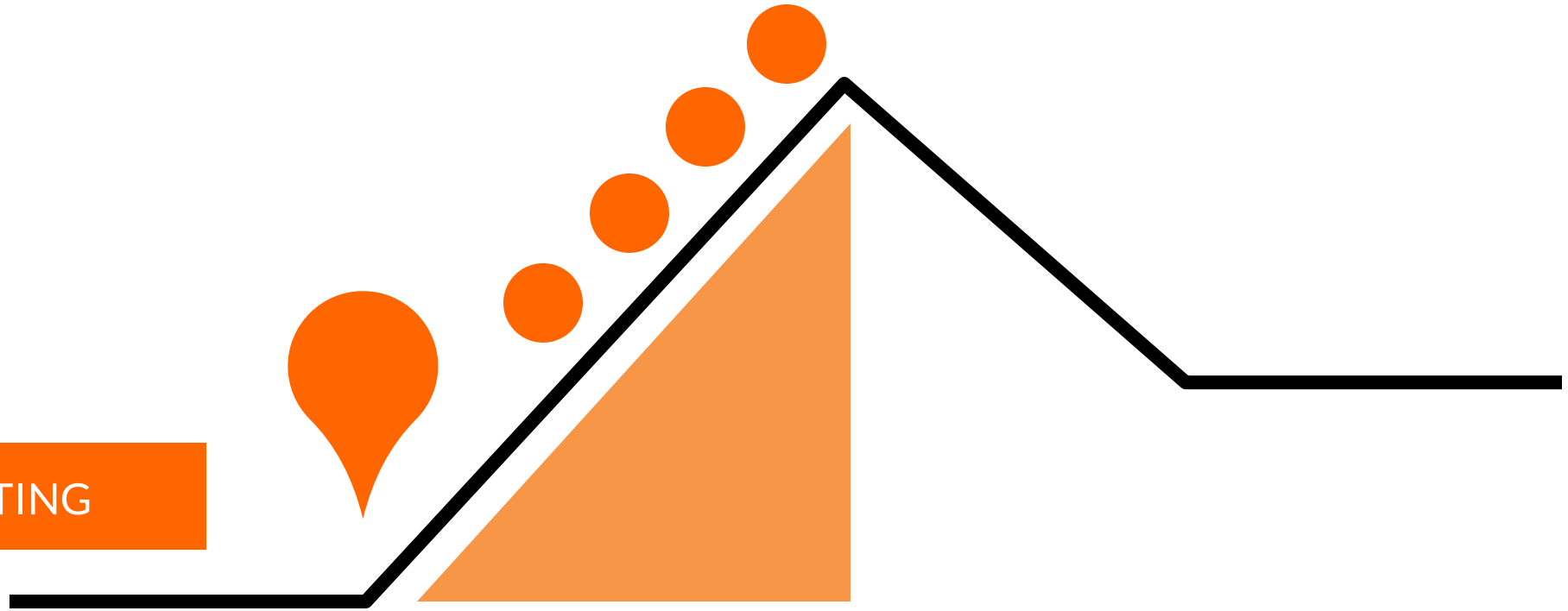


SETTING

Activity: reverse storyboarding

BEAT 4

- Contrast to gap around 9/11/01

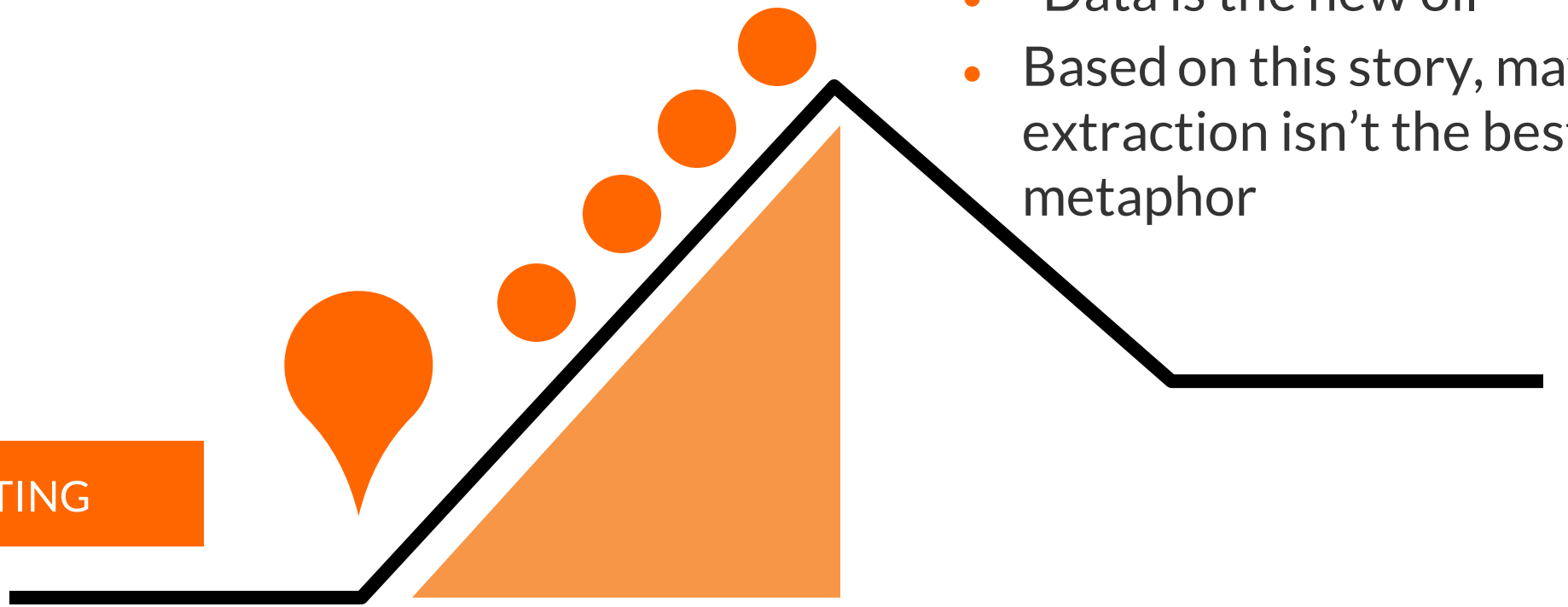


SETTING

Activity: reverse storyboarding

INSIGHT / OUTCOME

- "Data is the new oil"
- Based on this story, maybe extraction isn't the best metaphor

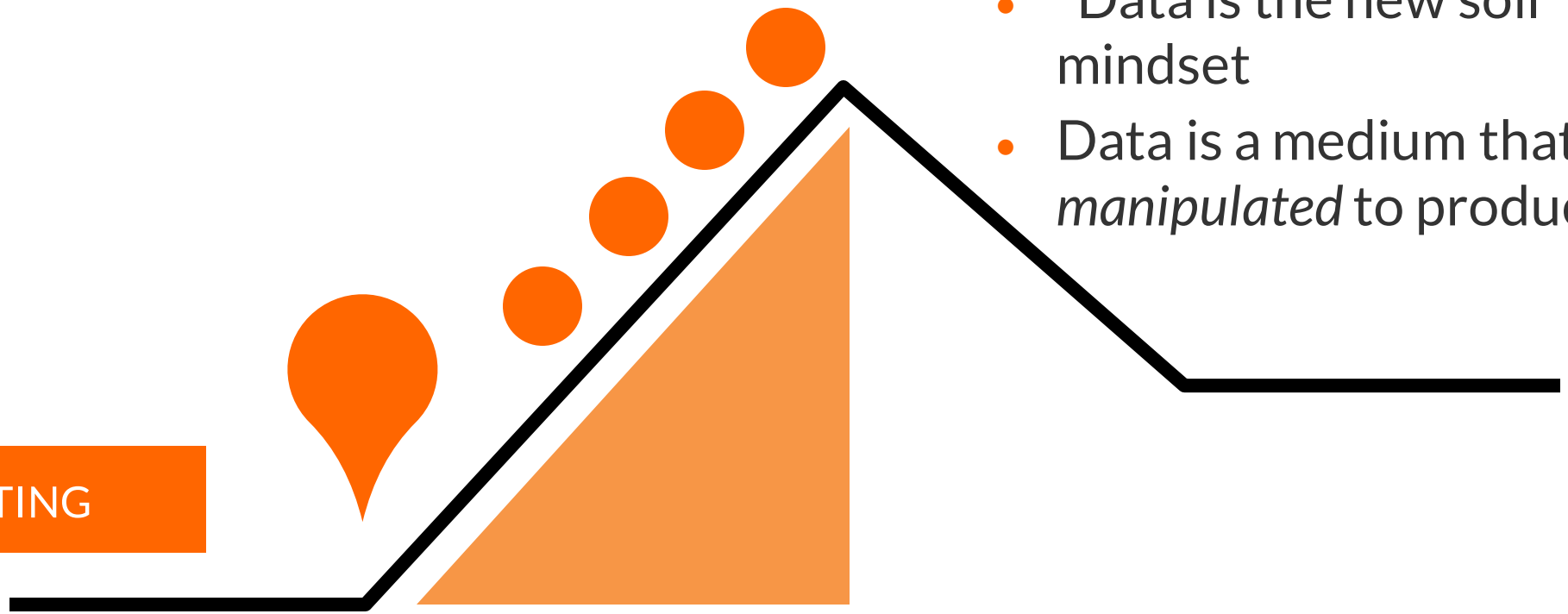


SETTING

Activity: reverse storyboarding

INSIGHT / OUTCOME

COURSE OF ACTION



- "Data is the new soil" – shift in mindset
- Data is a medium that must be *manipulated* to produce insights

SETTING

Q&A

DATA SOCIETY®

Thank
you!