Data Visualization

Vaibhav P. Vasani

Assistant Professor
Department of Computer Engineering
K. J. Somaiya College of Engineering
Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

Life Cycle of Data Analysis

- Phase 1: Discovery –
  - The Team learn and investigate the problem.
  - Develop context and understanding.
  - Come to know about data sources needed and available for the project.
  - Formulates initial hypothesis that can be later tested with data.

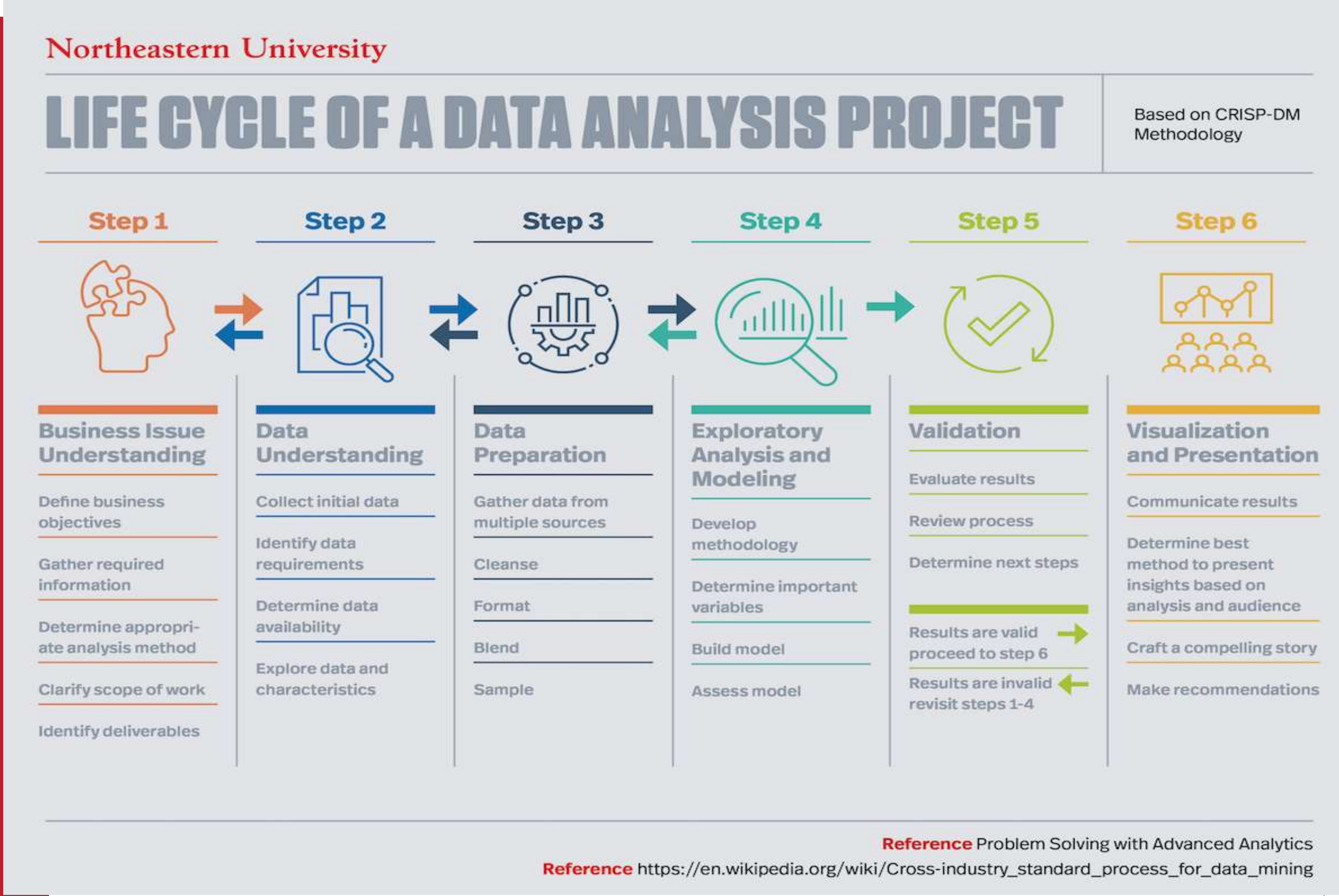
- **Phase 2: Data Preparation** –Steps to explore, preprocess, and condition data prior to modeling and analysis.
  - Execute, load, and transform the dataset
  - Data preparation tasks are likely to be performed multiple times and not in predefined order.
  - Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

- Phase 3: Model Planning –
  - Explore data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
  - Develop data sets for training, testing, and production purposes.
  - Team builds and executes models based on the work done in the model planning phase.
  - Several tools commonly used for this phase are – MATLAB, STASTICA.

- **Phase 4: Model Building –**Develops datasets for testing, training, and production purposes.
- Considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

- **Phase 5: Communication Results –**After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

- **Phase 6: Operationalize –**The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.



Northeastern University

LIFE CYCLE OF A DATA ANALYSIS PROJECT

Based on CRISP-DM Methodology

Step 1

Business Issue Understanding

- Define business objectives
- Gather required information
- Determine appropriate analysis method
- Clarify scope of work
- Identify deliverables

Step 2

Data Understanding

- Collect initial data
- Identify data requirements
- Determine data availability
- Explore data and characteristics

Step 3

Data Preparation

- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

Step 4

Exploratory Analysis and Modeling

- Develop methodology
- Determine important variables
- Build model
- Assess model

Step 5

Validation

- Evaluate results
- Review process
- Determine next steps

Results are valid proceed to step 6 →

Results are invalid revisit steps 1-4 ←

Step 6

Visualization and Presentation

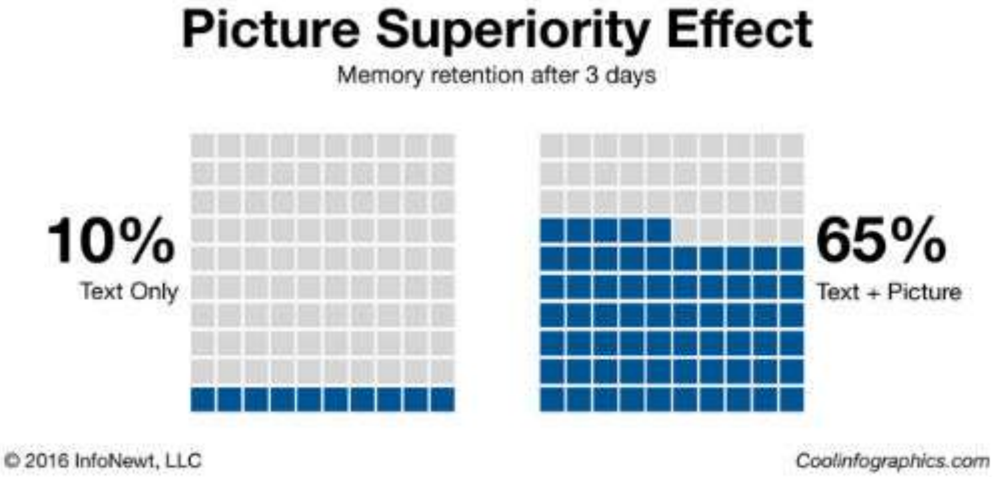
- Communicate results
- Determine best method to present insights based on analysis and audience
- Craft a compelling story
- Make recommendations

Reference Problem Solving with Advanced Analytics

Reference https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

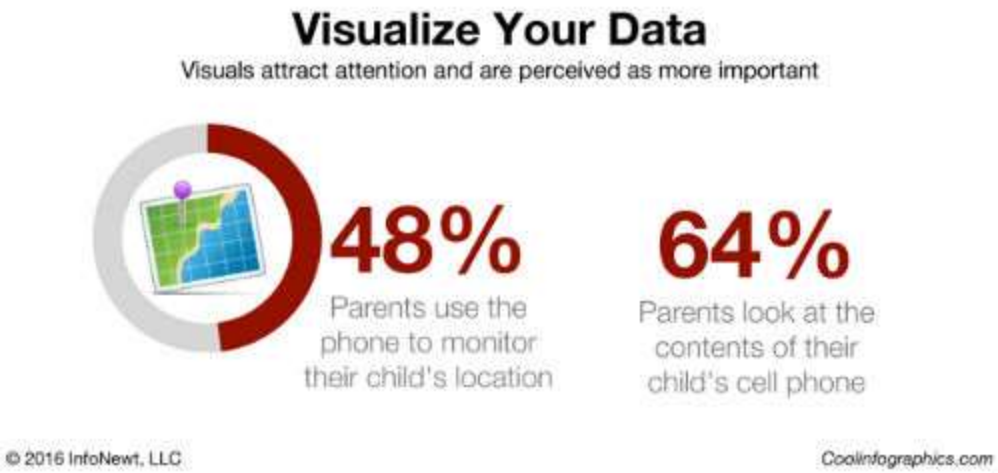
Visual Resolutions

- *“Big fonts are NOT data visualizations!”*
  - Our brains process visual information faster and more easily than text, and visual information is 650% more likely to be remembered by your audience than text alone (Brain Rules, John Medina, 2009). If you want to communicate a clear message, and you want your audience to remember that message, make it visual.



© 2016 InfoNewt, LLC

Coolinfographics.com



Visualize Your Data

Visuals attract attention and are perceived as more important

48%

Parents use the phone to monitor their child's location

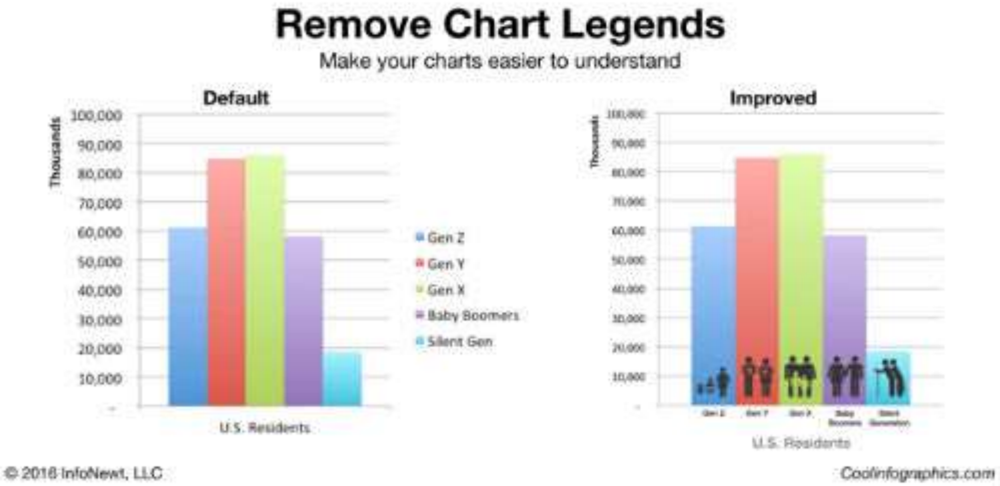
64%

Parents look at the contents of their child's cell phone

© 2016 InfoNewt, LLC

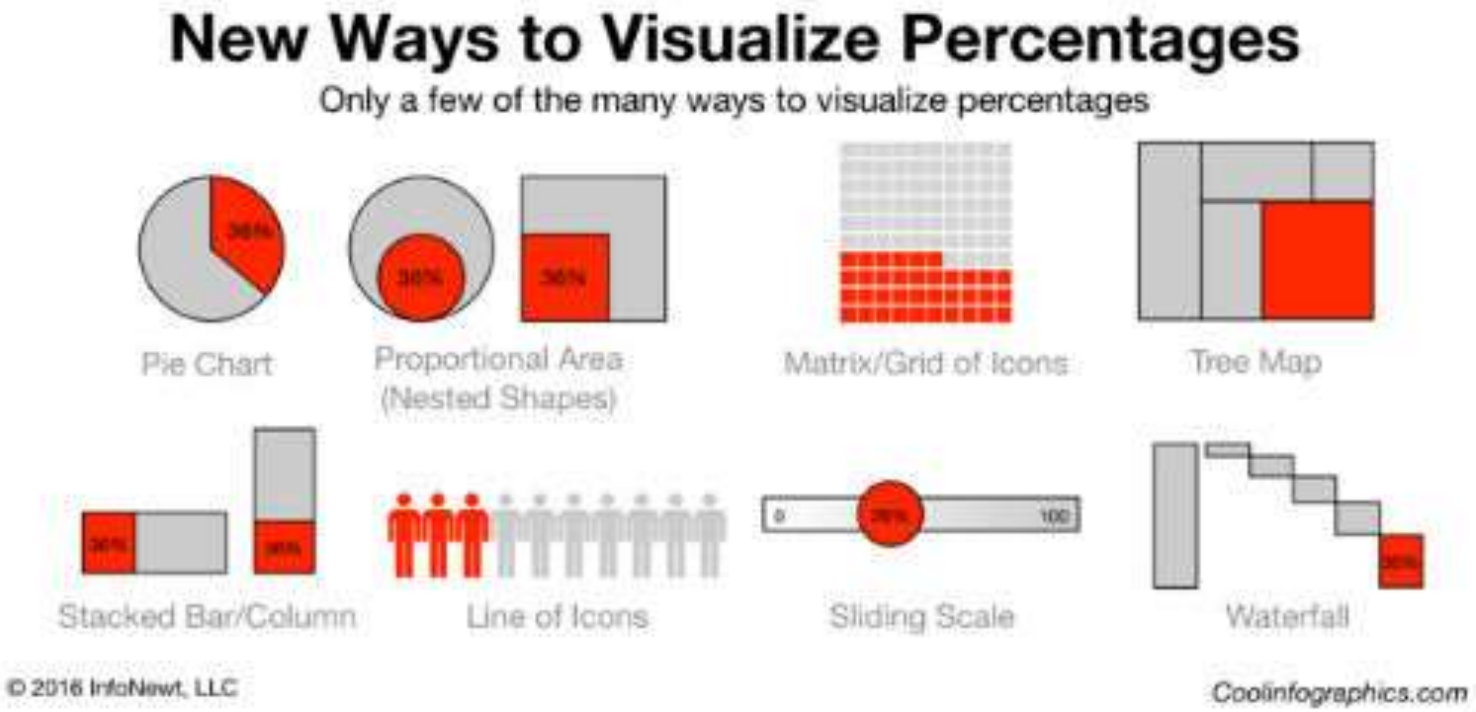
Coolinfographics.com

- **Remove Chart Legends**
  - Legends that are separate from the visualization of the data make your readers work much harder, looking back and forth between the data and the legend, to understand your visualization. Make understanding your data visualization much faster and easier by moving the data descriptions into the chart itself, and connected to the actual data.



Source: https://www.kdnuggets.com/2016/01/three-simple-resolutions-design-better-dataviz.html

- **Try New Ways to Visualize Your Data**



New Ways to Visualize Percentages

Only a few of the many ways to visualize percentages

© 2016 InfoNewt, LLC

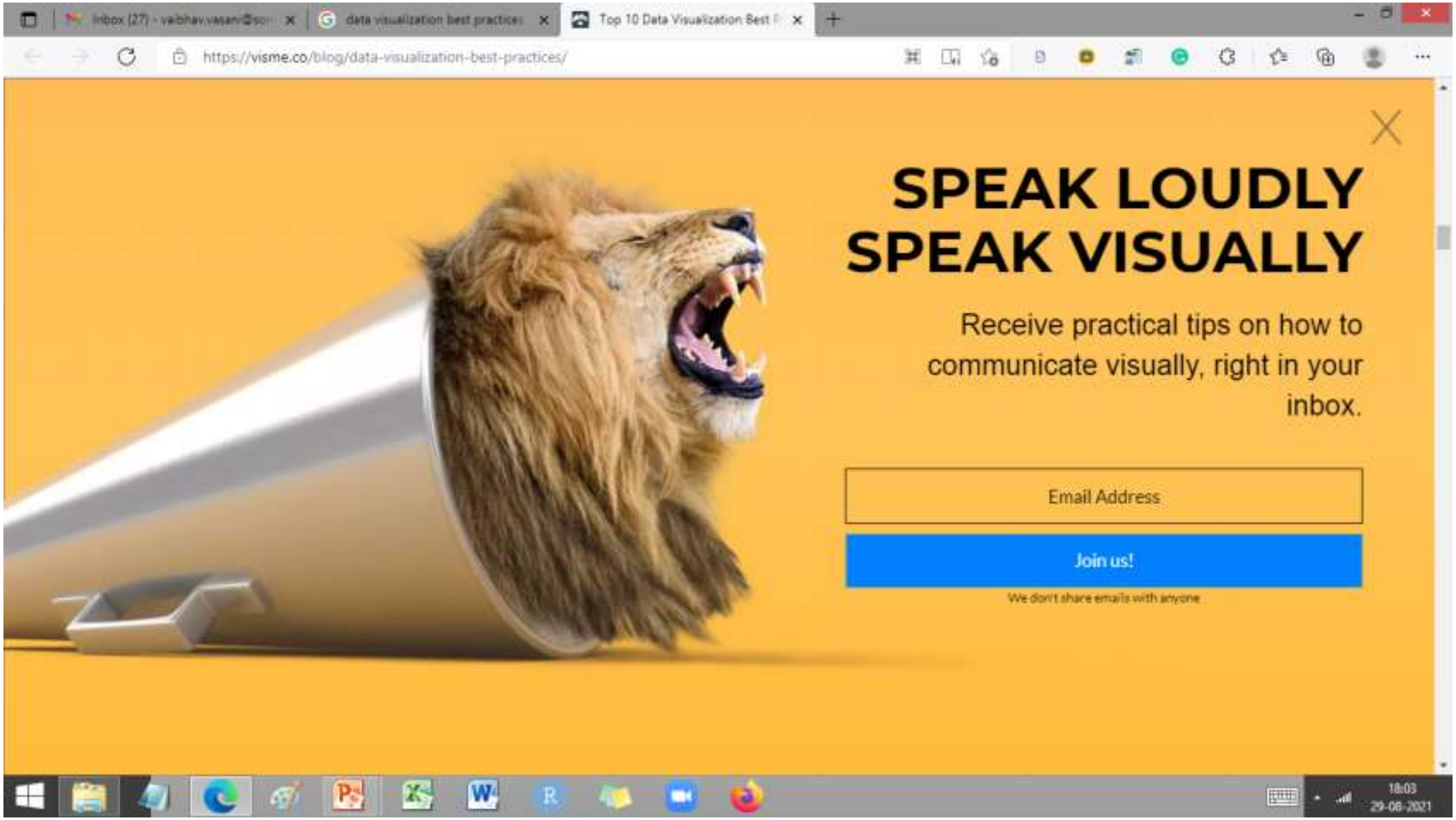
Coolinfographics.com

- https://datavizcatalogue.com/
- https://policyviz.com/2014/11/11/graphic-continuum-desktop-version/
- https://www.visual-literacy.org/periodic_table/periodic_table.html

- *Michael Friendly defines data visualization "as information which has been abstracted in some schematic form, including attributes or variables for the units of information." In other words, it is a coherent way to visually communicate quantitative content. Depending on its attributes, the data may be represented in many different ways, such as a line graph, bar chart, pie chart, scatter plot, or map.*

- *According to IBM, 2.5 quintillion bytes of data are created every day. The Research Scientist Andrew McAfee and Professor Erik Brynjolfsson of MIT* point *out that "more data cross the internet every second than were stored in the entire internet just 20 years ago."*

- *IDC predicts there will be 163 zettabytes (163 trillion gigabytes) of data by 2025.*



Inbox (27) - vaibhav.vasani@so
data visualization best practices
Top 10 Data Visualization Best P
https://visme.co/blog/data-visualization-best-practices/

SPEAK LOUDLY
SPEAK VISUALLY

Receive practical tips on how to communicate visually, right in your inbox.

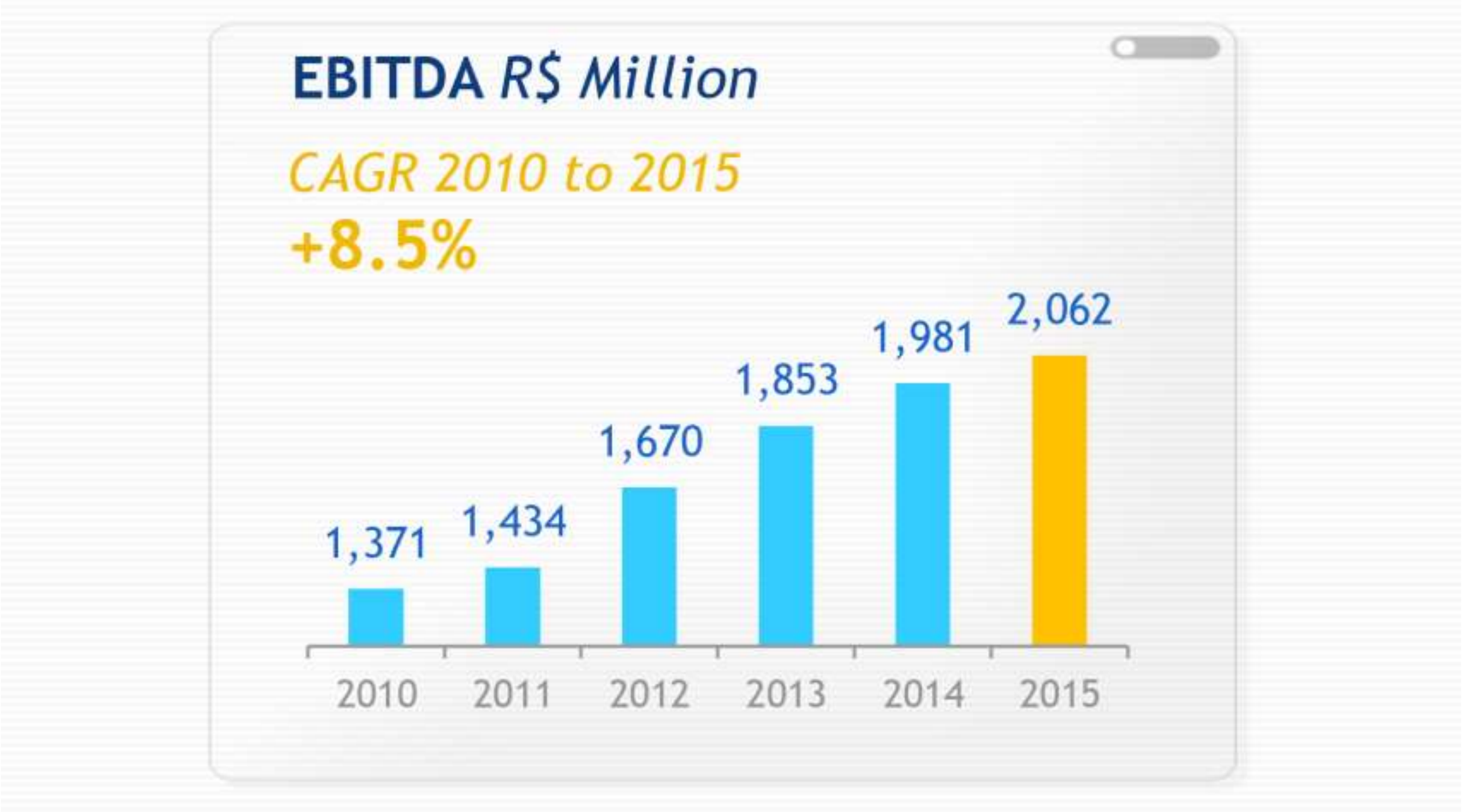
We don't share emails with anyone

Data visualization best practices

- Define a Clear Purpose
- Data visualizations should be used to empower a specific audience and address their needs
- Choose the right visual for your purpose
- Provide Context: Context engenders trust, which leads to action
- Keep visualizations and dashboards simple and digestible
- Design to keep users engaged
- Don’t Distort the Data
- Accept feedback.

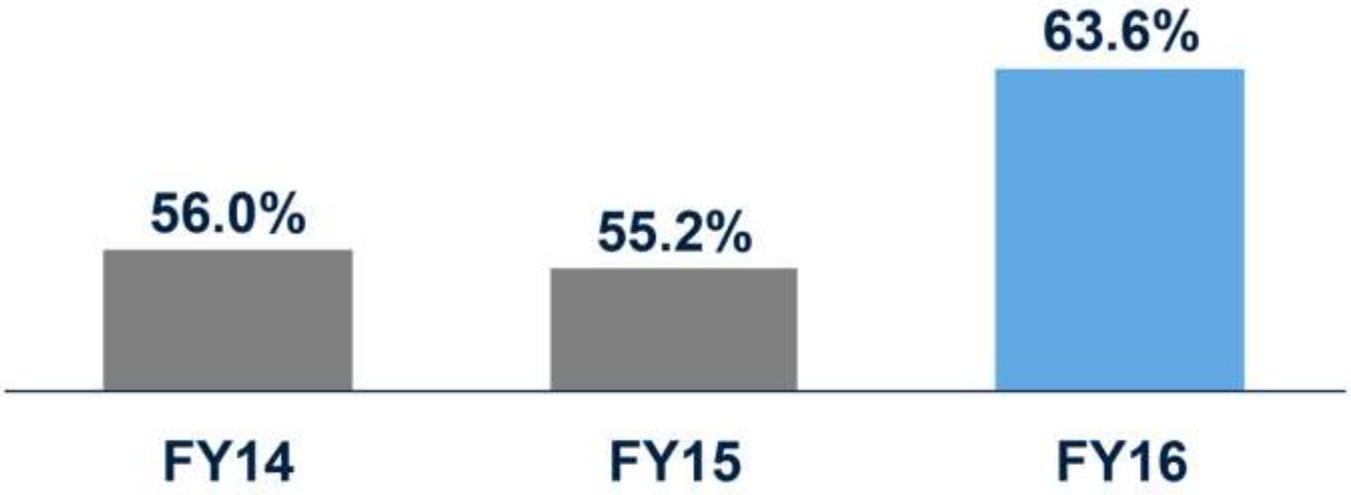
Don’t Distort the Data

Identify problem

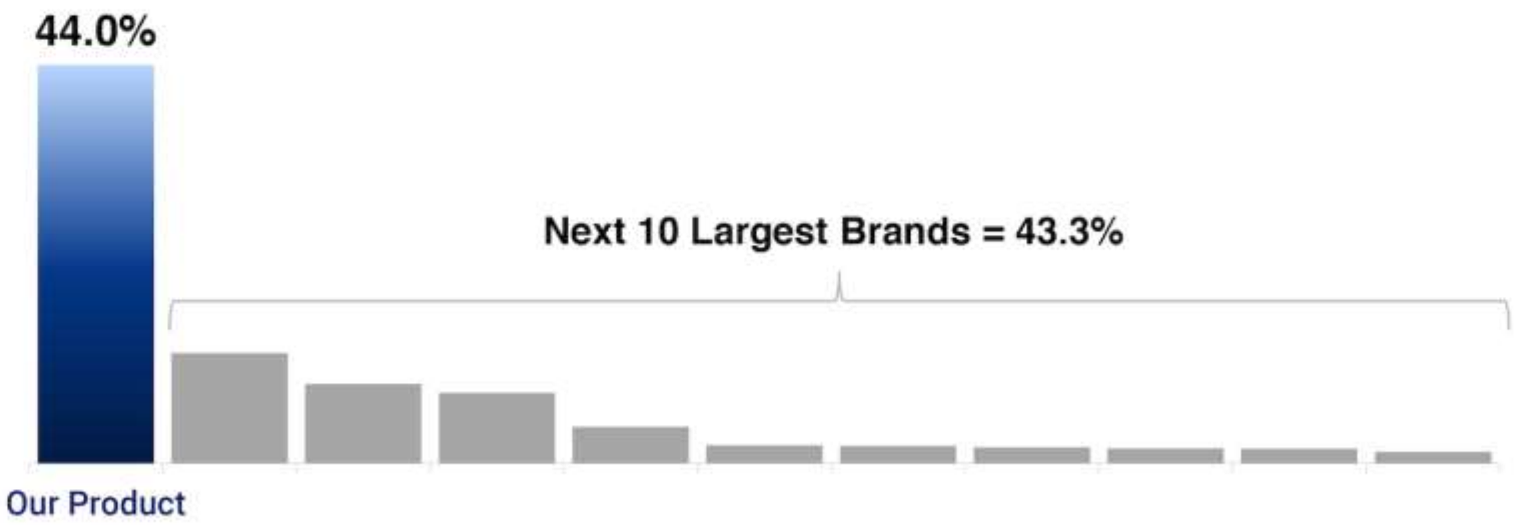


Significant progress in 2016…

Departure:00[1]

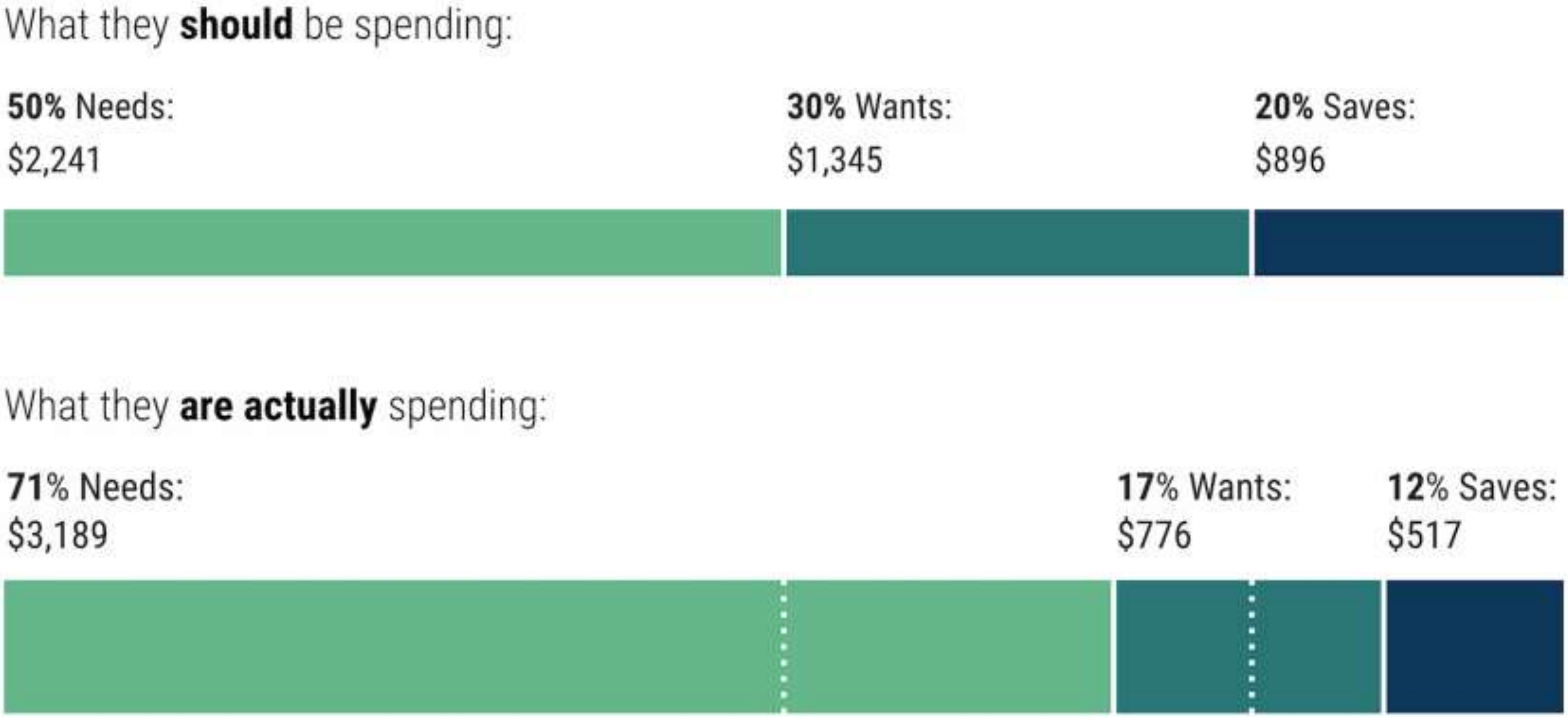


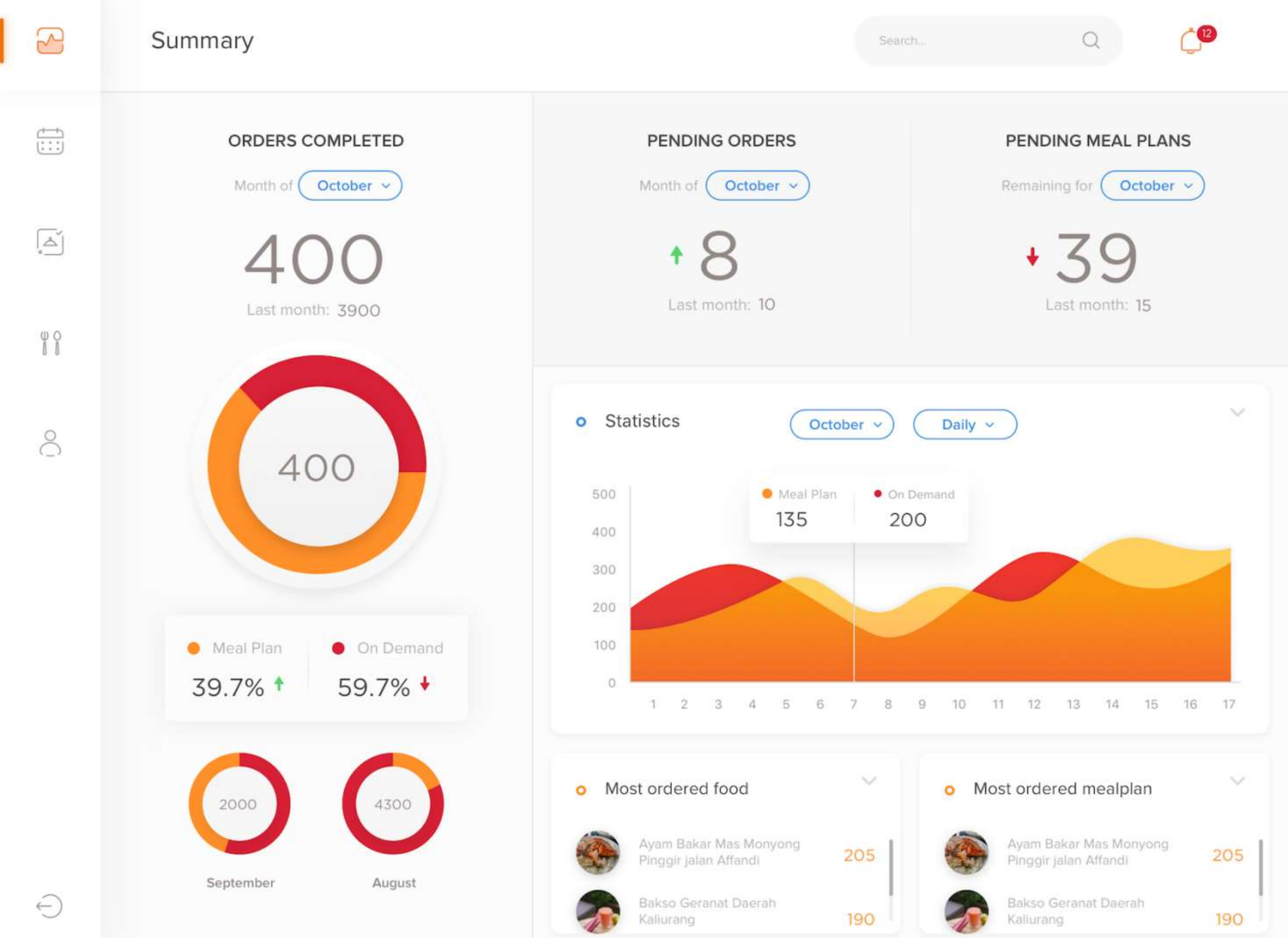
2015 Retail Share



Previous iPad
New iPad
GIZMODO

Right or Wrong?





Summary

Search...

12

ORDERS COMPLETED

Month of October

400

Last month: 3900

PENDING ORDERS

Month of October

8

Last month: 10

PENDING MEAL PLANS

Remaining for October

39

Last month: 15

● Meal Plan	● On Demand
39.7%	59.7%

Statistics

October

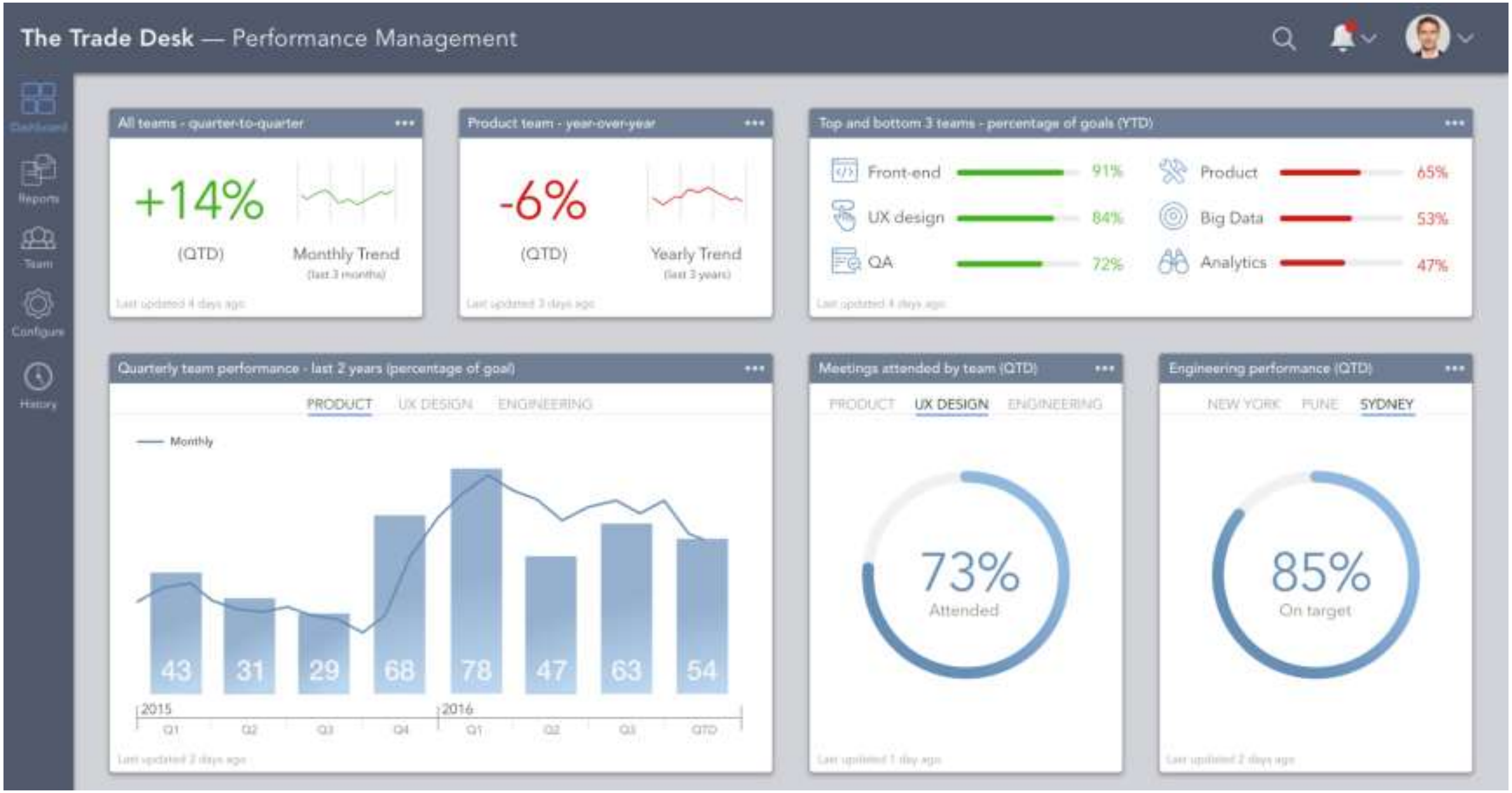
Daily

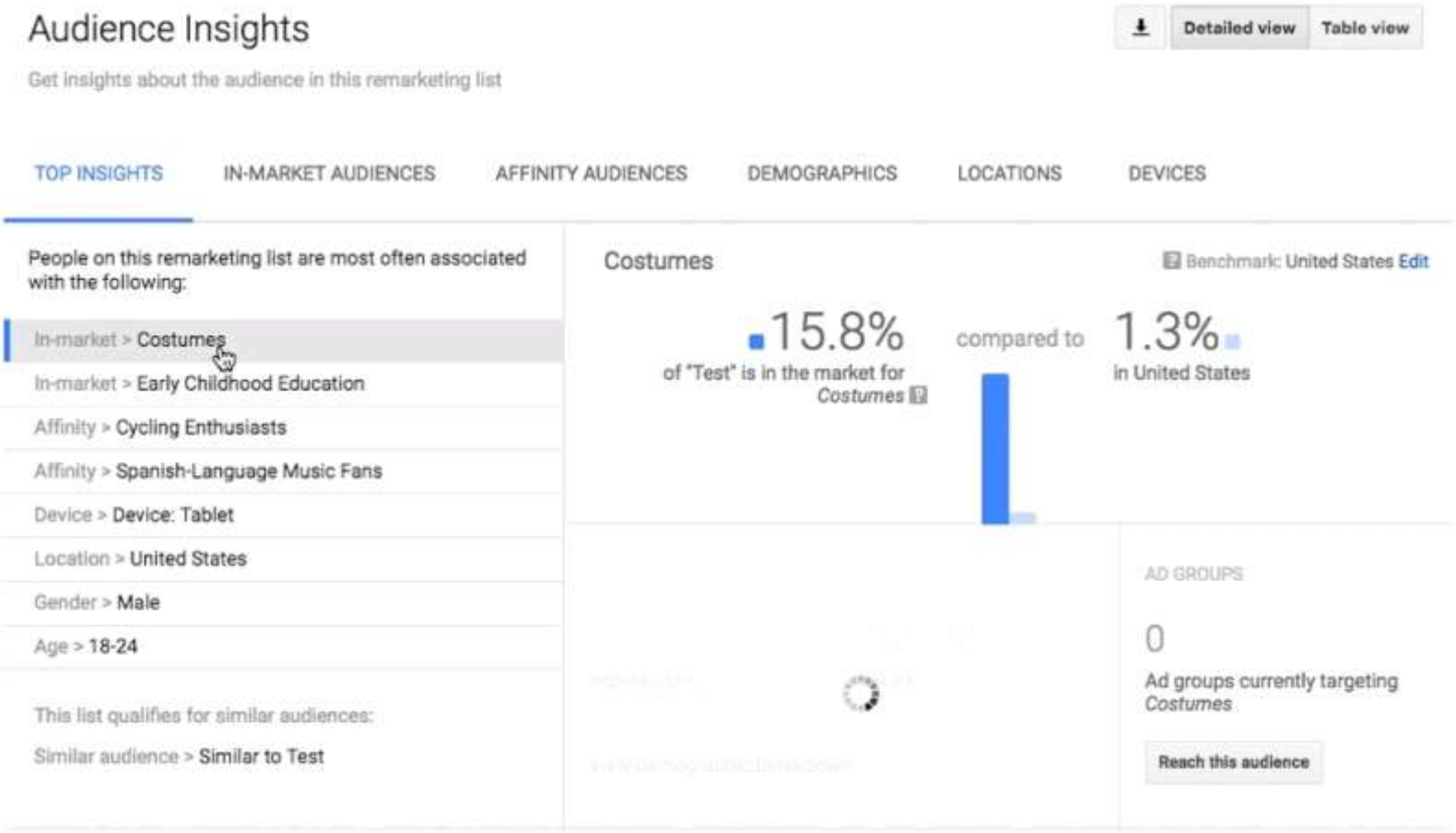
Most ordered food

Ayam Bakar Mas Monyong Pinggir jalan Affandi	205
Bakso Geranat Daerah Kaliurang	190

Most ordered mealplan

Ayam Bakar Mas Monyong Pinggir jalan Affandi	205
Bakso Geranat Daerah Kaliurang	190


The Trade Desk — Performance Management
Dashboard
Reports
Team
Configure
History
All teams - quarter-to-quarter
+14%
(QTD)
Monthly Trend
(last 3 months)
Last updated 4 days ago
Product team - year-over-year
-6%
(QTD)
Yearly Trend
(last 3 years)
Last updated 3 days ago
Top and bottom 3 teams - percentage of goals (YTD)
Front-end 91%
UX design 84%
QA 72%
Product 65%
Big Data 53%
Analytics 47%
Last updated 4 days ago
Quarterly team performance - last 2 years (percentage of goal)
PRODUCT
UX DESIGN
ENGINEERING
Monthly
43
31
29
68
78
47
63
54
2015
Q1
Q2
Q3
Q4
2016
Q1
Q2
Q3
QTD
Last updated 2 days ago
Meetings attended by team (QTD)
PRODUCT
UX DESIGN
ENGINEERING
73%
Attended
Last updated 1 day ago
Engineering performance (QTD)
NEW YORK
PUNE
SYDNEY
85%
On target
Last updated 2 days ago



Audience Insights

Get insights about the audience in this remarketing list

People on this remarketing list are most often associated with the following:

- In-market > Costumes
- In-market > Early Childhood Education
- Affinity > Cycling Enthusiasts
- Affinity > Spanish-Language Music Fans
- Device > Device: Tablet
- Location > United States
- Gender > Male
- Age > 18-24

This list qualifies for similar audiences:

Similar audience > Similar to Test

Costumes

Benchmark: United States Edit

15.8% of "Test" is in the market for *Costumes* compared to 1.3% in United States

AD GROUPS

0

Ad groups currently targeting *Costumes*

Effective Data Visualisation with Design Principles

benefit

- Use of pictures, graphics, or images, that make it easier to understand new information and find useful trends.
- Reduces your overall time and effort spent on data analysis.
- Useful to a wider set of users including sales & marketing and finance heads.
- Enables faster decision making in designing or revising key business strategies and taking business actions.
- Improves your business ROI from data.

- Identify their business strengths or areas of improvement.
- Determine the factors that influence the online behavior of customers.
- Arrive at the right strategy for product placement and pricing.
- Predict future trends including sales volumes.

Balance the Design

- A balanced design is one with visual elements like shape, colour, negative space, and texture equally distributed across the plot.
- There are three different types of balances in design:
  - Symmetrical–Each side of the visual is the same as the other
  - Asymmetrical–Both sides are different but still have a similar visual weight
  - Radial–Elements are placed around a central object which acts as an anchor
- You will have to figure out which type of balance works the best for your data visualization and apply that.

ARCH BAY

John Adams Middle School
12 : 1
Jackson Kenneth

Emphasise the Key Areas

- The user’s attention should be drawn to the right data points by carefully choosing the size, colours, contrast, and negative space.
- The goal of the data visualization is to make sure that the important data doesn’t go unnoticed and emphasizing it helps.
- Since the attention of a user first falls in the top-left corner of a plot, you should place the important data points there.

Illustrating Movement

- Movement directs the user's attention in a certain direction, just like emphasis.
- Your visual elements should mimic movement in an "F" pattern, which is how people read.
- Starting from top left to right, and gradually down the page.
- You could also illustrate movement across the page by using complementary colours that can catch the viewer's gaze and take it across the page.
- This principle is more applicable to static visualizations.
- If your data visualization tool is capable of animation and interactive designs, the movement aspect should already be covered.

Photo by hannah cauhepe on Unsplash

Smart Use of Patterns

- Repeated design elements form a pattern.
- When it comes to visualizing your data, patterns make for a great way to display similar types of information spread across the page as one.
- If the data on the page is too much for emphasizing, establishing a pattern by using similar colours, chart types and elements are the way to go.
- Patterns also make it easier to communicate an anomaly, since any disruption in the pattern will naturally draw the viewer’s attention and curiosity.
- Using patterns is one of the simplest and most effective design principles when it comes to data visualization.

Proportion

- If you are going to draw a picture of a bird on a tree, the tree will be significantly bigger compared to the bird.
- In data visualization, the proportion is made up of the size of each element on the page.
- Proportions in data visualization can indicate the weight of different data sets and the relationship between their values.
- If you need to emphasize the importance of a certain data point, all you have to do is to make it bigger than the rest.
- In addition to this, you should ensure that the chart reflects the interrelationship of various numbers as accurately as possible.
- For example, if a slice in a pie chart is marked 36%, it should actually use 36% of the area inside the chart.

Proper Rhythm

- Rhythm is a rather vague design principle that is closely associated with movement.
- A design is said to have a balanced rhythm when the design elements together create a pleasing movement to the eye.
- If the design elements like shapes, colors, or proportions together create a “choppiness”, you might want to rearrange them to facilitate smooth eye movement across the data.

Variety

- Variety is an important factor that keeps viewers engaged and interested in your data.
- It’s all about finding ways to visualize your data using different and interesting design elements to avoid repetition.
- The result will be a data visualization which is not only eye-catching but also helps the viewer retain the information presented for longer.

Theme

- A unified theme ensures every part of your design is consistent and follows a standard.
- This should happen naturally if you have taken care of the aforementioned design principles.
- You can incorporate a theme for your company or based on the niche of the visualization.
- This helps connect with the user on a deeper level and augments the visual design.

Unity

- Much like balance, unity ensures that every part of your design is congruent.
- This gives your dashboard the impression of having an overarching “theme” and it will happen naturally if you’ve already implemented the other core principles of data visualization.
- If your dashboard doesn’t feel unified, revisit the other eight design principles and make sure you’re employing them.

Design your visualization for the mobile phone.

- With the increase in the number of mobile phone users, the demand for data visualization on smartphones is growing. Visualized data in the form of stock price changes or industrial production-related data on a smartphone is now driving better actions and decisions.
- As a data analyst, you need to design and optimize your data visualization for the mobile phone. This could include design practices such as:
- Designing charts with the most critical information on the top left corner.
- Use of appropriate color coding.
- Zoom-in features online graphs to show data changes in detail.
- Avoiding the use of graph titles and axis labels

The Bar Chart

- A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

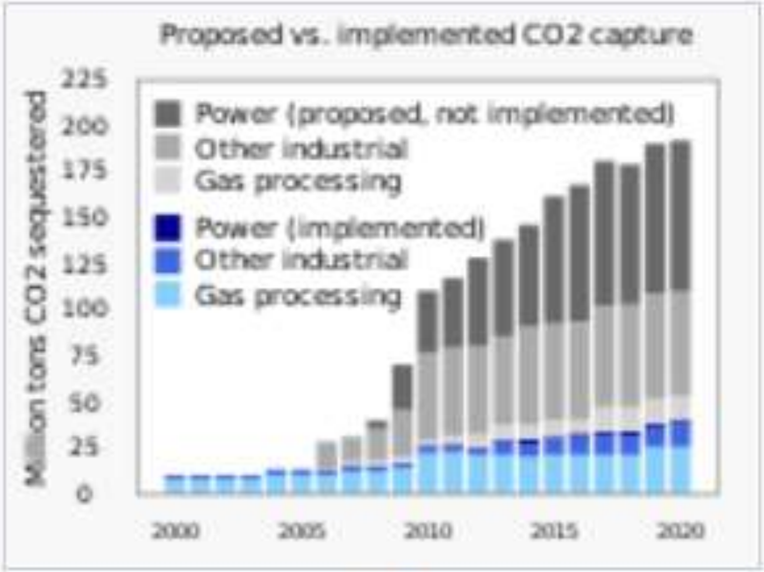
- Bar graphs/charts provide a visual presentation of categorical data.
- Categorical data is a grouping of data into discrete groups, such as months of the year, age group, shoe sizes, and animals.
- These categories are usually qualitative.
- In a column (vertical) bar chart, categories appear along the horizontal axis and the height of the bar corresponds to the value of each category.

- Bar charts have a discrete domain of categories, and are usually scaled so that all the data can fit on the chart.
- When there is no natural ordering of the categories being compared, bars on the chart may be arranged in any order.
- Bar charts arranged from highest to lowest incidence are called Pareto charts.

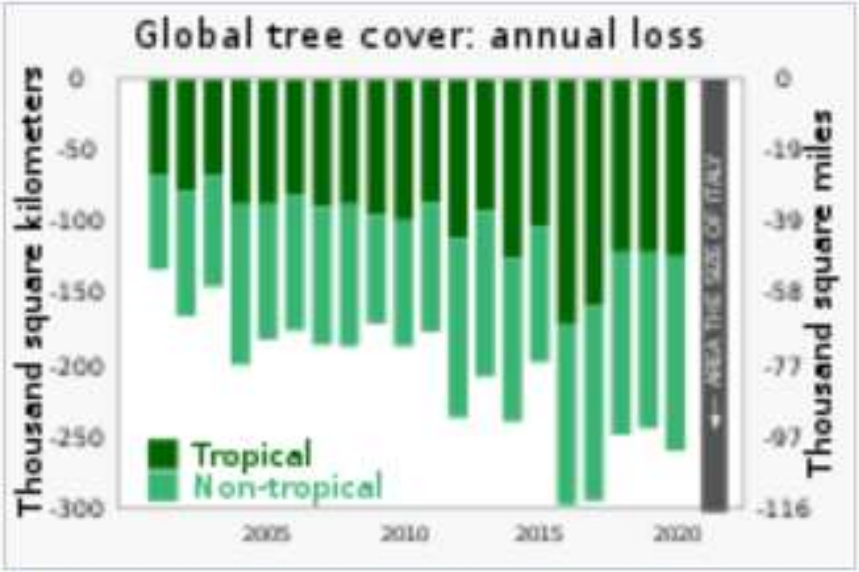
Grouped (clustered) and stacked

- Bar graphs can also be used for more complex comparisons of data with grouped (or "clustered") bar charts, and stacked bar charts.
- In **grouped (clustered) bar charts**, for each categorical group there are two or more bars color-coded to represent a particular grouping.
  - For example, a business owner with two stores might make a grouped bar chart with different colored bars to represent each store: the horizontal axis would show the months of the year and the vertical axis would show revenue.

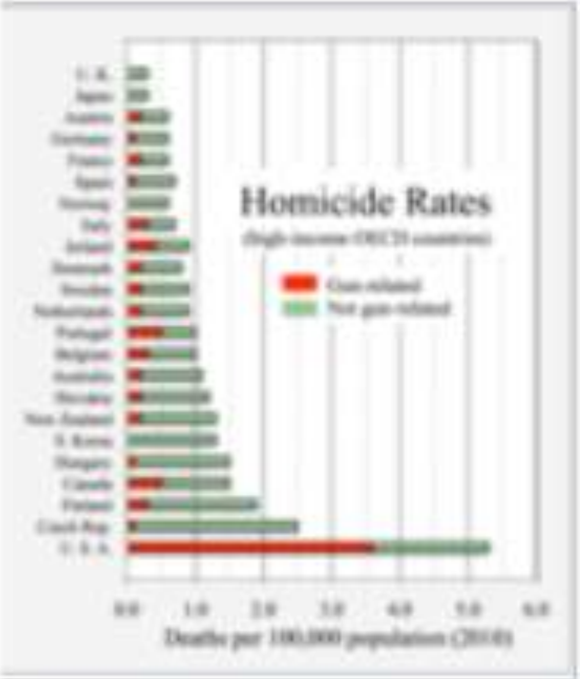
- a **stacked bar chart** stacks bars on top of each other so that the height of the resulting stack shows the combined result. Stacked bar charts are not suited to data sets having both positive and negative values.
- Grouped bar charts usually present the information in the same order in each grouping. Stacked bar charts present the information in the same sequence on each bar.
-



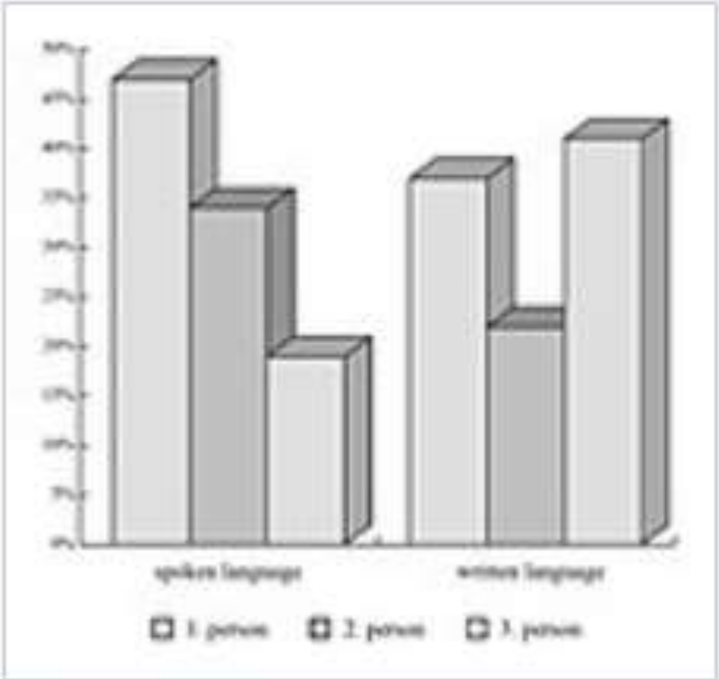
A vertical stacked bar chart with positive values



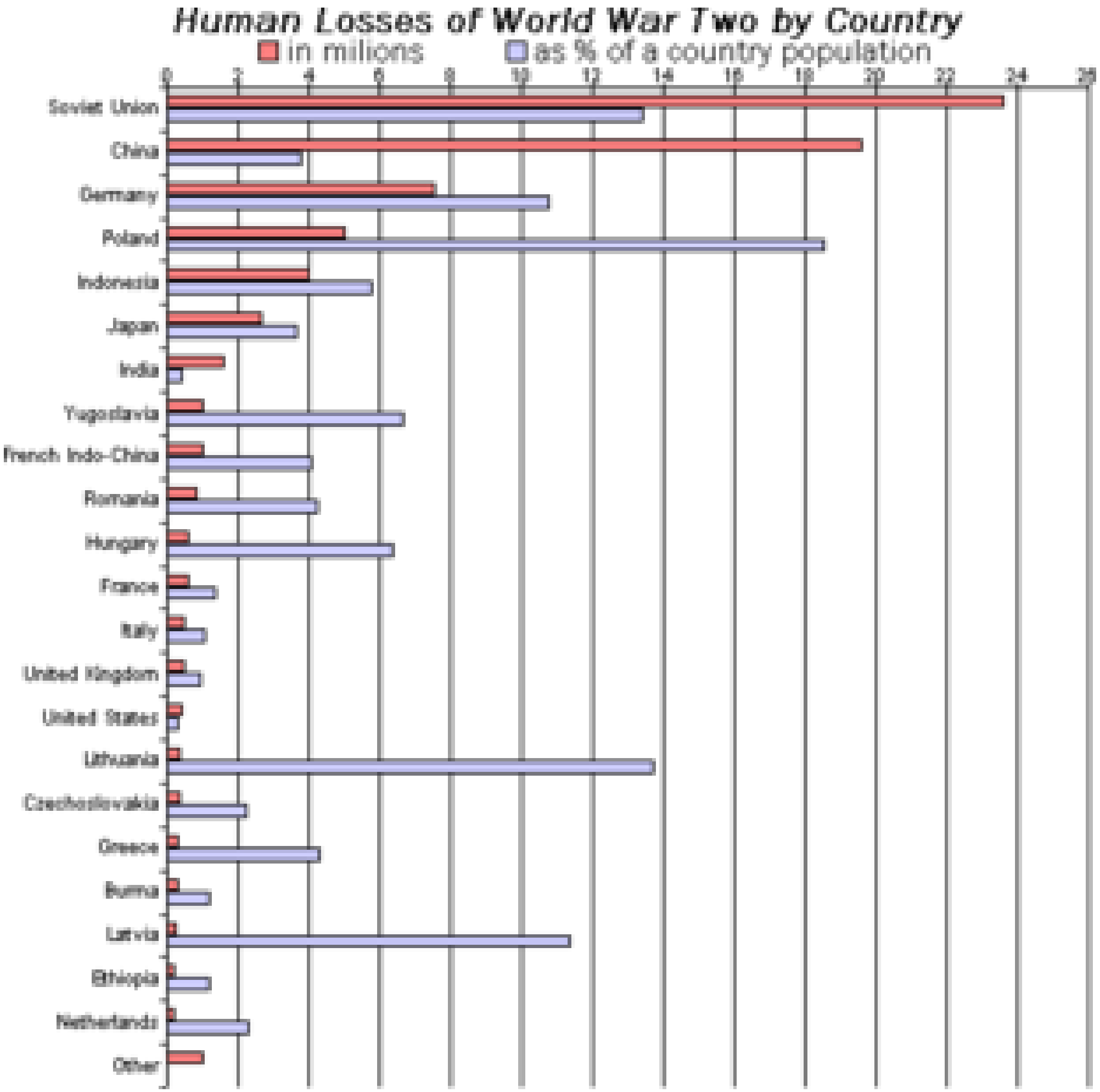
A vertical stacked bar chart with negative values

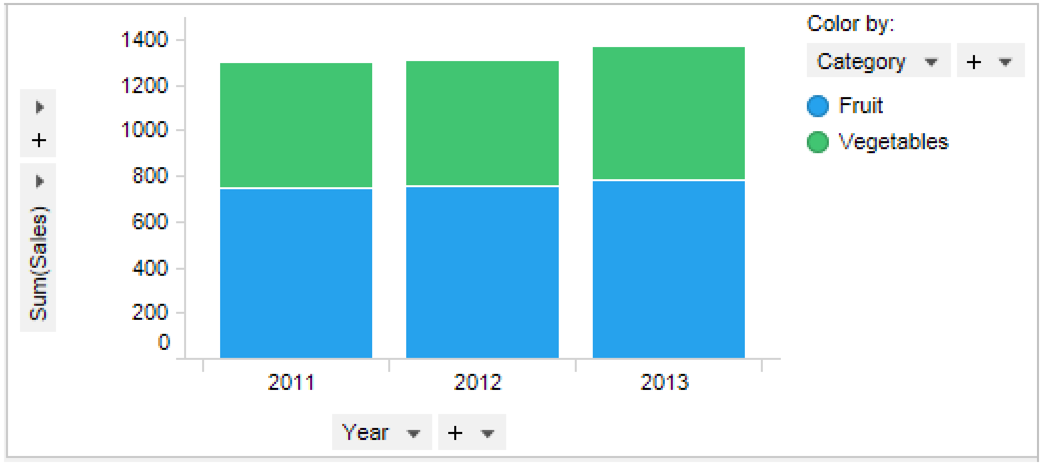


A horizontal stacked bar chart



A vertical, grouped (clustered) 3D bar chart


Human Losses of World War Two by Country
🟥 in millions
🟦 as % of a country population
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Soviet Union
China
Germany
Poland
Indonesia
Japan
India
Yugoslavia
French Indo-China
Romania
Hungary
France
Italy
United Kingdom
United States
Lithuania
Czechoslovakia
Greece
Burma
Latvia
Ethiopia
Netherlands
Other

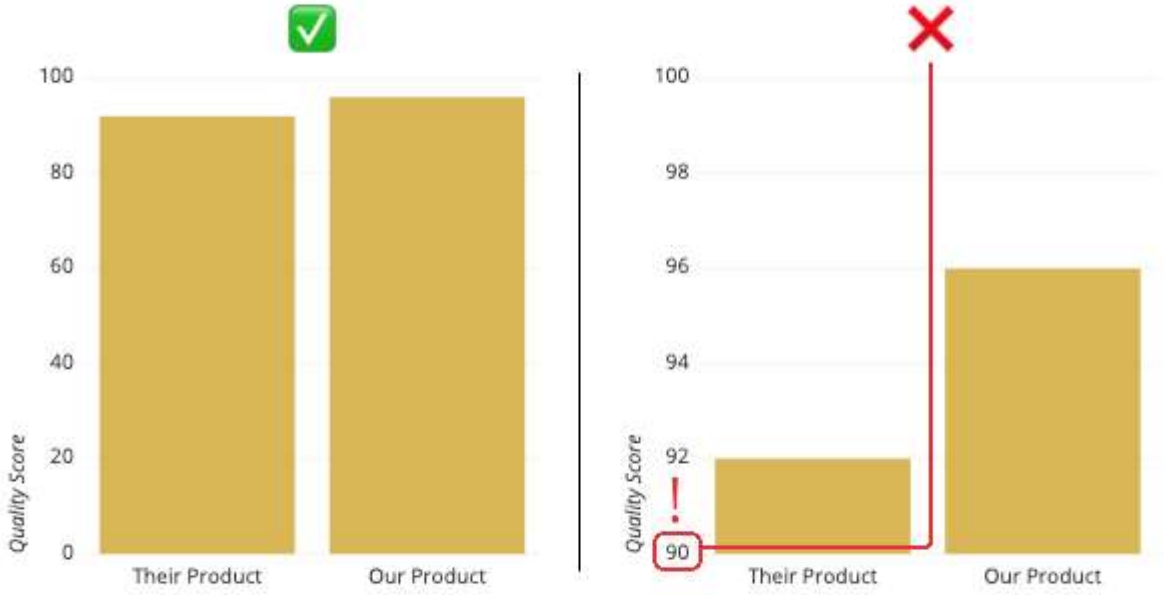

Color by:
Category
+
Fruit
Vegetables
1400
1200
1000
800
600
400
200
0
Sum(Sales)
+
2011
2012
2013
Year
+

When you should use a bar chart

- A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data.
- From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others.
- Since this is a fairly common task, bar charts are a fairly ubiquitous chart type.

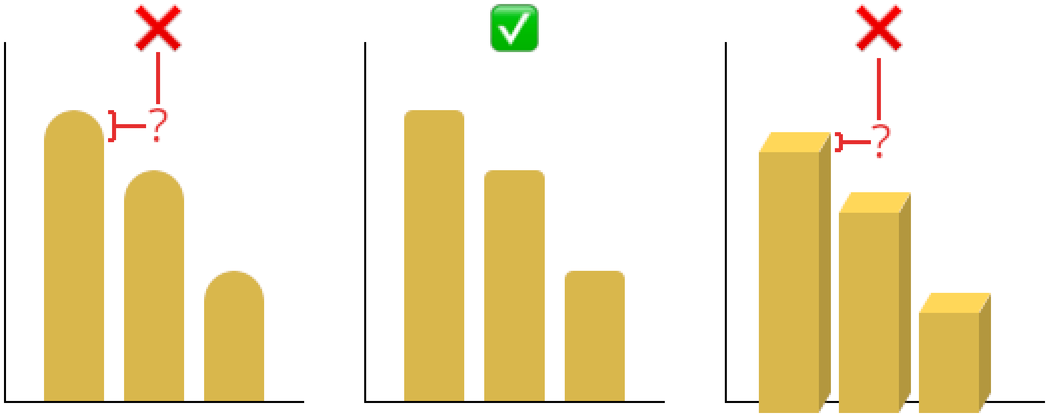
Best practices for using bar charts

- **Use a common zero-valued baseline**
  - First and foremost, make sure that all of your bars are being plotted against a zero-value baseline. Not only does that baseline make it easier for readers to compare bar lengths, it also maintains the truthfulness of your data visualization. A bar chart with a non-zero baseline or some other gap in the axis scale can easily misrepresent the comparison between groups since the ratio in bar lengths will not match the ratio in actual bar values.


✓
100
80
60
40
20
0
Quality Score
Their Product
Our Product
✗
100
98
96
94
92
!
90
Quality Score
Their Product
Our Product

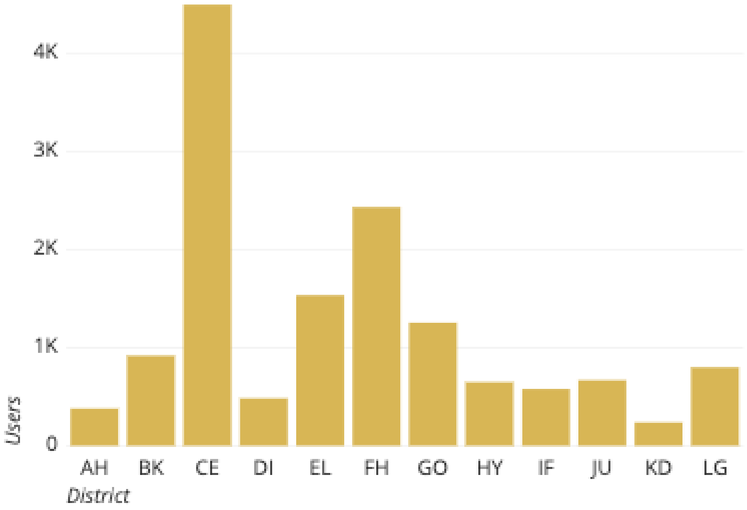
Maintain rectangular forms for your bars

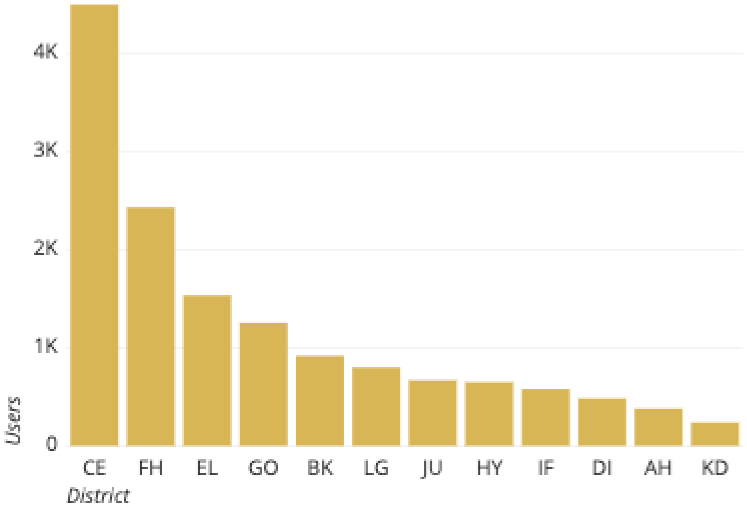
- Another major no-no is to mess with the shape of the bars to be plotted. Some tools will allow for the rounding of the bar caps, rather than just have straight edges. This rounding means that it’s difficult for the reader to tell where to read the actual value: from the top of the semicircle, or somewhere in the middle? A little bit of rounding of the corners can be okay, but make sure each bar is flat enough to discern its true value and provide an easy comparison between bars.
- Similarly, you should avoid including 3-d effects on your bars. As with heavy rounding, this can make it harder to know how to measure bar lengths, and as a bonus, might cause baselines to not be aligned (see the above point).



Consider the ordering of category levels

- One consideration you should have when putting together a bar chart is what order in which you will plot the bars. A standard convention to take is to sort the bars from longest to shortest: while it is always possible to compare the bar lengths no matter the order, this can reduce the burden on the reader to make those comparisons themselves. The major exception to this is if the category labels are inherently ordered in some way. In cases like that, the inherent ordering usually takes precedence.
-

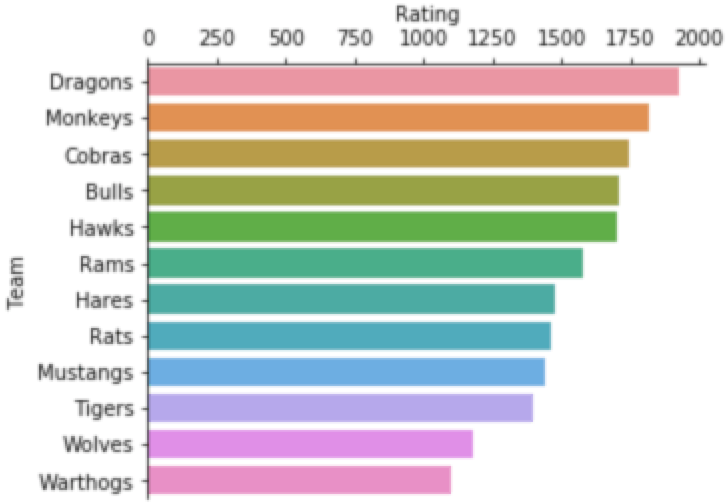

4K
3K
2K
1K
0
Users
AH
BK
CE
DI
EL
FH
GO
HY
IF
JU
KD
LG
District


4K
3K
2K
1K
0
Users
CE
FH
EL
GO
BK
LG
JU
HY
IF
DI
AH
KD
District

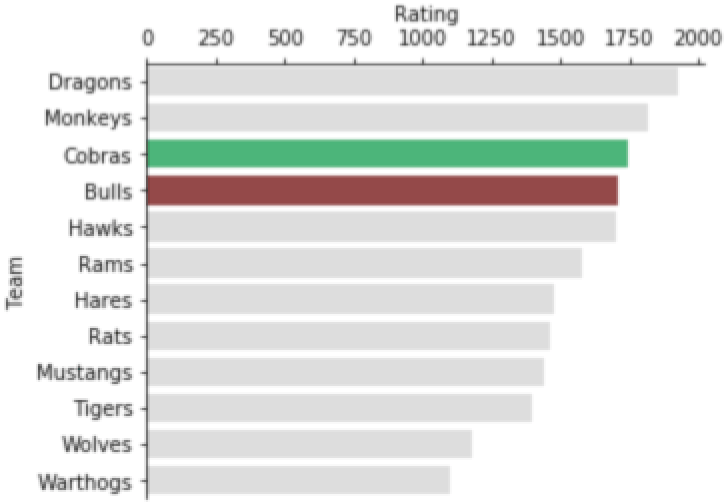
Use color wisely

- Another consideration is on how you should use color in your bar charts. Certain tools will color each bar differently by default, but this can distract the reader by implying additional meaning where none exists. Instead, color should be used with purpose. For example, you might use color to highlight specific columns for storytelling. Colors can also be used if they are meaningful for the categories posted (e.g. to match company or team colors).

✗


Rating
0
250
500
750
1000
1250
1500
1750
2000
Team
Dragons
Monkeys
Cobras
Bulls
Hawks
Rams
Hares
Rats
Mustangs
Tigers
Wolves
Warthogs

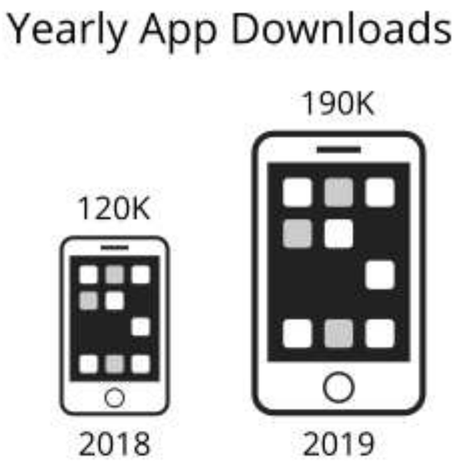
✓


Rating
0
250
500
750
1000
1250
1500
1750
2000
Team
Dragons
Monkeys
Cobras
Bulls
Hawks
Rams
Hares
Rats
Mustangs
Tigers
Wolves
Warthogs

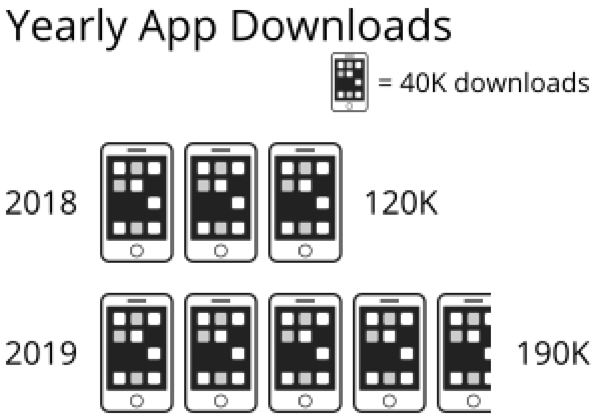
Common misuses

- Replacing bars with images

- It may be tempting to replace bars with pictures that depict what is being measured (e.g. bags of money for money amounts), be careful that you do not misrepresent your data in this way. If your choice of symbol scales both width and height with value, differences will look much larger than they actually are, since people will end up comparing the areas of the bars rather than just their widths or heights. In the example below, there is a 58% growth in downloads from 2018 to 2019. However, this growth is exaggerated with the icon-based representation, since the surface area of the 2019 icon is more than 2.5 times the size of the 2018 icon.


Yearly App Downloads
190K
120K
2018
2019

- If you feel the need to use icons to depict value, then a better – though still not great – option is to use the *pictogram chart* type instead. In a pictogram chart, each category’s value is indicated by a series of icons, with each icon representing a certain quantity. In a certain sense, this is like changing the texture of its corresponding bar to a repeating image. One major caution with this chart type is that it can make values harder to read, since the reader needs to perform some mental mathematics to gauge the relative values of each category.



Yearly App Downloads

Common bar chart options

- Horizontal bars vs. vertical bars

Advantages

- show each data category in a frequency distribution
- display relative numbers or proportions of multiple categories
- summarize a large data set in visual form
- clarify trends better than do tables
- estimate key values at a glance
- permit a visual check of the accuracy and reasonableness of calculations
- be easily understood due to widespread use in business and the media

Disadvantages

- require additional explanation
- be easily manipulated to yield false impressions
- fail to reveal key assumptions, causes, effects, or patterns

Question

?