Data Visualization

Visual Encodings :

color, size, shape, lines, axes, scaling, annotation

Taxonomy of data visualization(Some Types of charts, but not limited to):

Comparison charts – Bar chart, Box plots, Histograms, Gantt charts, Glyph chart, Sanky diagam, Word Cloud etc.

Hierarchies and relationships – Pie chart, stacked bar, Tree map etc.

Changes over time – Line chart, sparklines, candlestick/ohlc etc.

Connections and relationships – scatter lots, bubble plots, radial network, heat maps, etc.

Defining Data visualization

- Visual display of quantitative information
- Mapping data to visual elements
- Encoding data with size, shape, color...
- Storytelling / narrative elements

Types of Data Visualization

Exploratory

- Find insights
- Conversation between data and “you”

Explanatory

- Present insights

Exploratory data visualization

Statistical approaches:

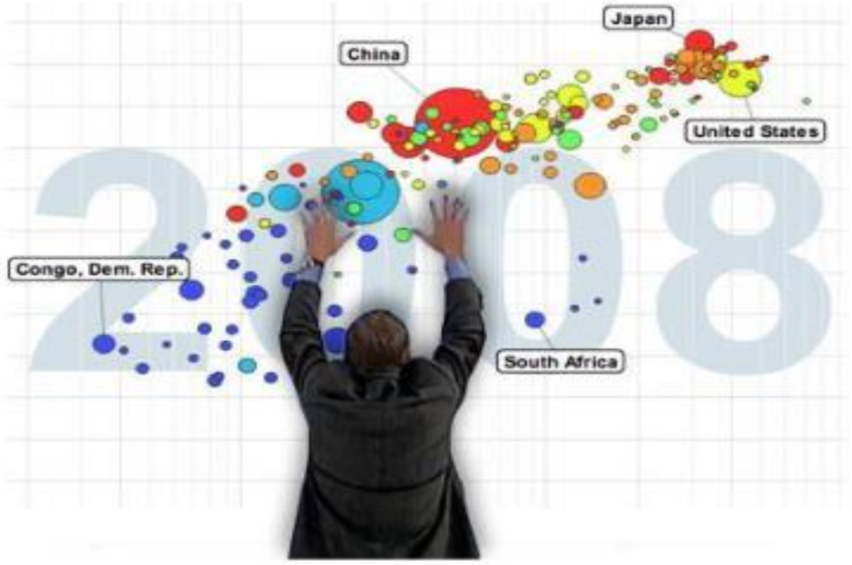
- Quantitative
  - Hypothesis testing
  - Analysis of variance (ANOVA)
  - Point estimates and confidence intervals
  - Least squares regression
- Graphical
  - Scatter plots
  - Histograms
  - Probability plots
  - Residual plots
  - Box plots
  - Block plots

Exploratory data visualization

Graphical analysis procedures:

- Testing assumptions
- Model selection
- Model validation
- Estimator selection
- Relationship identification
- Factor effect determination
- Outlier detection

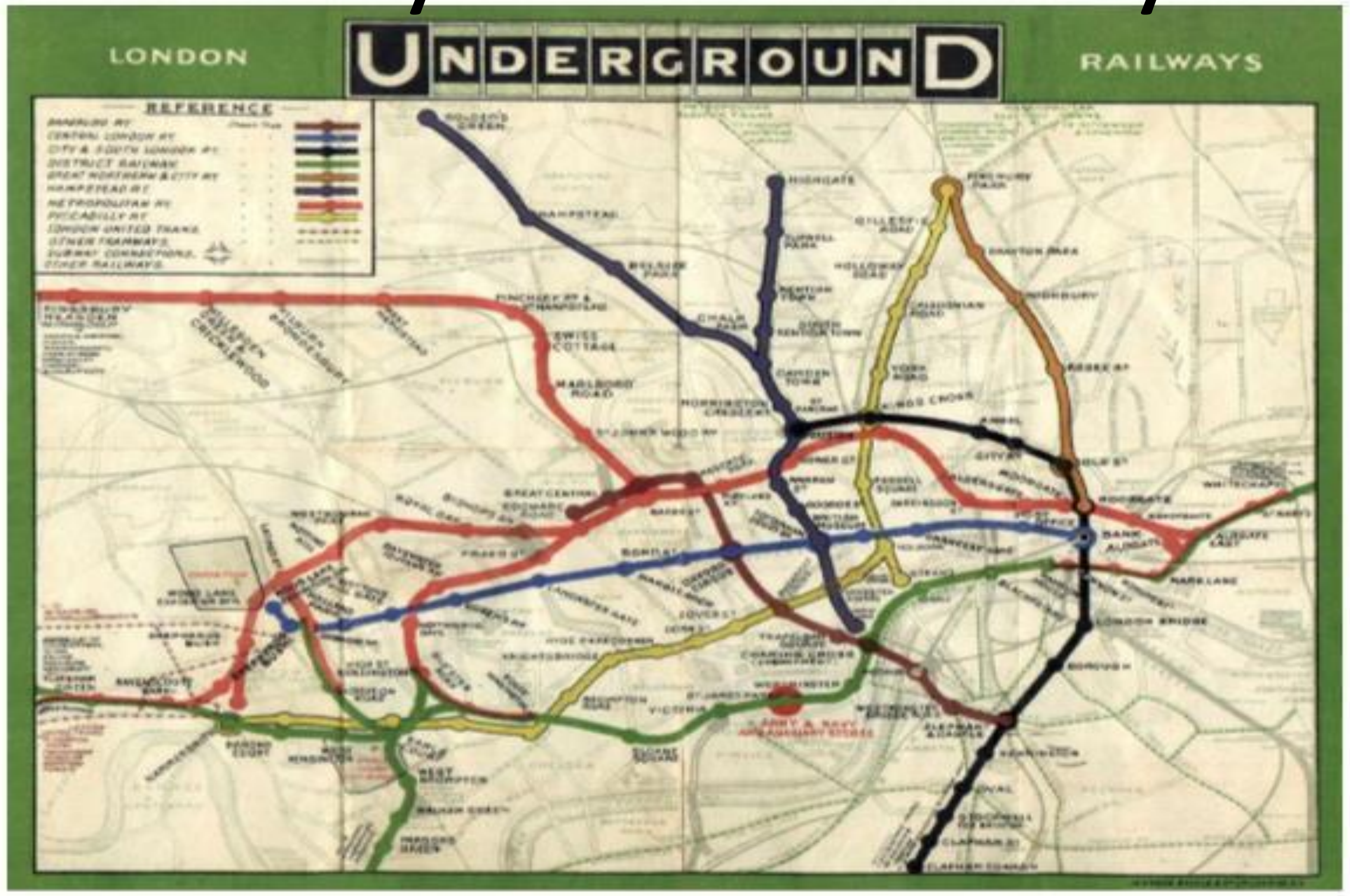
MUST USE for deriving insights from data

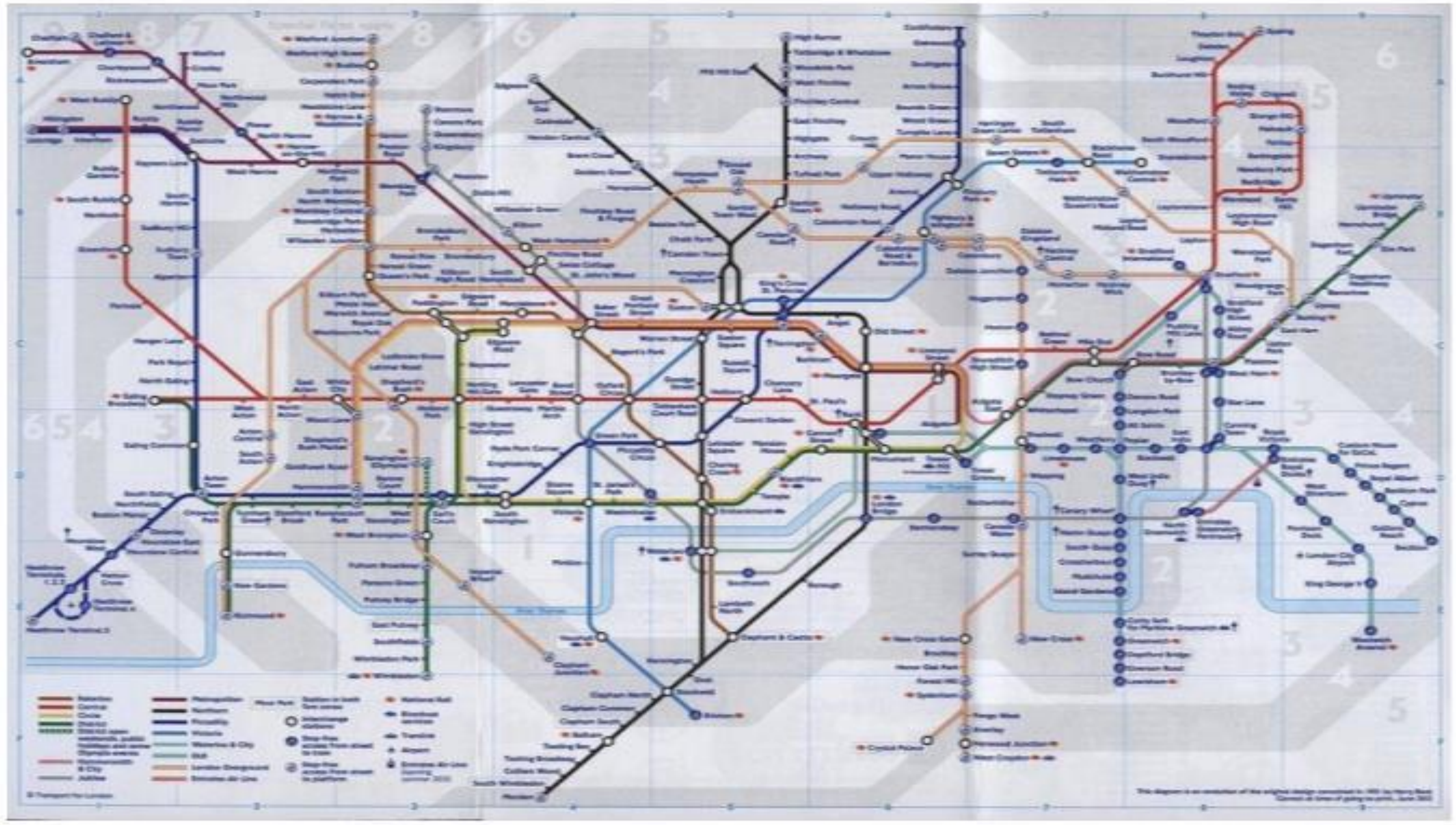


Explanatory data visualization

Visualization is both an art and science

- Harry Beck's subway map of London





Visual encoding of data

Data Types

- Quantitative
  - Continuous, Discrete
- Categorical
  - Nominal, Ordered, Interval

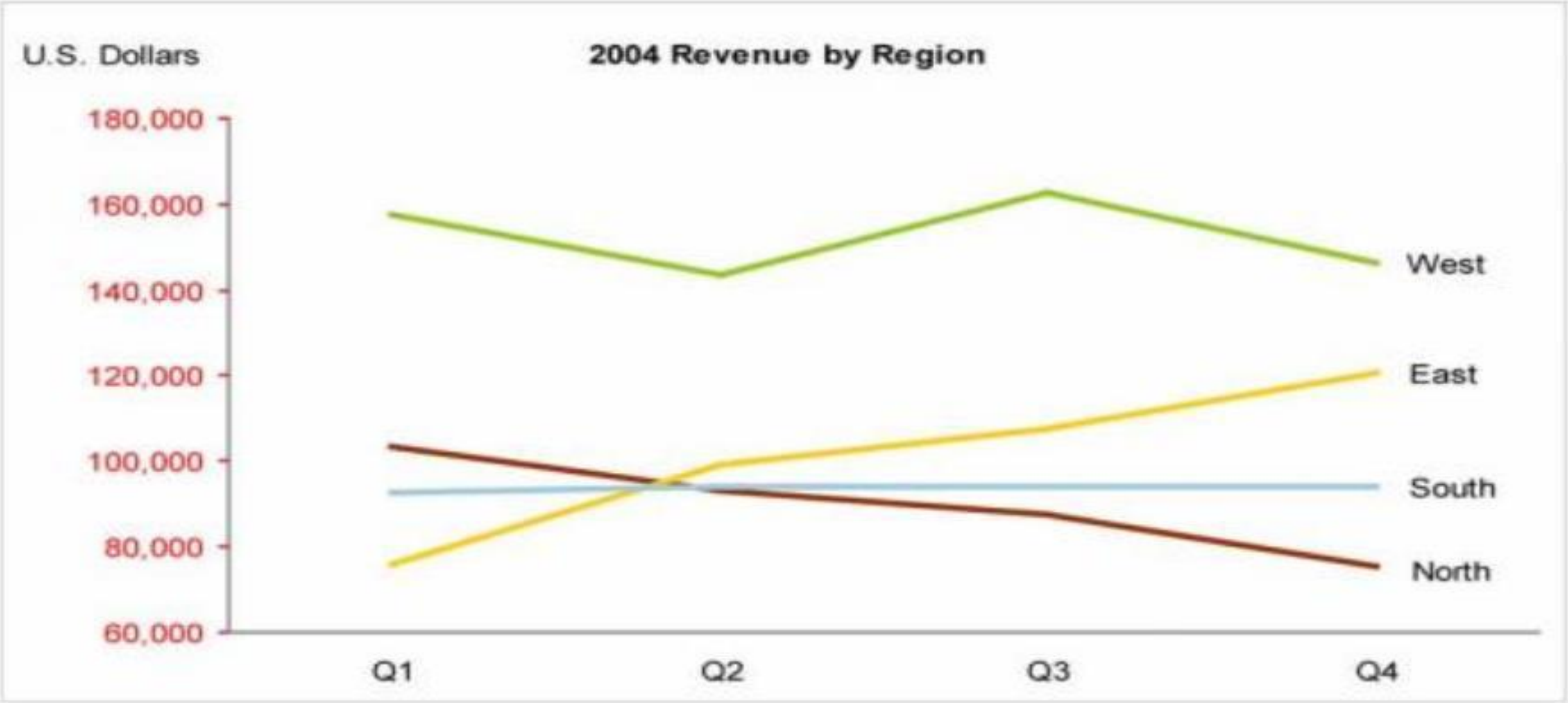
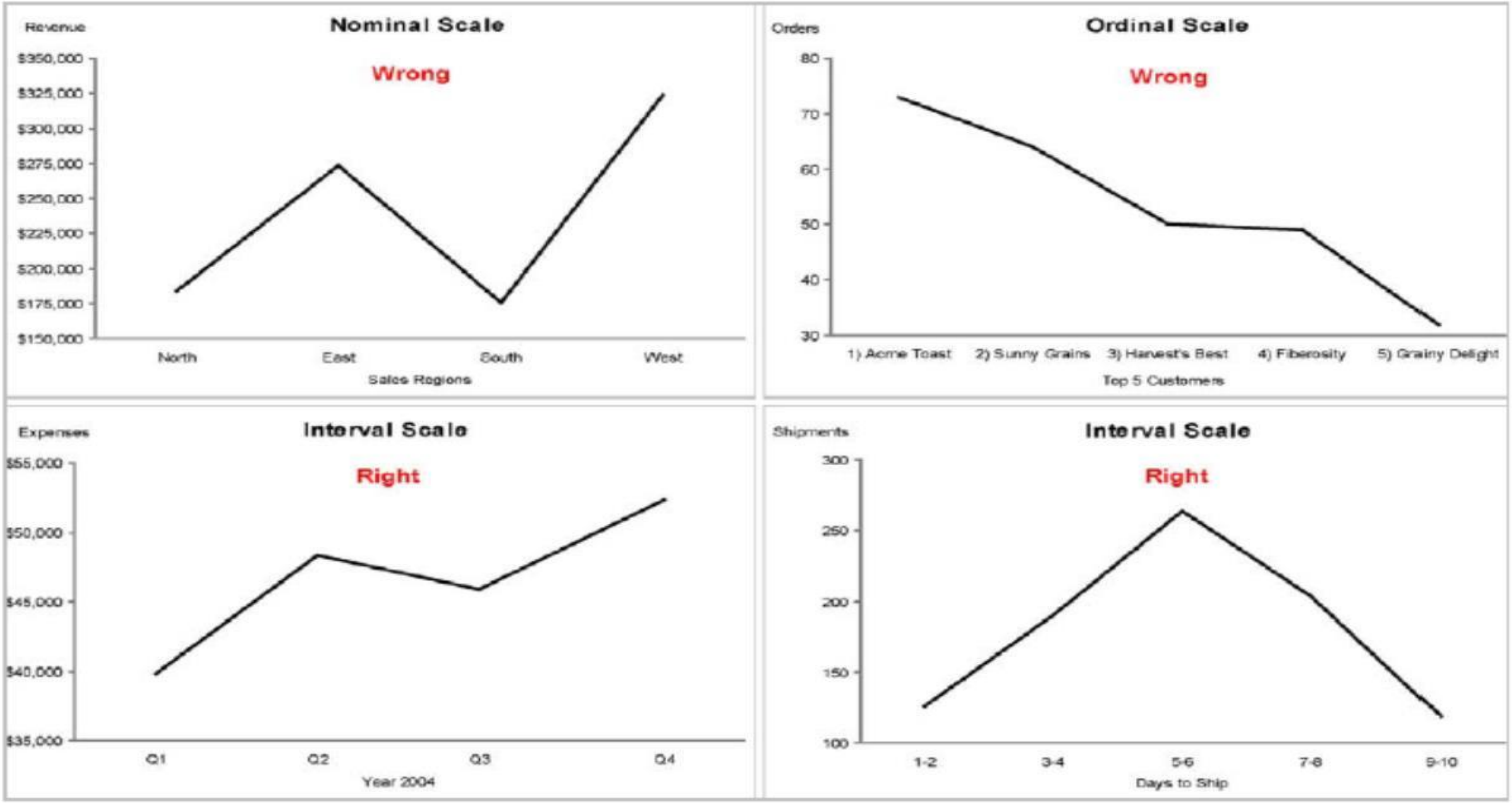


Figure 1: Illustration of the difference between quantitative data (red) and categorical data (black).

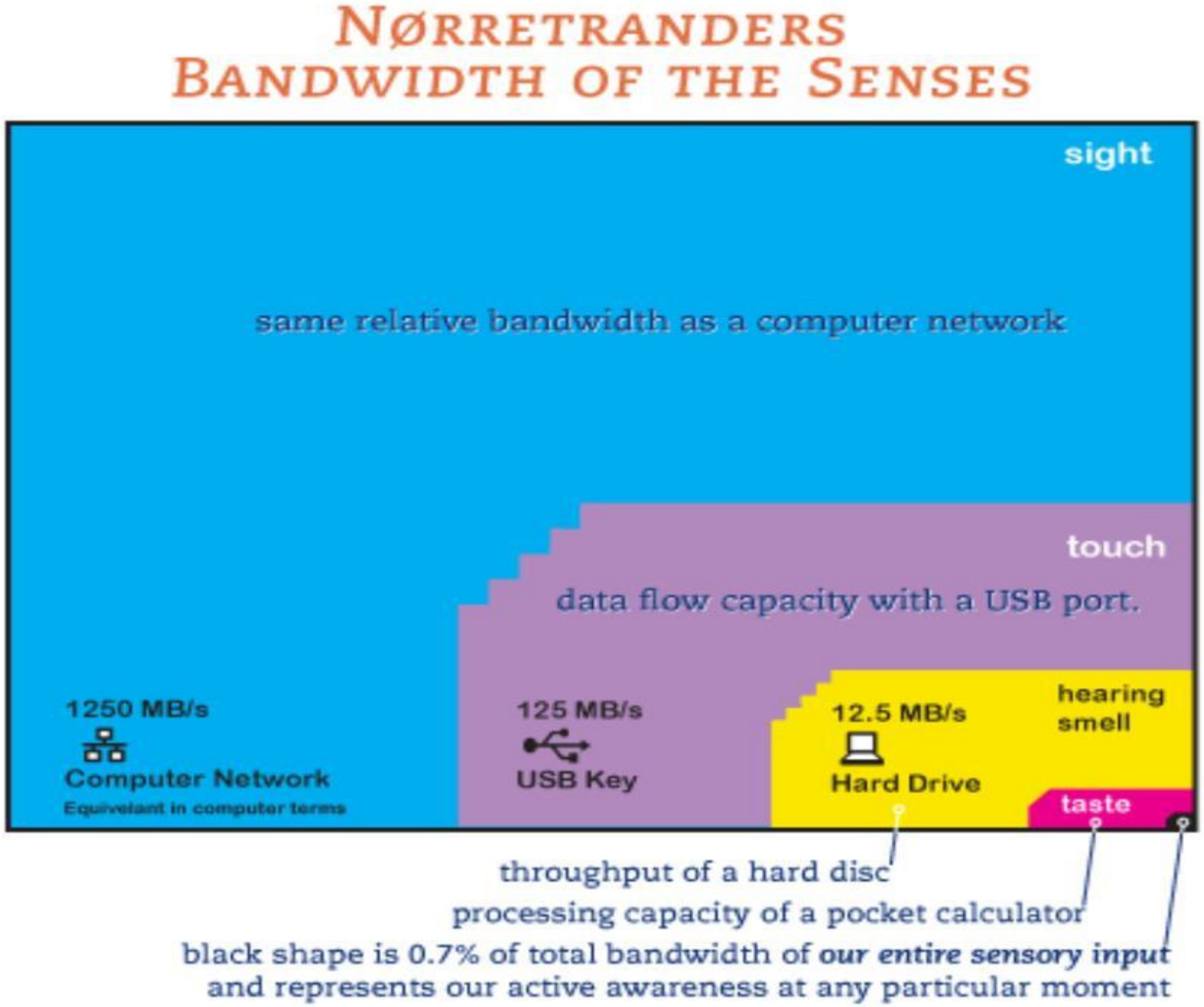
Visual encoding of data

Categorical scales and graph design



Visual encoding of data

Bandwidth of our senses: **[Tor Norretranders]**

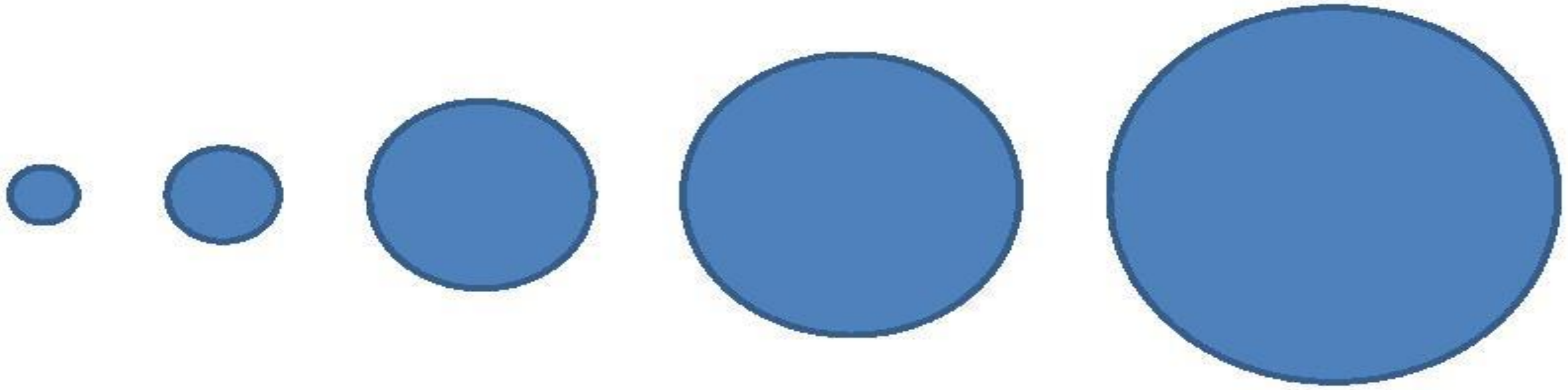


Source: http://chitownmediapsych.blogspot.ie/2010/09/context-will-be-king-working-title.html
²David McCandless : Information Is Beautiful

Visual encoding of data

Data → visual display elements

- Position x
- Position y
- Retinal variables
  - Size, Orientation **(ordered data)**
  - Color Hue, Shape **(nominal data)**
- Animation



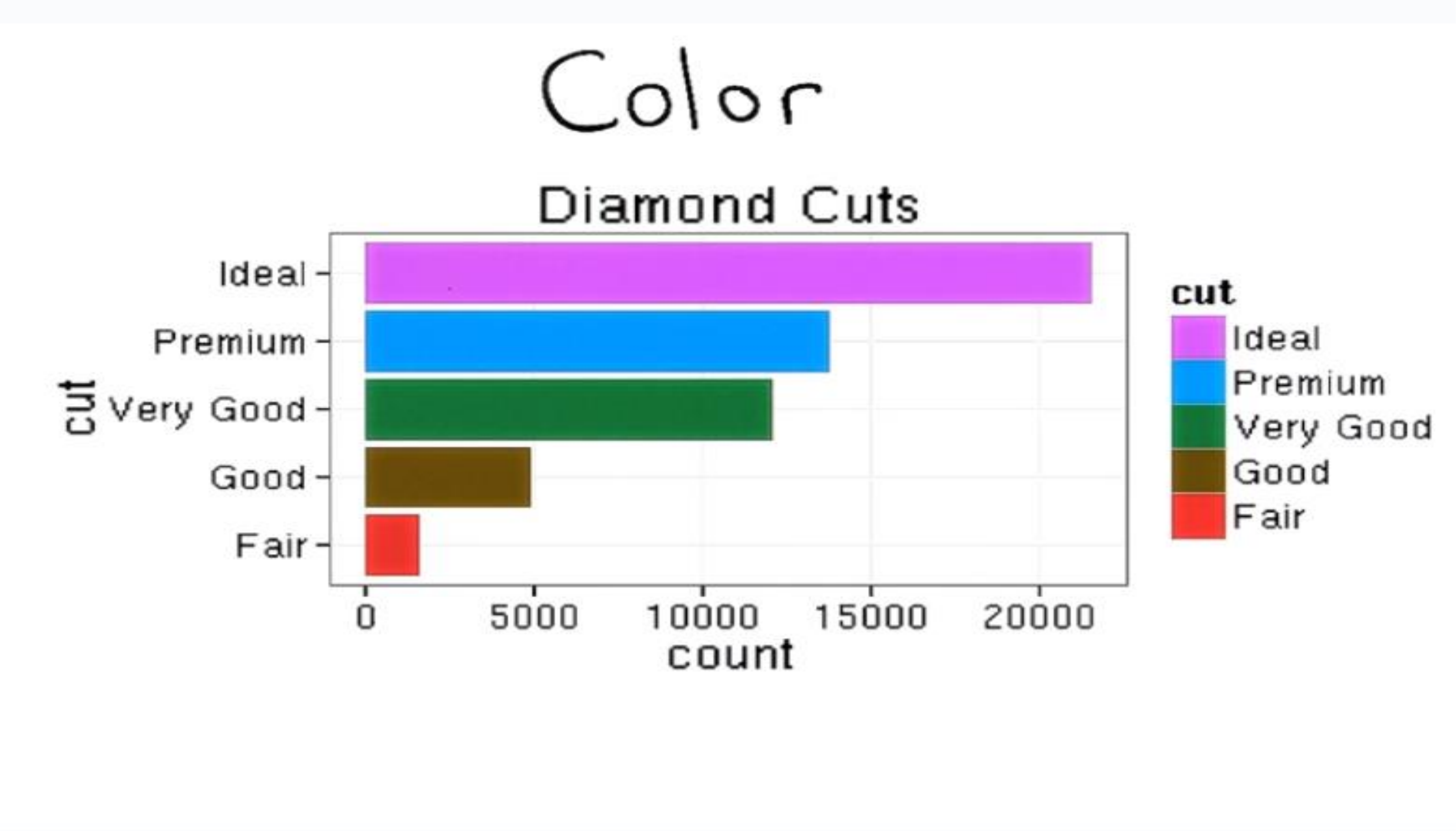
Visual encoding of data

Ranking visual display elements (framework):

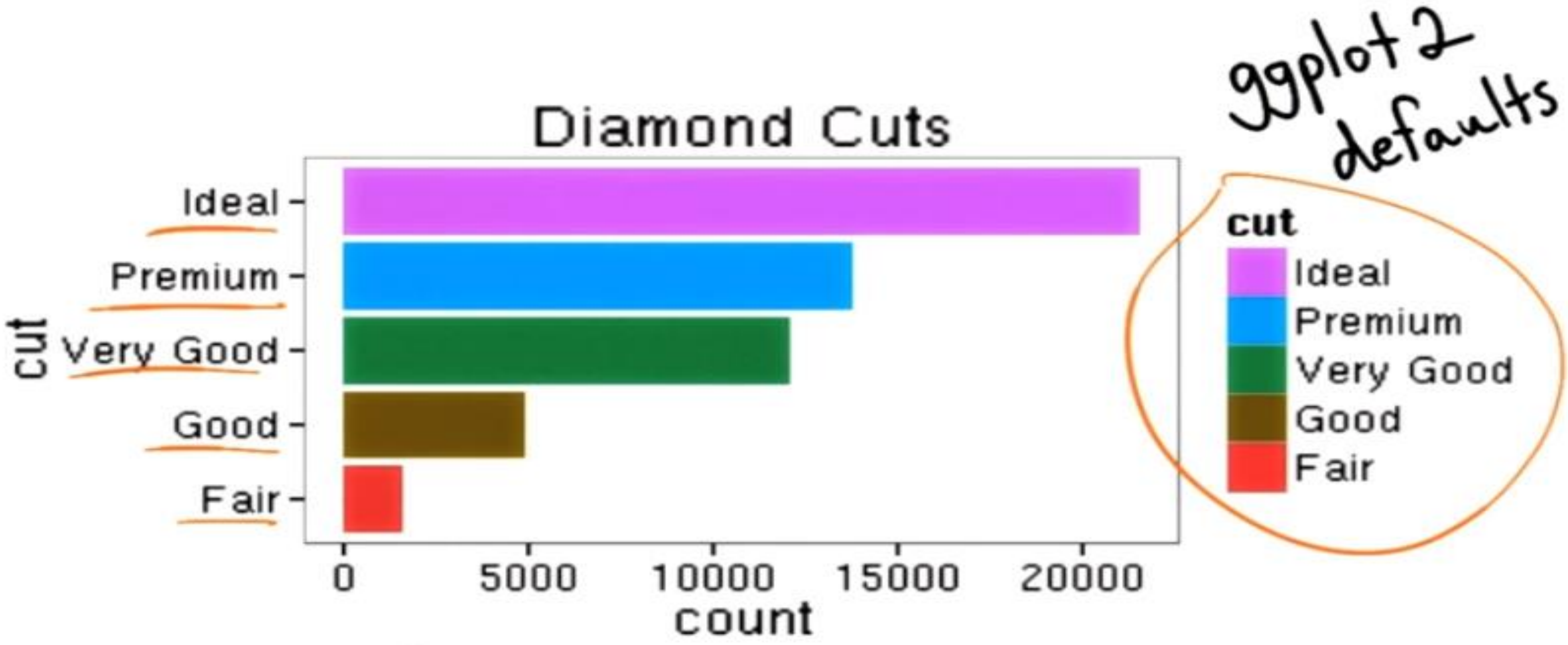
1. Position along a common-scale e.g. scatter plots
2. Position on identical but non-aligned scales
   E.g. multiple scatter plots
3. Length e.g. bar chart
4. Angle & Slope e.g. pie-chart
5. Area e.g. bubbles
6. Volume, density & color saturation e.g. heat-map
7. Color hue e.g. highlights

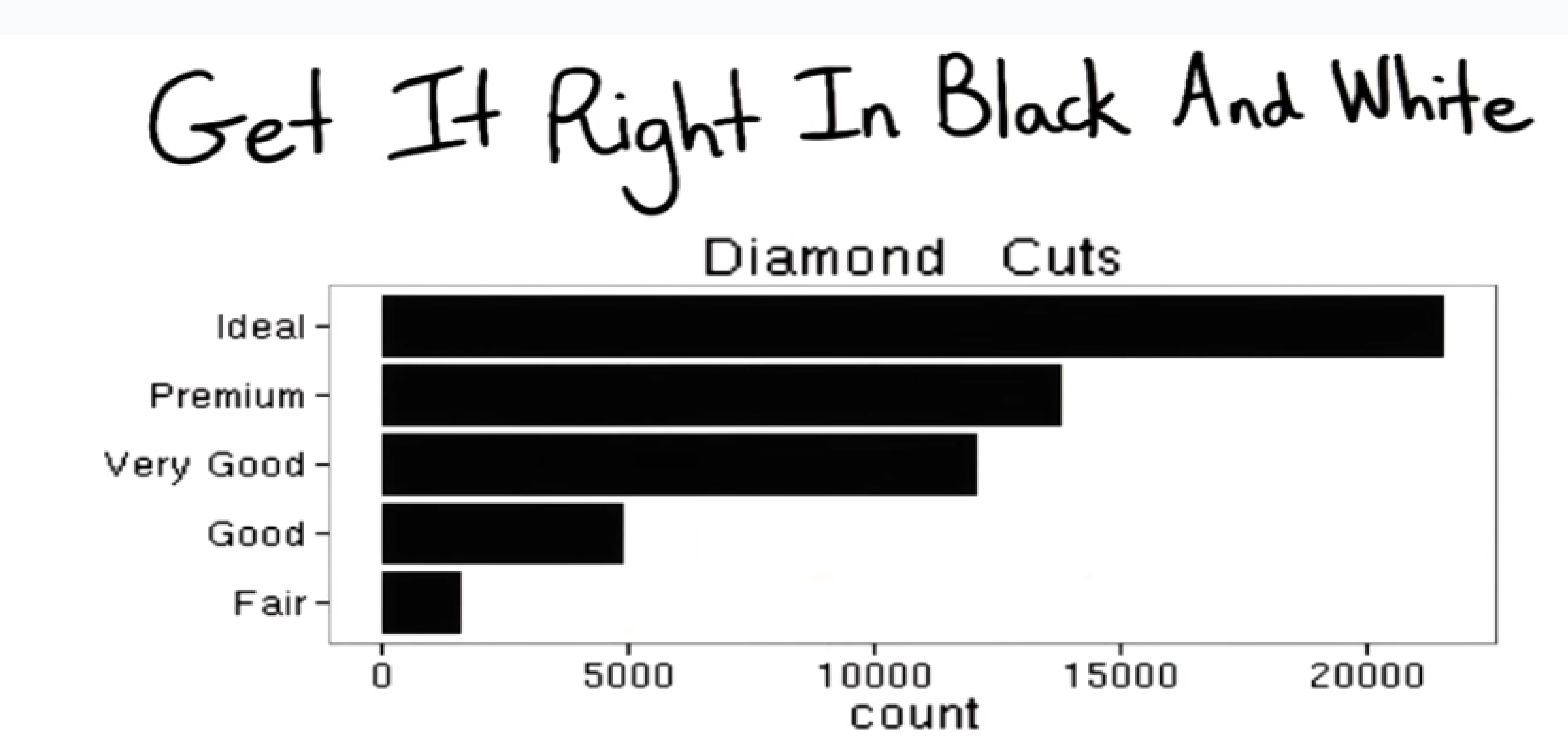
Ref. Graphical Perception & graphical methods for analyzing scientific data – William Cleveland & Robert McGill (1985)

Careful with color



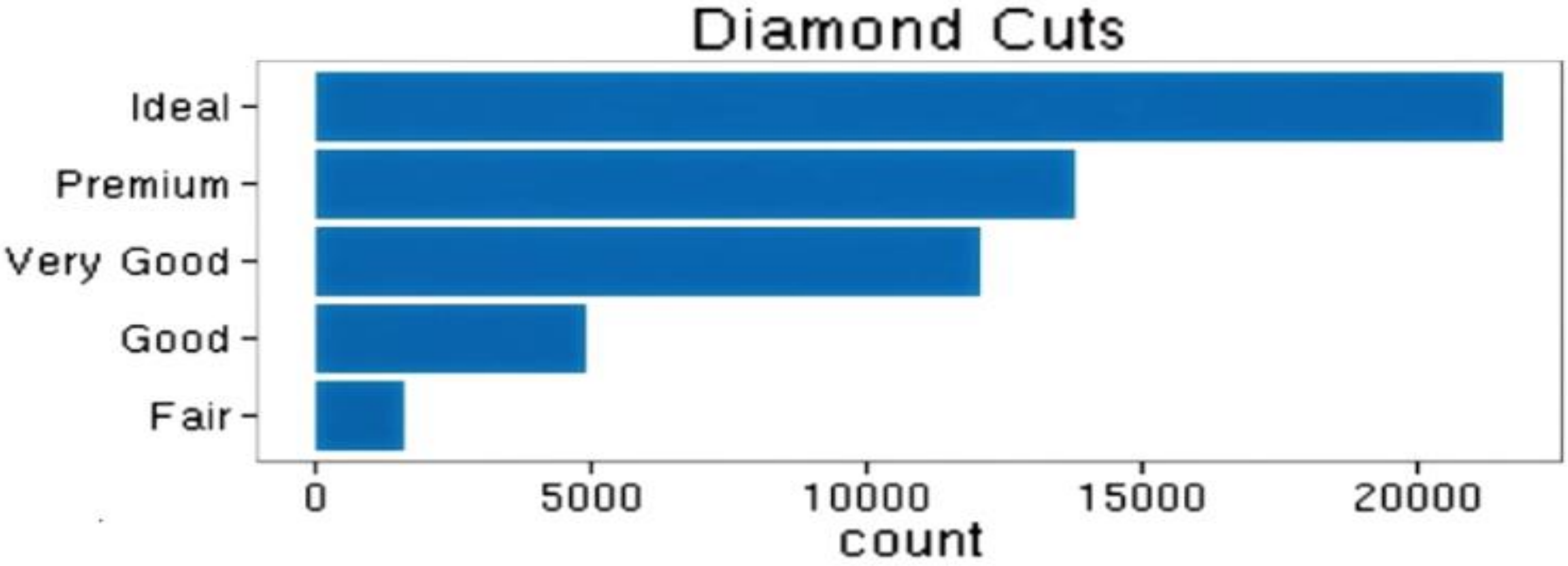
Redundant encodings



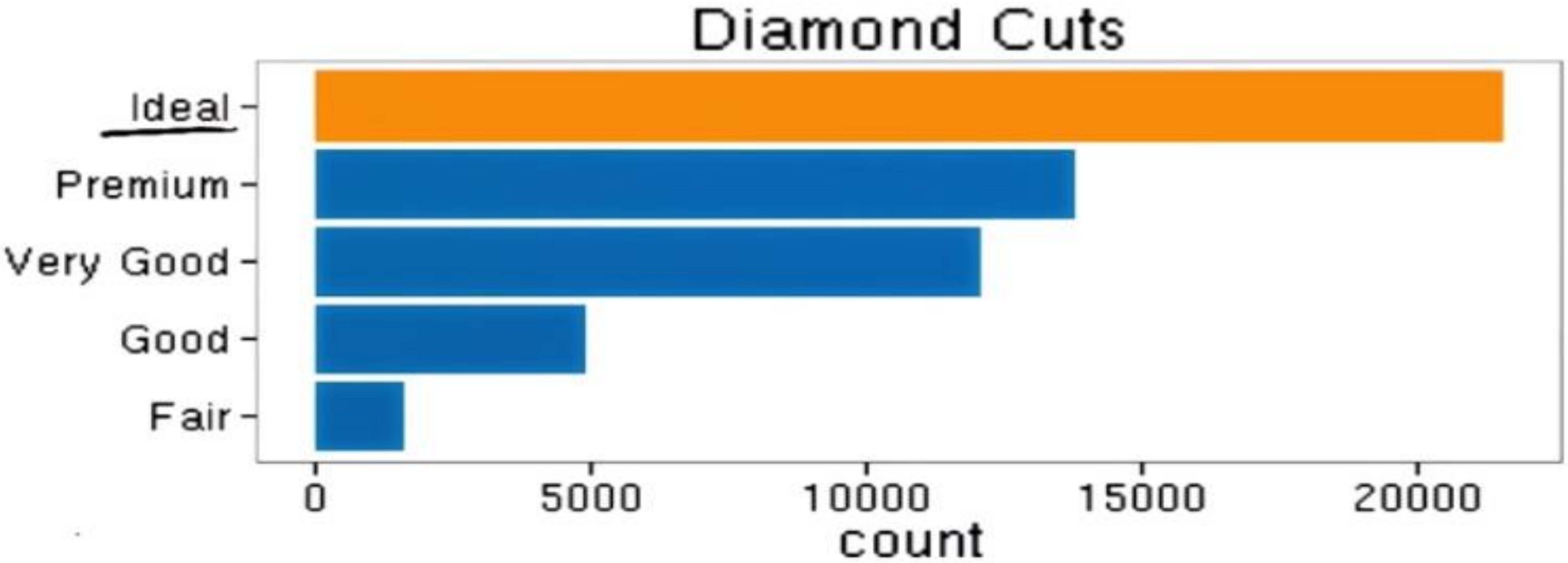


Get It Right In Black And White

Use Medium Hues or Pastels



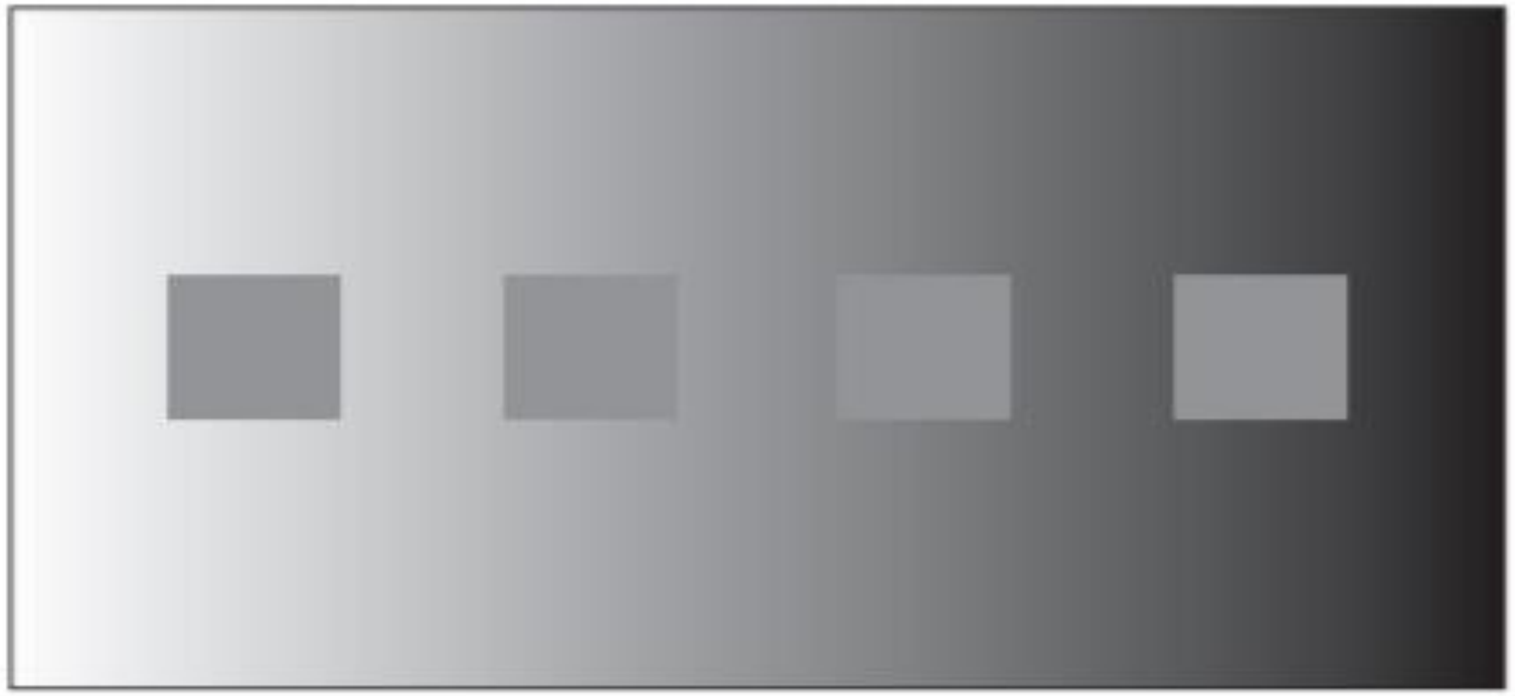
Use color to highlight.



Color in context

Rule #1 If you want different objects of the same color in a table or graph to look the same, make sure that the background—the color that surrounds them—is consistent.

Rule #2 If you want objects in a table or graph to be easily seen, use a background color that contrasts sufficiently with the object.

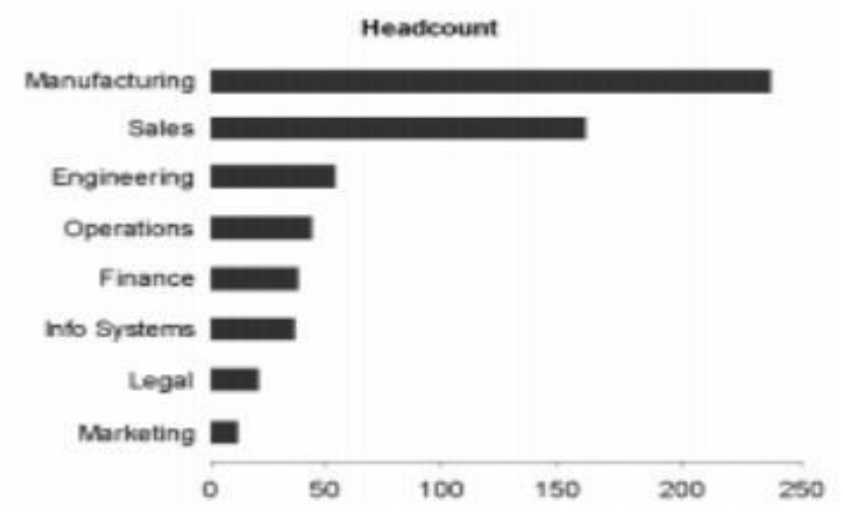


Design principles

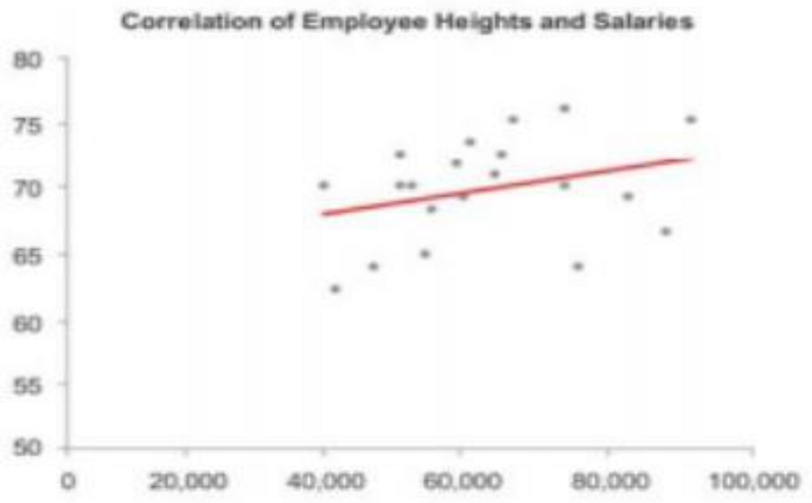
- Choose the right type of chart
  - Trends / Change over time → Line charts
  - Distributions → Histograms
  - Summary Information → Table
  - Relationships → Scatter Plots
- Get it right in black & white (before adding color)
- Prefer 2D to 3D for statistical charts
- Use color to highlight
- Avoid rainbow palette
- Avoid chartjunk : “less is more”
- Try to have a high data-ink ratio

Design principles

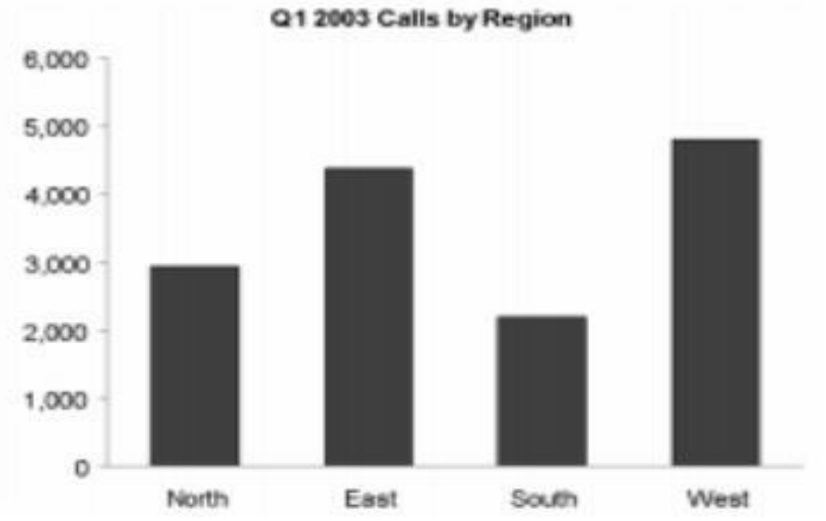
- Choose the right type of chart



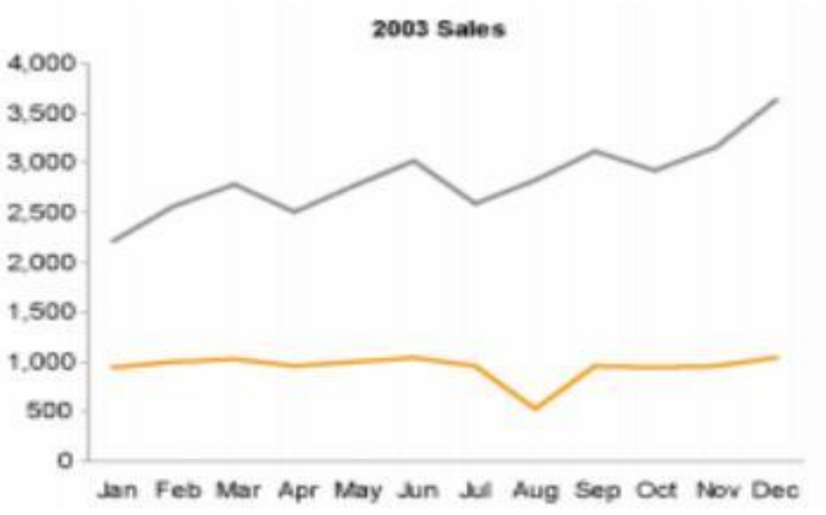
Ranking



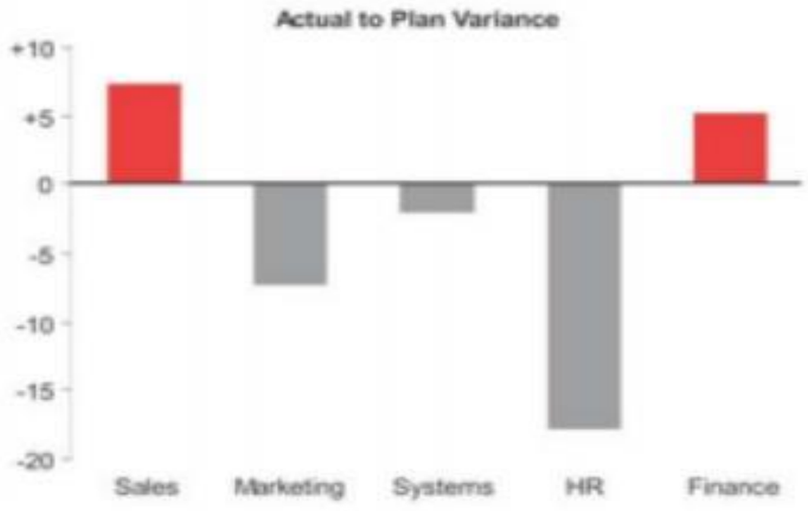
Correlation



Nominal comparison

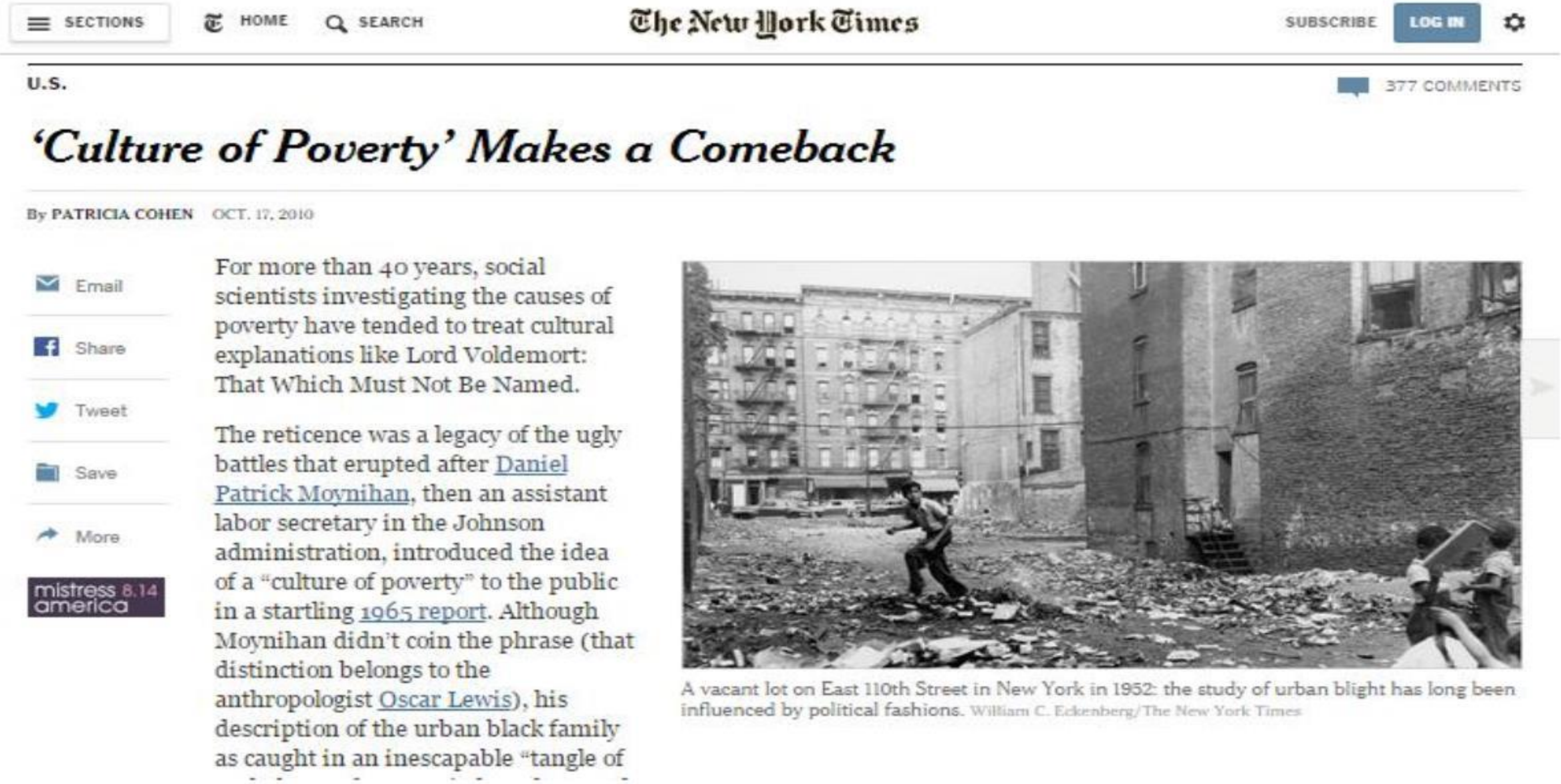


Time-series



Deviation

Narrative structures



U.S.

‘Culture of Poverty’ Makes a Comeback

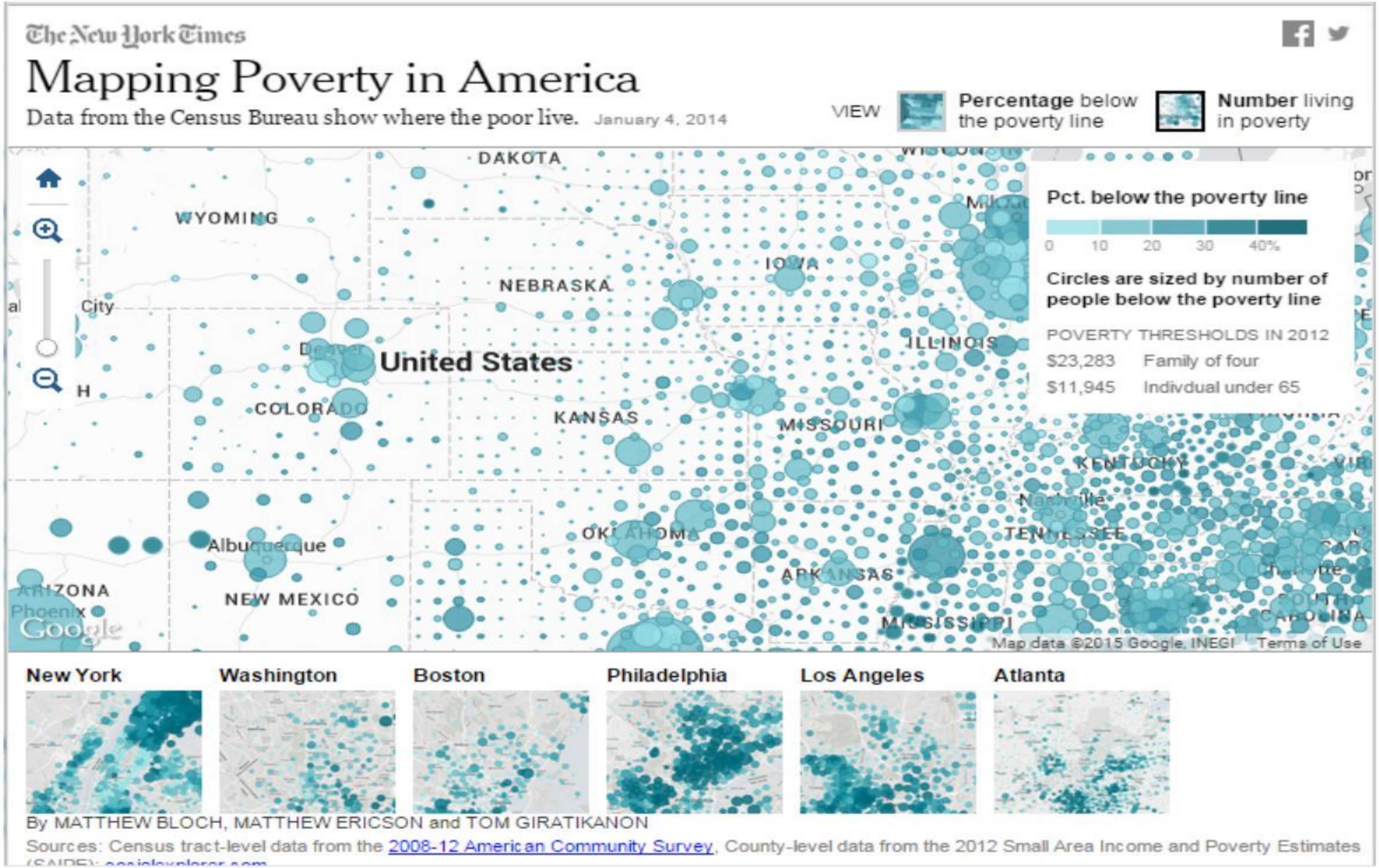
By PATRICIA COHEN OCT. 17, 2010

For more than 40 years, social scientists investigating the causes of poverty have tended to treat cultural explanations like Lord Voldemort: That Which Must Not Be Named.

The reticence was a legacy of the ugly battles that erupted after Daniel Patrick Moynihan, then an assistant labor secretary in the Johnson administration, introduced the idea of a “culture of poverty” to the public in a startling 1965 report. Although Moynihan didn’t coin the phrase (that distinction belongs to the anthropologist Oscar Lewis), his description of the urban black family as caught in an inescapable “tangle of

A vacant lot on East 110th Street in New York in 1952: the study of urban blight has long been influenced by political fashions. William C. Eckenberg/The New York Times

Narrative structures



The New York Times

Mapping Poverty in America

Data from the Census Bureau show where the poor live. January 4, 2014

VIEW **Percentage** below the poverty line **Number** living in poverty

Pct. below the poverty line

0 10 20 30 40%

Circles are sized by number of people below the poverty line

POVERTY THRESHOLDS IN 2012

$23,283	Family of four
$11,945	Indivdual under 65

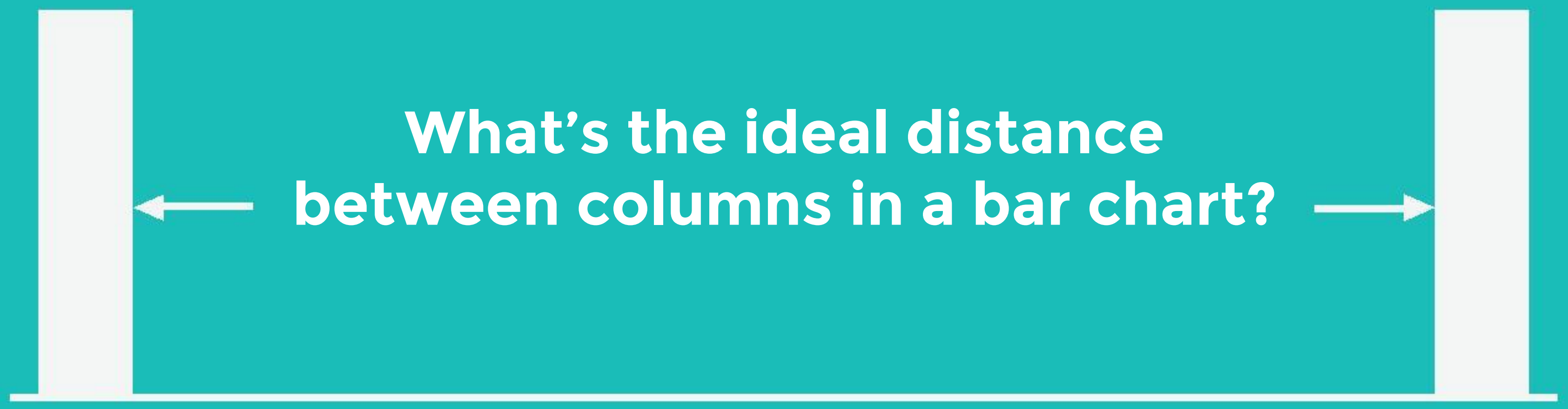
New York Washington Boston Philadelphia Los Angeles Atlanta

By MATTHEW BLOCH, MATTHEW ERICSON and TOM GIRATIKANON

Sources: Census tract-level data from the 2008-12 American Community Survey, County-level data from the 2012 Small Area Income and Poverty Estimates (SAIPE); socialexplorer.com

Your data is only as good as your ability to understand and communicate it, which is why choosing the right visualization is essential.

If your data is misrepresented or presented ineffectively, key insights and understanding are lost, which hurts both your message and your reputation. The good news is that you don't need a PhD in statistics to crack the data visualization code. This guide will walk you through the most common charts and visualizations, help you choose the right presentation for your data, and give you practical design tips and tricks to make sure you avoid rookie mistakes. It's everything you need to help your data make a big impact.



What's the ideal distance between columns in a bar chart?

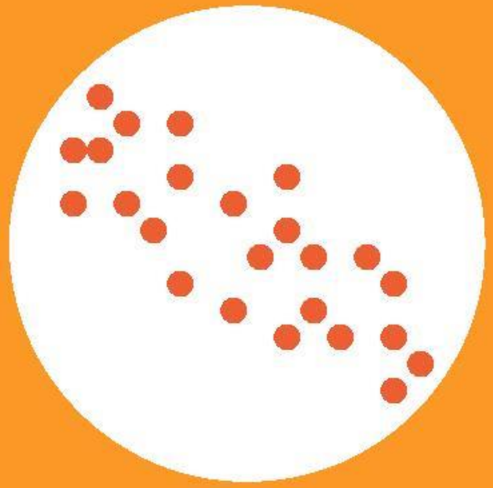
You're about to find out.

FINDING THE STORY IN YOUR DATA

Information can be visualized in a number of ways, each of which can provide a specific insight. When you start to work with your data, it's important to identify and understand the story you are trying to tell and the relationship you are looking to show. Knowing this information will help you select the proper visualization to best deliver your message.

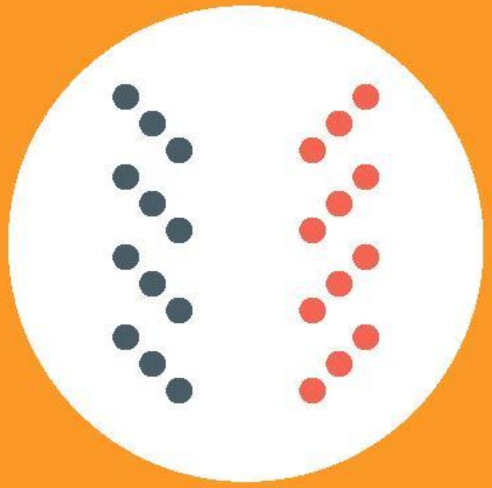
When analyzing data, search for patterns or interesting insights that can be a good starting place for finding your story, such as:

TRENDS



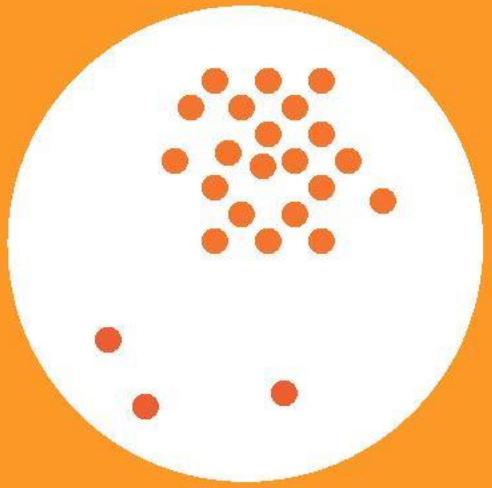
Example:
Ice cream sales over time

CORRELATIONS



Example:
Ice cream sales vs. temperature

OUTLIERS

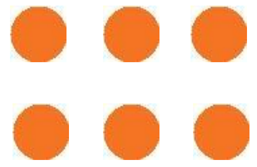


Example:
Ice cream sales in an unusual region

KNOW YOUR DATA

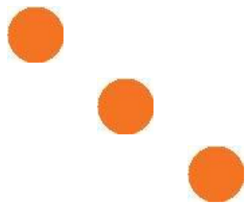
Before understanding visualizations, you must understand the types of data that can be visualized and their relationships to each other. Here are some of the most common you are likely to encounter.

DATA TYPES



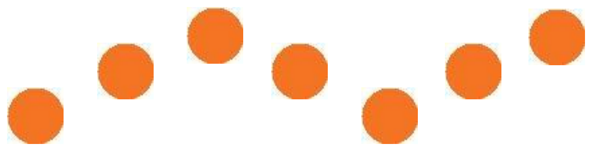
QUANTITATIVE

Data that can be counted or measured; all values are numerical.



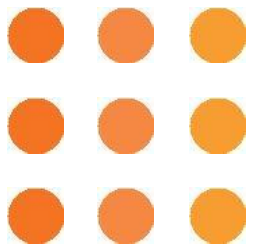
DISCRETE

Numerical data that has a finite number of possible values. Example: Number of employees in the office.



CONTINUOUS

Data that is measured and has a value within a range. Example: Rainfall in a year.



CATEGORICAL

Data that can be sorted according to group or category. Example: Types of products sold.

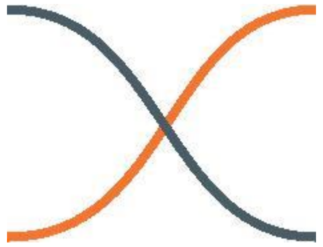
DATA RELATIONSHIPS

NOMINAL COMPARISON

This is a simple comparison of the quantitative values of subcategories. Example: Number of visitors to various websites.

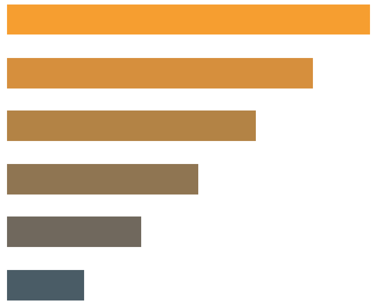
TIME-SERIES

This tracks changes in values of a consistent metric over time. Example: Monthly sales.



CORRELATION

This is data with two or more variables that may demonstrate a positive or negative correlation to each other. Example: Salaries according to education level.



RANKING

This shows how two or more values compare to each other in relative magnitude. Example: Historic weather patterns, ranked from the hottest months to the coldest.

DEVIATION

This examines how data points relate to each other, particularly how far any given data point differs from the mean. Example: Amusement park tickets sold on a rainy day vs. a regular day.

DISTRIBUTION

This shows data distribution, often around a central value. Example: Heights of players on a basketball team.

PART-TO-WHOLE RELATIONSHIPS

This shows a subset of data compared to the larger whole. Example: Percentage of customers purchasing specific products.

Now that you've got a handle on the most common data types and relationships you'll most likely have to work with, let's dive into the different ways you can visualize that data to get your point across.

GUIDE TO CHART TYPES

In this section, we'll cover the uses, variations, and best practices for some of the most common data visualizations:

BAR CHART

PIE CHART

LINE CHART

AREA CHART

SCATTER PLOT

BUBBLE CHART

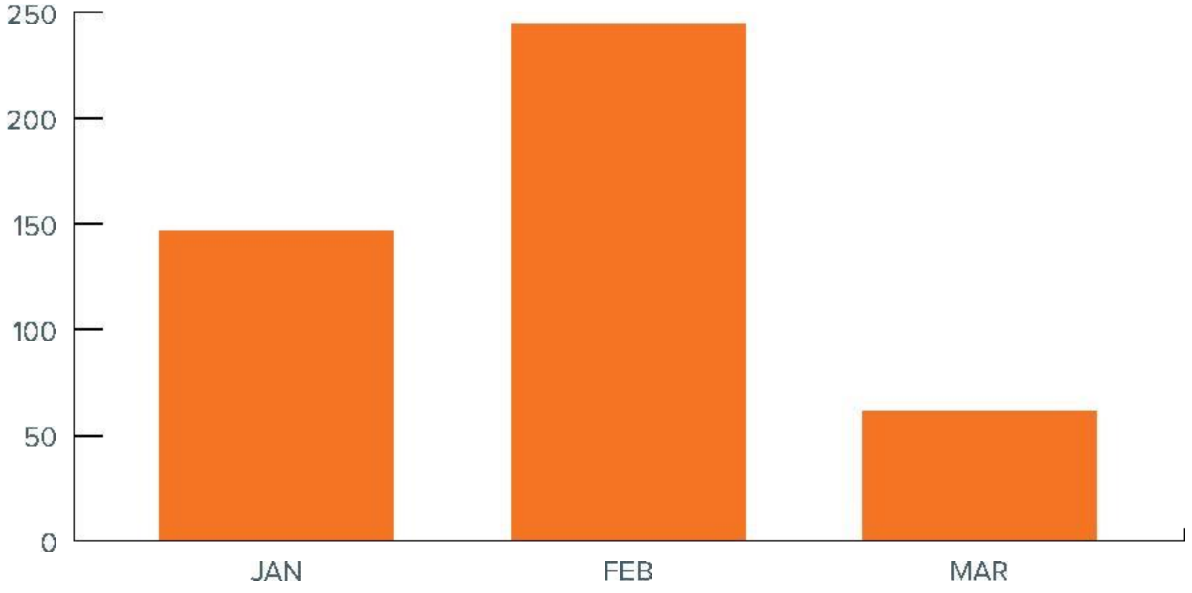
HEAT MAP

BAR CHART

Bar charts are very versatile. They are best used to show change over time, compare different categories, or compare parts of a whole.

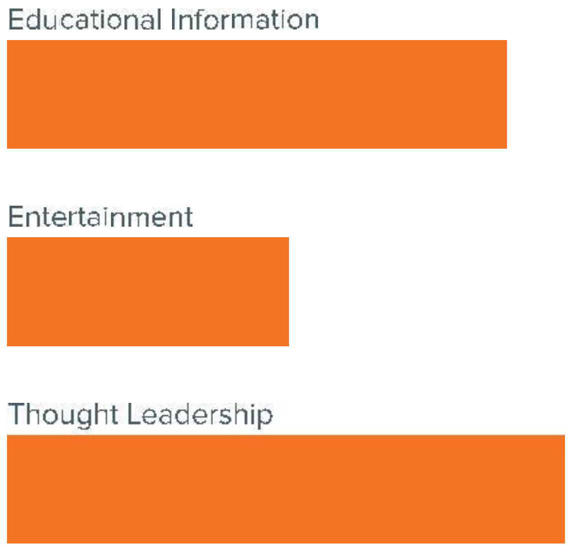
VARIATIONS OF BAR CHARTS

PAGE VIEWS, BY MONTH



VERTICAL (COLUMN CHART)
Best used for chronological data (time-series should always run left to right), or when visualizing negative values below the x-axis.

CONTENT PUBLISHED, BY CATEGORY

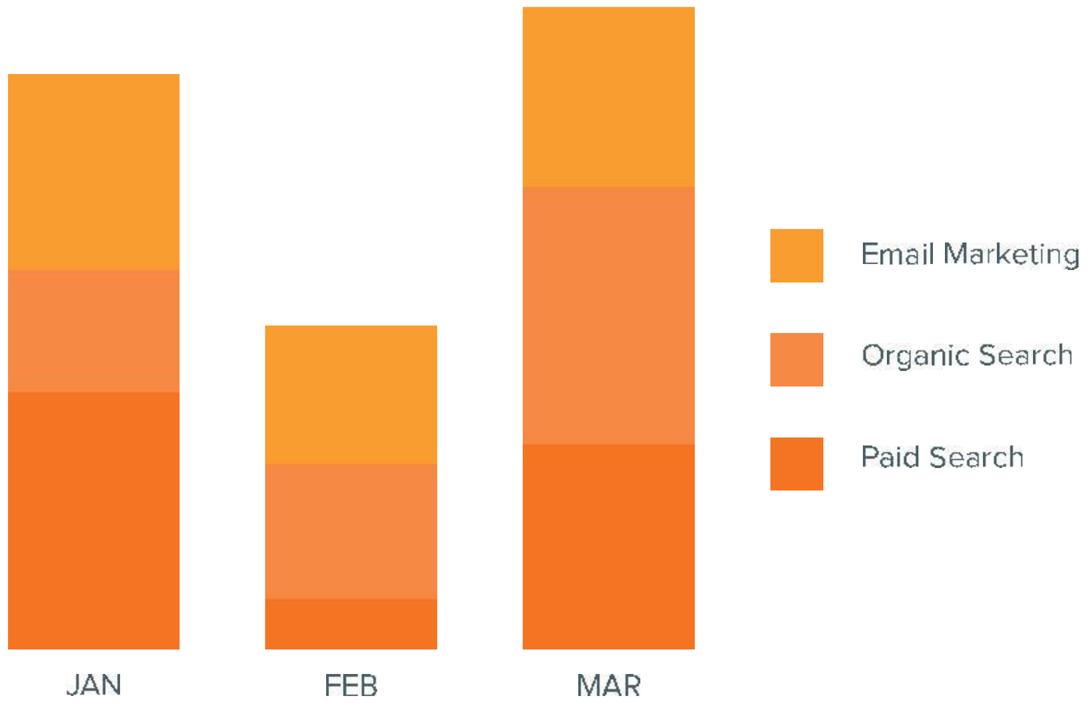


HORIZONTAL
Best used for data with long category labels.

BAR CHART

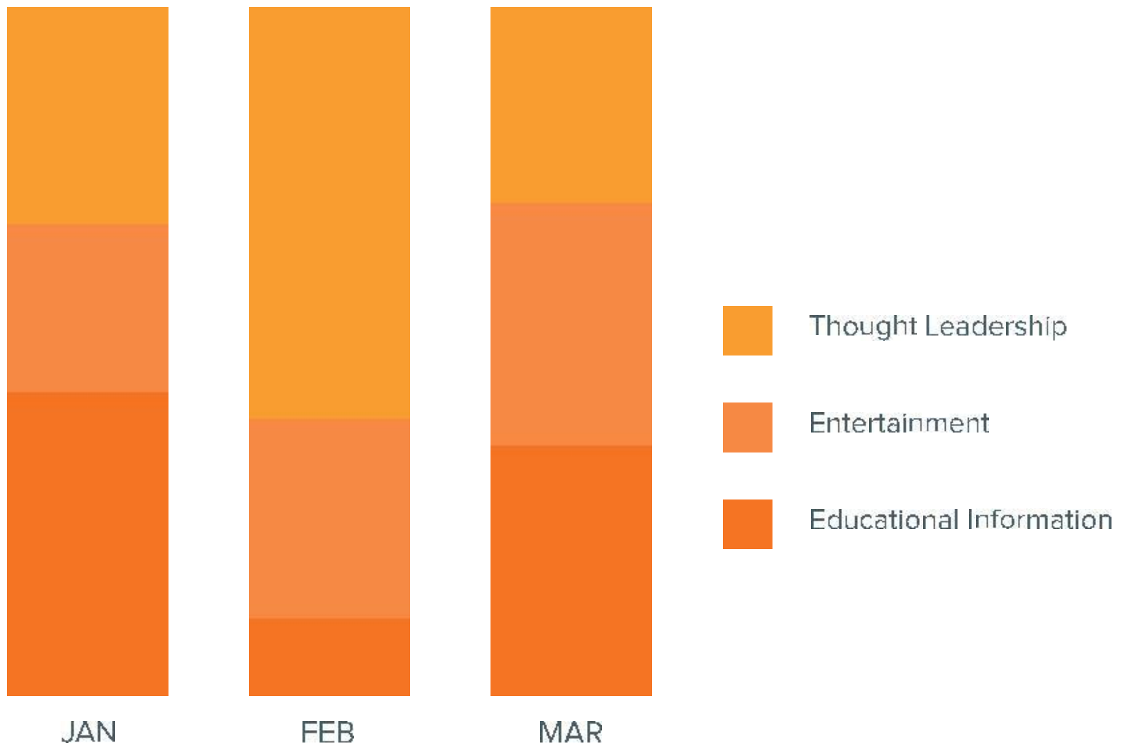
VARIATIONS OF BAR CHARTS (CONT.)

MONTHLY TRAFFIC, BY SOURCE



STACKED
Best used when there is a need to compare multiple part-to-whole relationships. These can use discrete or continuous data, oriented either vertically or horizontally.

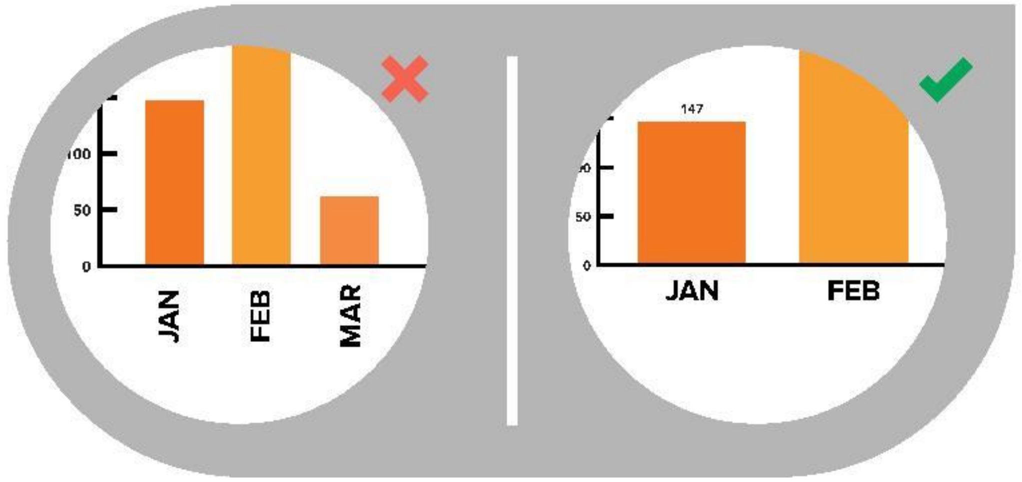
PERCENTAGE OF CONTENT PUBLISHED, BY MONTH



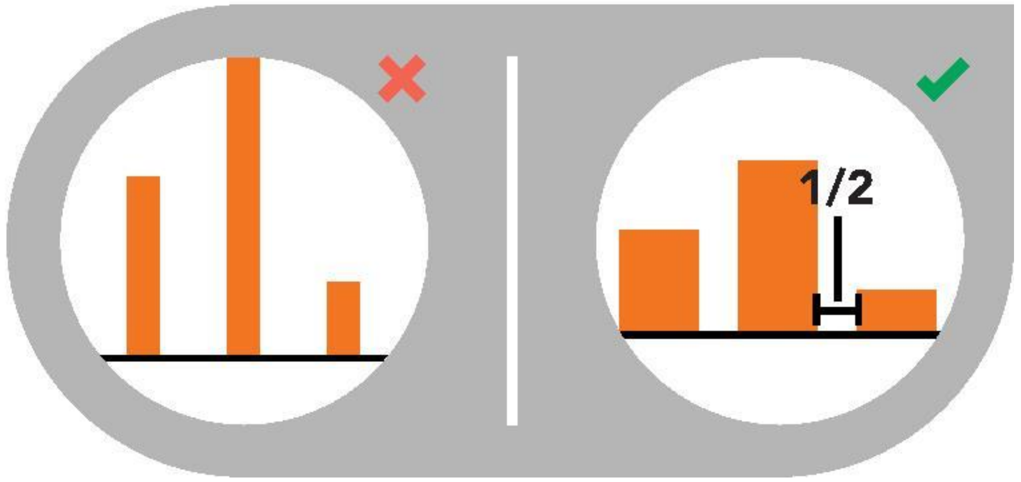
100% STACKED
Best used when the total value of each category is unimportant and percentage distribution of subcategories is the primary message.

BAR CHART

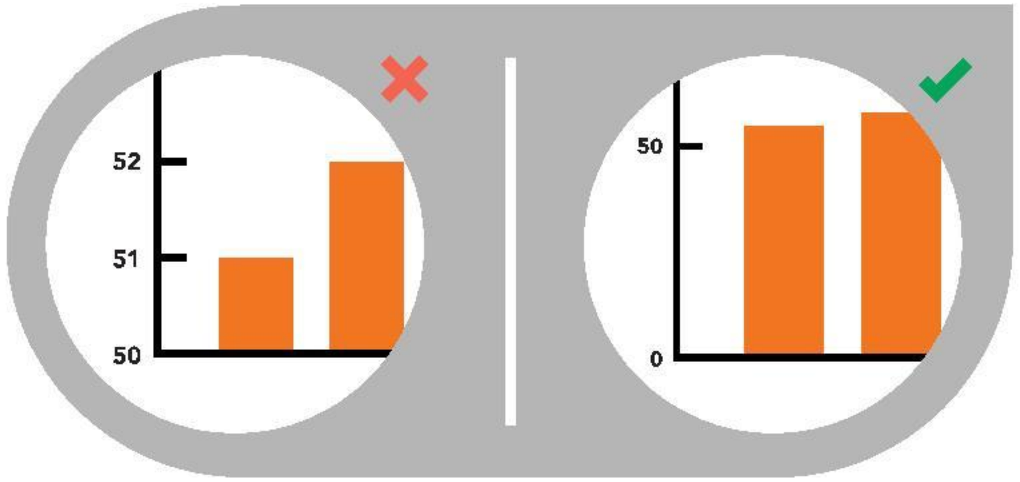
DESIGN BEST PRACTICES



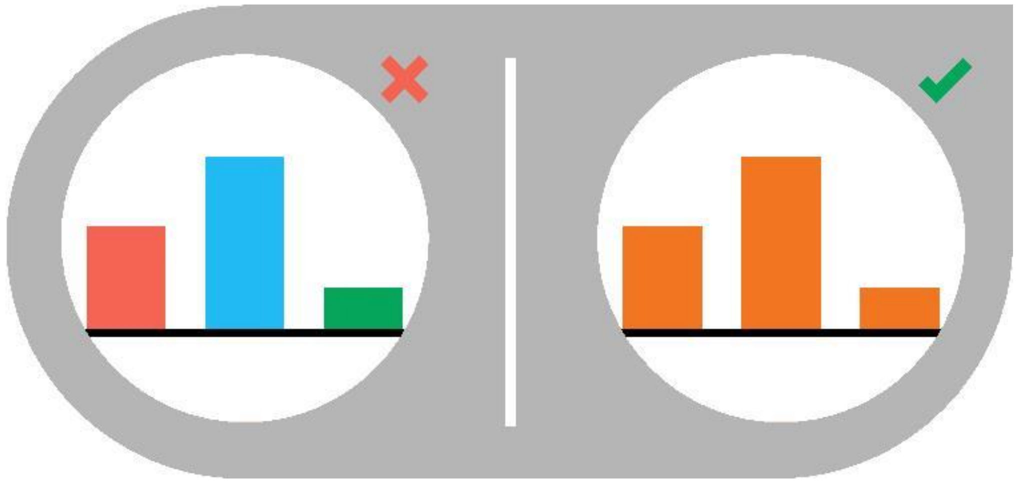
USE HORIZONTAL LABELS
Avoid steep diagonal or vertical type, as it can be difficult to read.



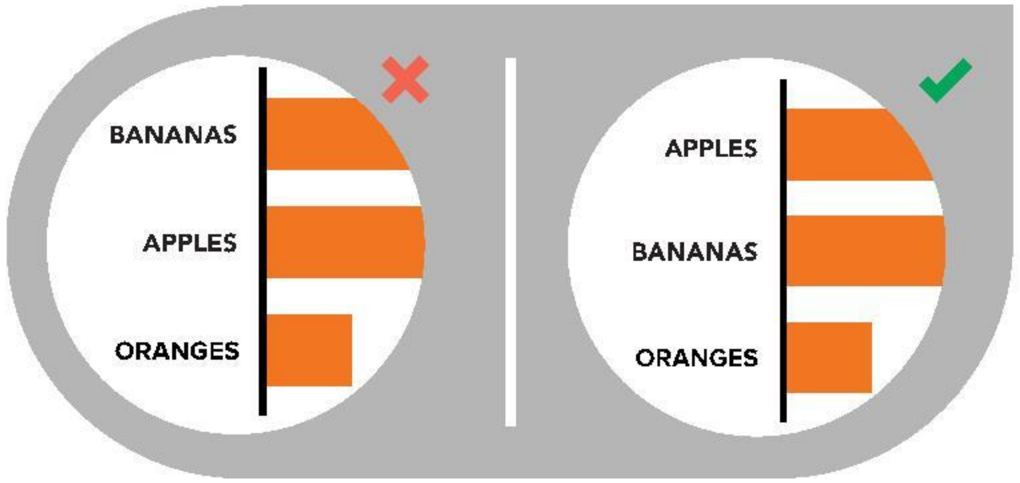
SPACE BARS APPROPRIATELY
Space between bars should be ½ bar width.



START THE Y-AXIS VALUE AT 0
Starting at a value above zero truncates the bars and doesn’t accurately reflect the full value.



USE CONSISTENT COLORS
Use one color for bar charts. You may use an accent color to highlight a significant data point.



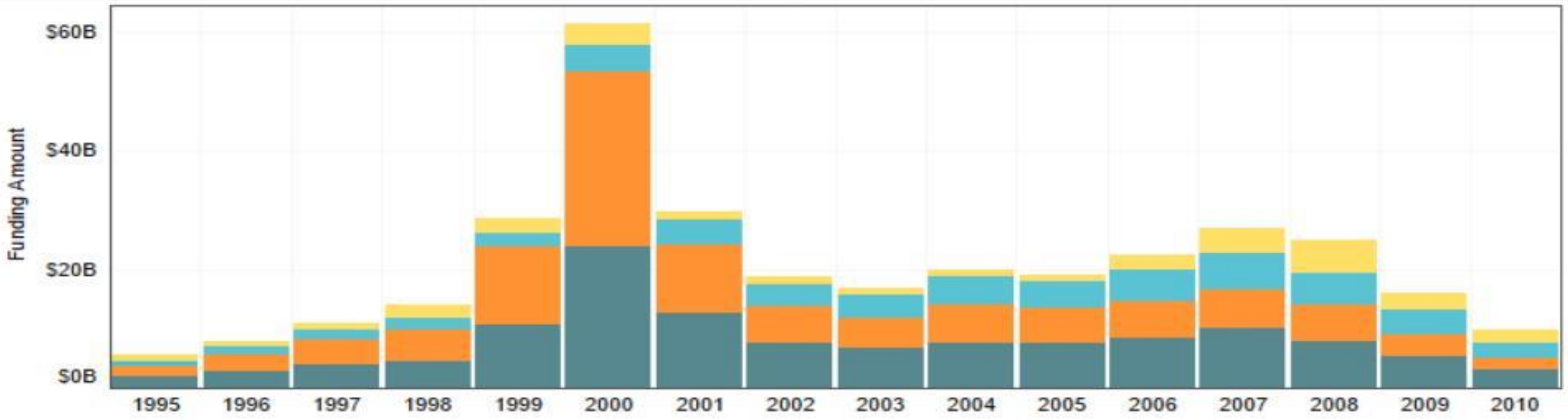
ORDER DATA APPROPRIATELY
Order categories alphabetically, sequentially, or by value.

Venture Financing

🟦 Hardware 🟧 Software 🟦 Biotech 🟨 Industrial

Although software funding has dramatically declined after the dot-com period, it still receives more funding than its competing sectors.

Funding History

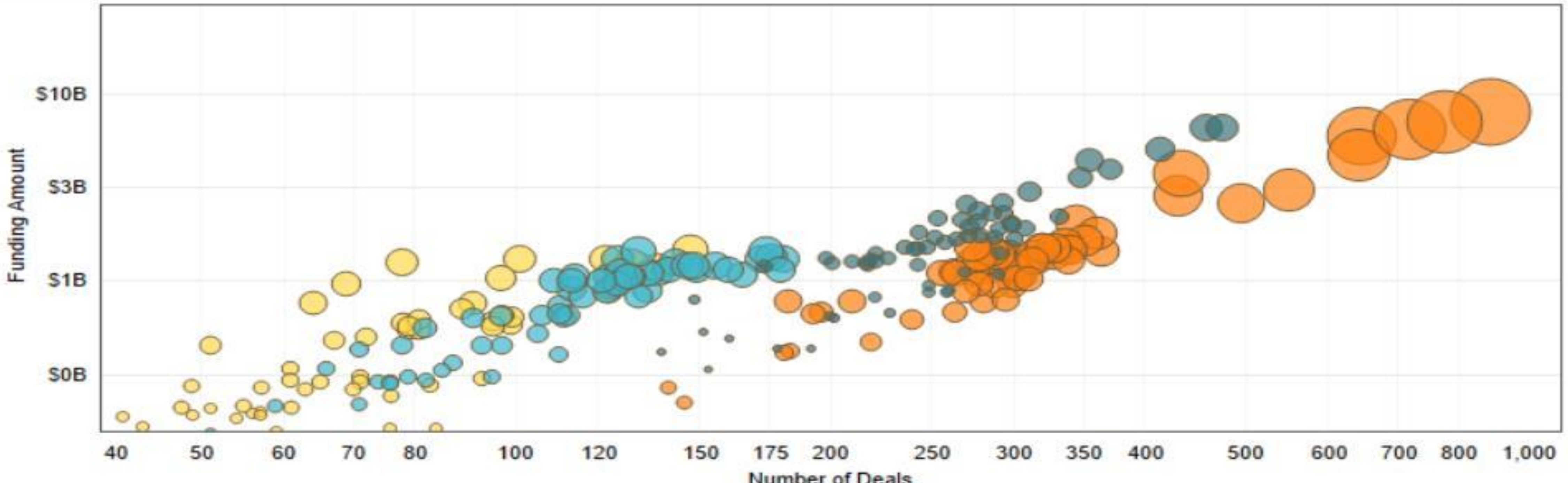


Select Date Range:
1995 to 2010

Select Quarter
☑ Q1
☑ Q2
☑ Q3
☑ Q4

Select Sector:
☑ Biotech
☑ Hardware
☑ Industrial
☑ Software

Funding Details

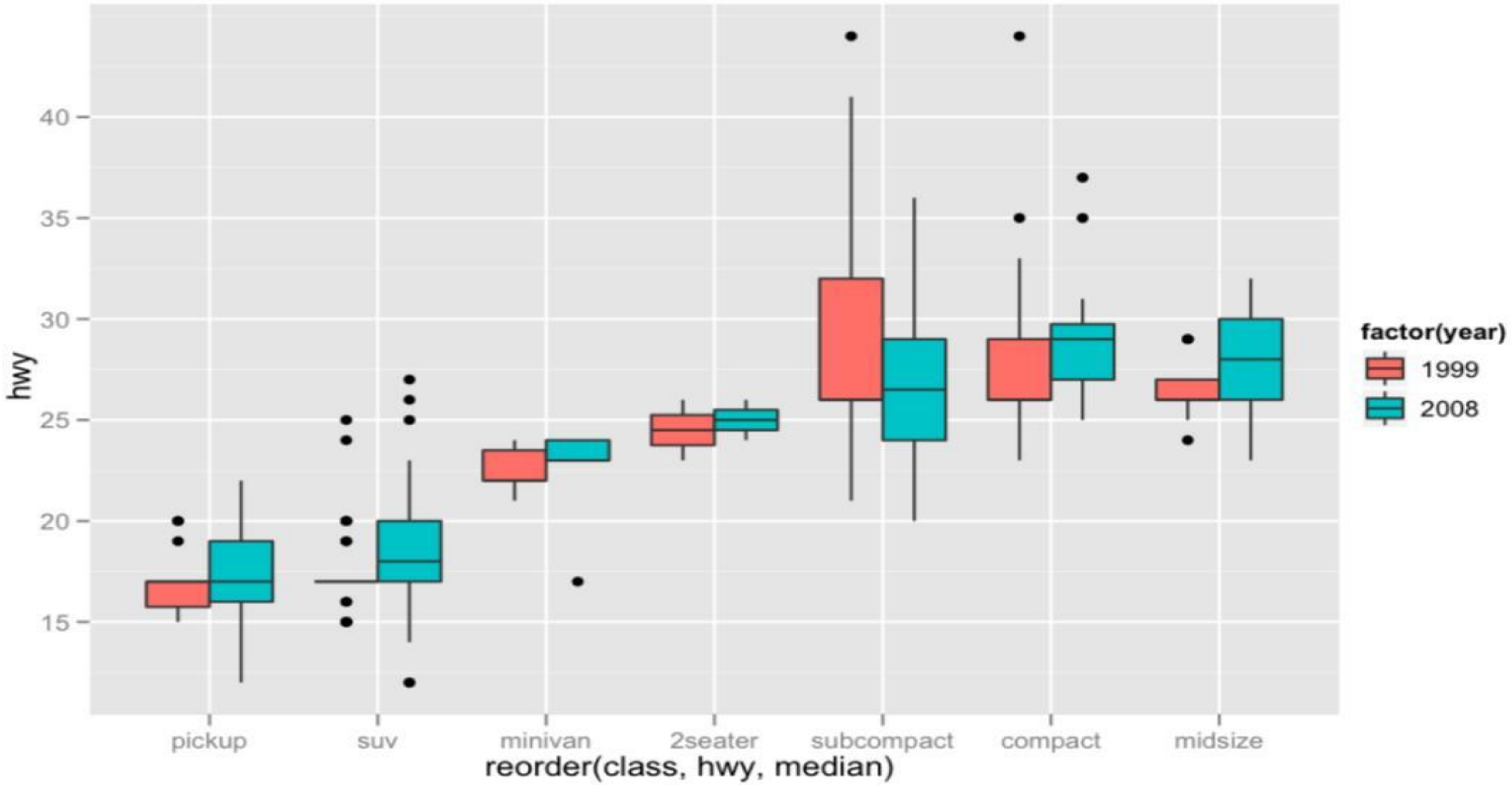


Select Funding Type:
☑ First sequence
☑ Total

Avg. Funding Amount
● $0B
● $1B
● $2B
● $3B
● $4B

Box Plot (using R ggplot 2)

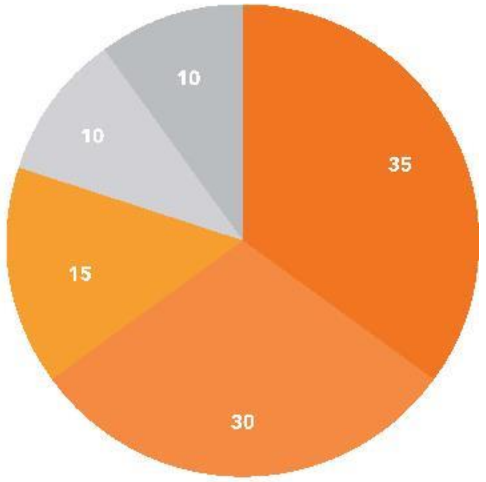
Chart in R ggplot2



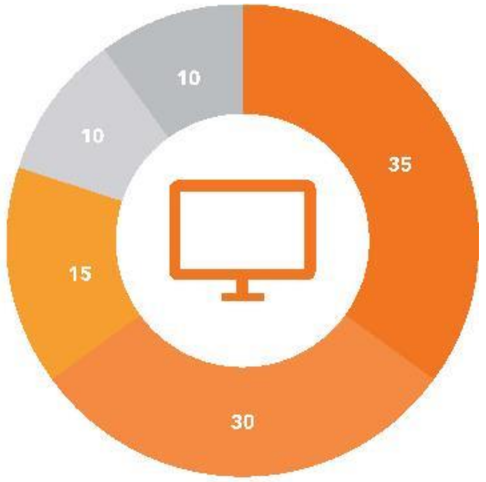
PIE CHART

Pie charts are best used for making part-to-whole comparisons with discrete or continuous data. They are most impactful with a small data set.

VARIATIONS OF PIE CHARTS



STANDARD
Used to show part-to-whole relationships.



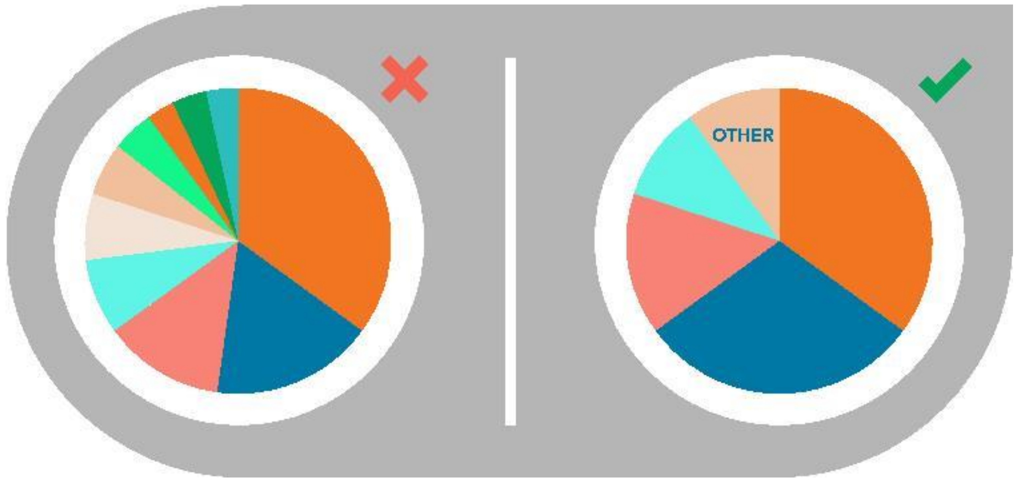
DONUT
Stylistic variation that enables the inclusion of a total value or design element in the center.

THE CASE AGAINST THE PIE CHART

The pie chart is one of the most popular chart types. However, some critics, such as data visualization expert Stephen Few, are not fans. They argue that we are really only able to gauge the size of pie slices if they are in familiar percentages (25%, 50%, 75%, 100%) and positions, because they are common angles. We interpret other angles inconsistently, making it difficult to compare relative sizes and therefore less effective.

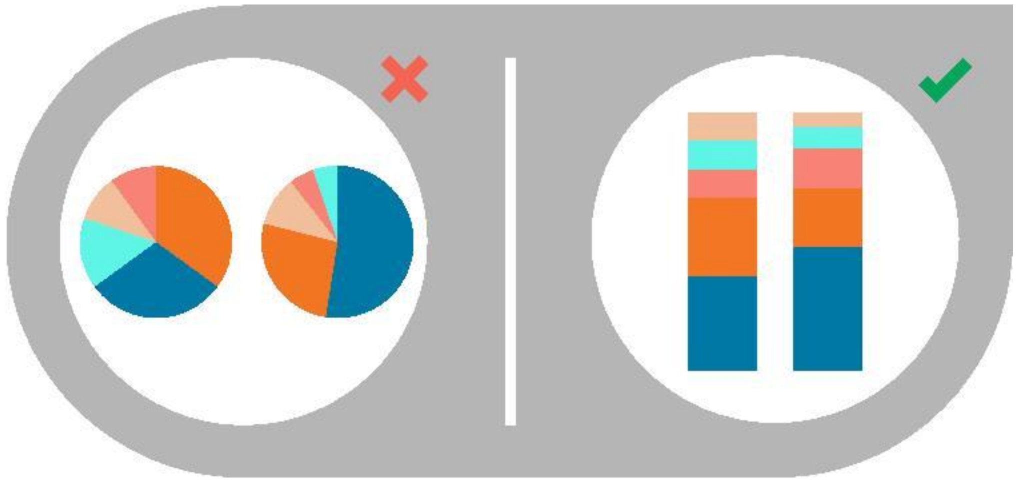
PIE CHART

DESIGN BEST PRACTICES



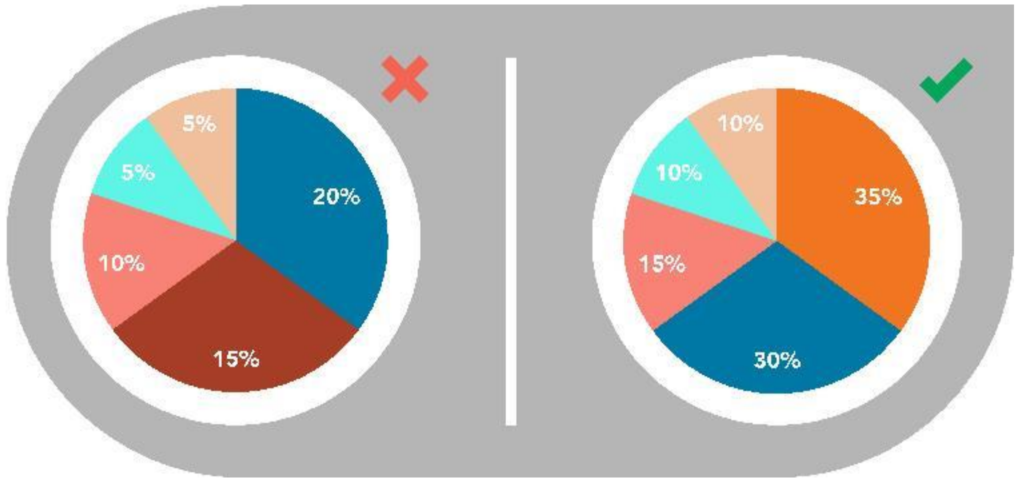
VISUALIZE NO MORE THAN 5 CATEGORIES PER CHART

It is difficult to differentiate between small values; depicting too many slices decreases the impact of the visualization. If needed, you can group smaller values into an “other” or “miscellaneous” category, but make sure it does not hide interesting or significant information.



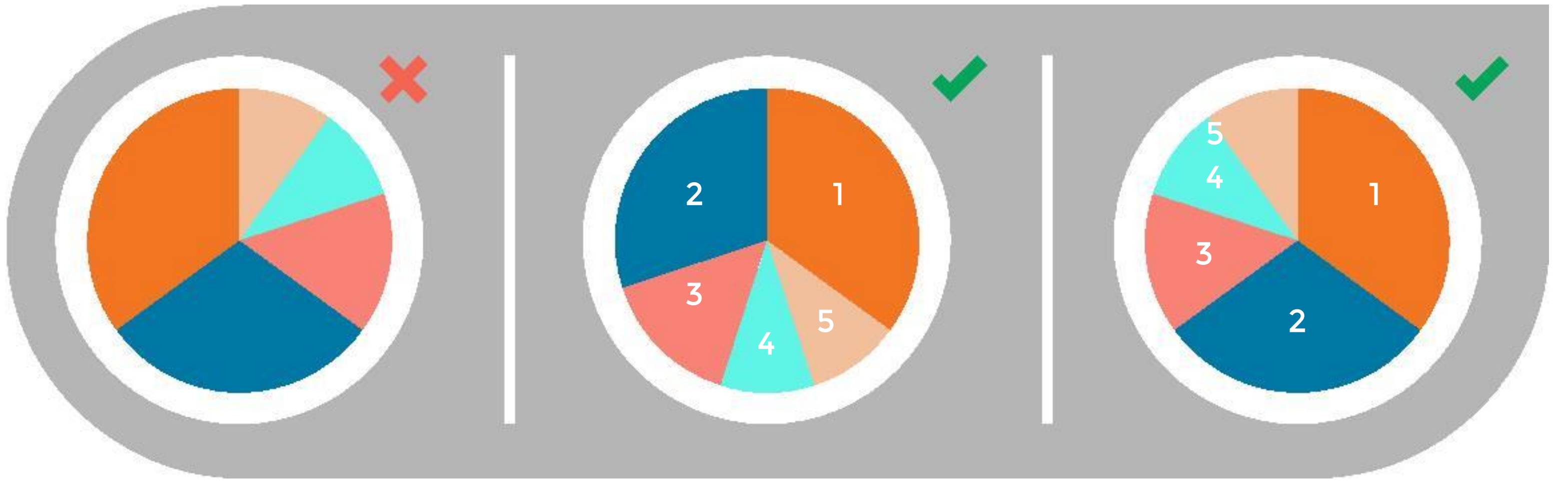
DON’T USE MULTIPLE PIE CHARTS FOR COMPARISON

Slice sizes are very difficult to compare side-by-side. Use a stacked bar chart instead.



MAKE SURE ALL DATA ADDS UP TO 100%

Verify that values total 100% and that pie slices are sized proportionate to their corresponding value.



ORDER SLICES CORRECTLY

There are two ways to order sections, both of which are meant to aid comprehension:

OPTION 1

Place the largest section at 12 o’clock, going clockwise. Place the second largest section at 12 o’clock, going counterclockwise. The remaining sections can be placed below, continuing counterclockwise.

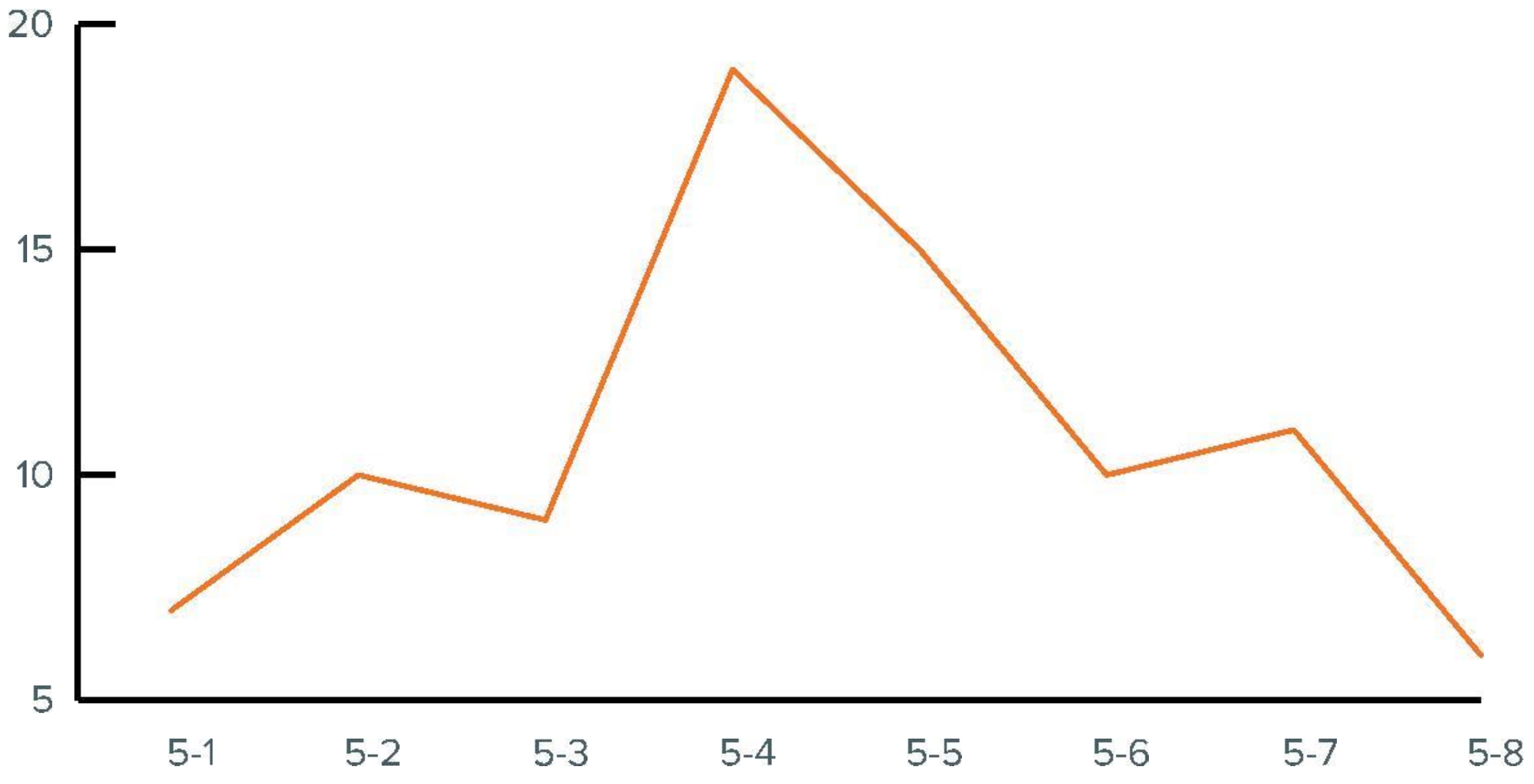
OPTION 2

Start the largest section at 12 o’clock, going clockwise. Place remaining sections in descending order, going clockwise.

LINE CHART

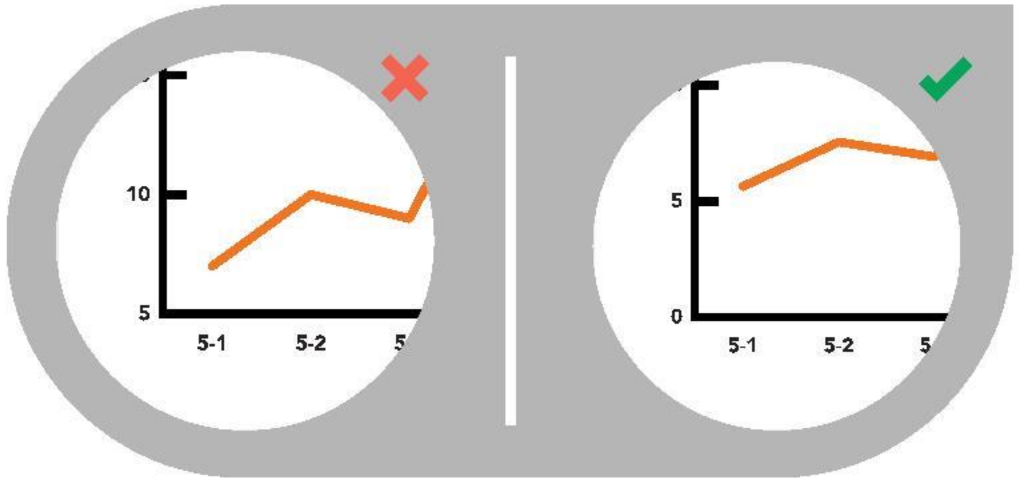
Line charts are used to show time-series relationships with continuous data. They help show trend, acceleration, deceleration, and volatility.

DIRECT MARKETING VIEWS, BY DATE



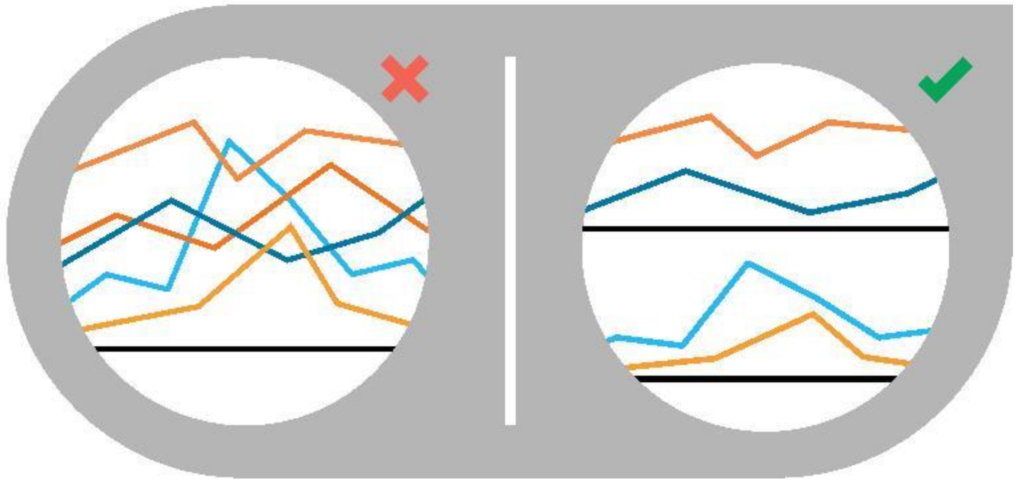
LINE CHART

DESIGN BEST PRACTICES



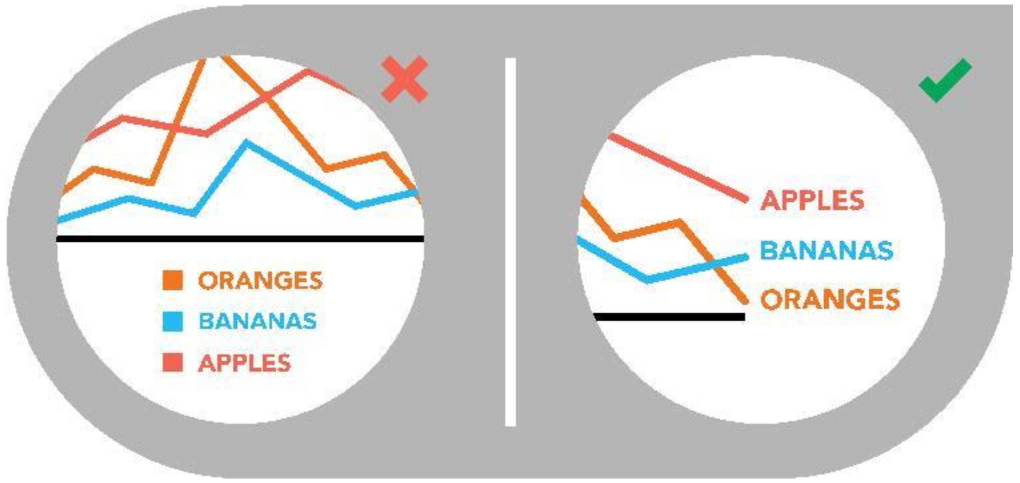
INCLUDE A ZERO BASELINE IF POSSIBLE

Although a line chart does not have to start at a zero baseline, it should be included if possible. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances.



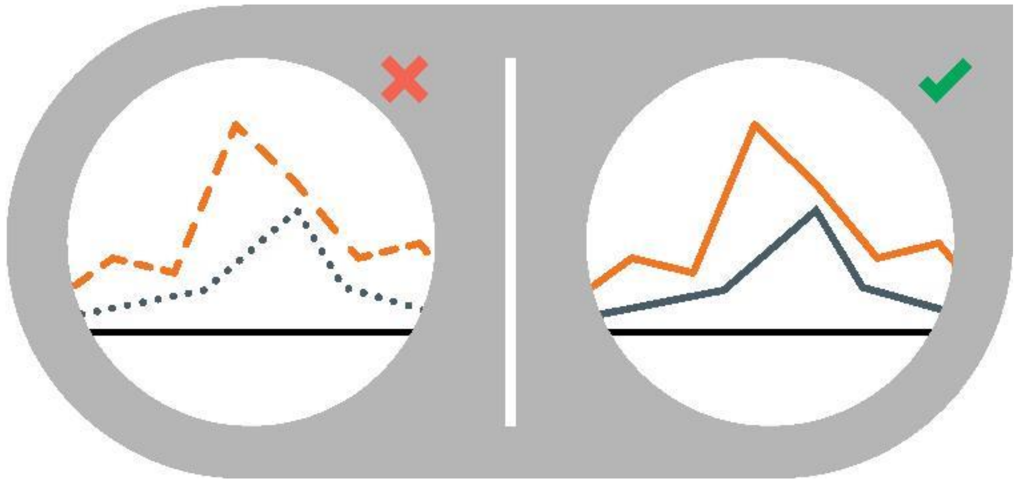
DON'T PLOT MORE THAN 4 LINES

If you need to display more, break them out into separate charts for better comparison.



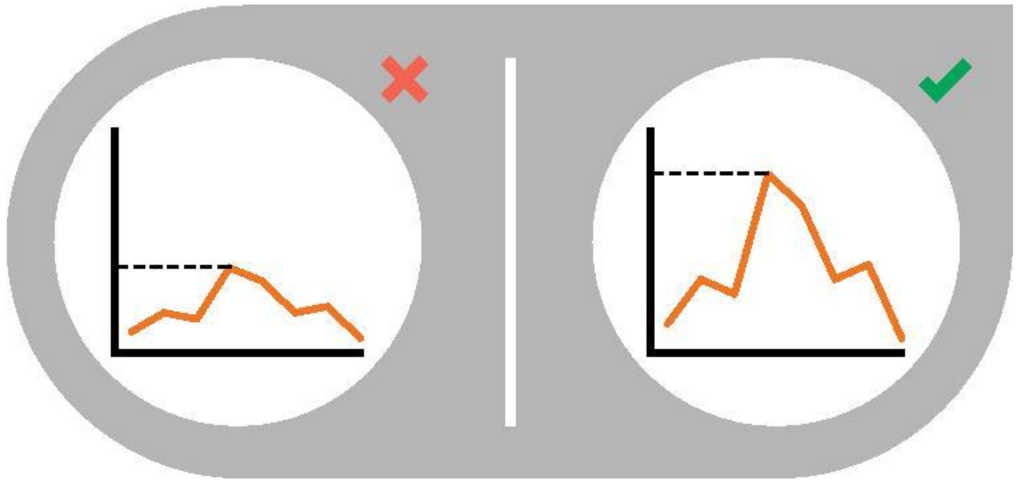
LABEL THE LINES DIRECTLY

This lets readers quickly identify lines and corresponding labels instead of referencing a legend.



USE SOLID LINES ONLY

Dashed and dotted lines can be distracting.



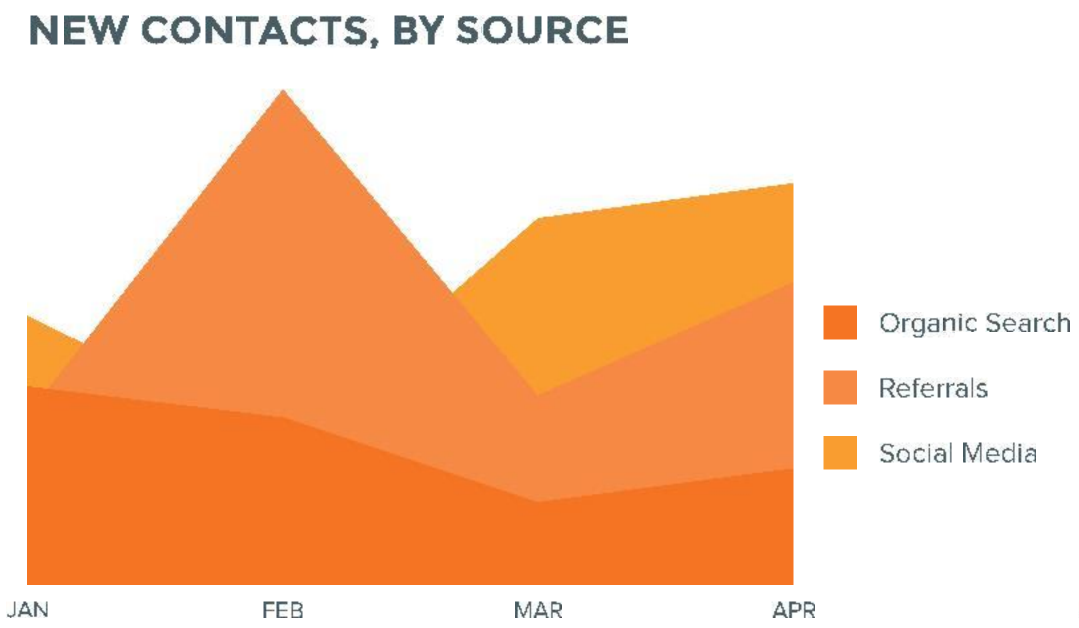
USE THE RIGHT HEIGHT

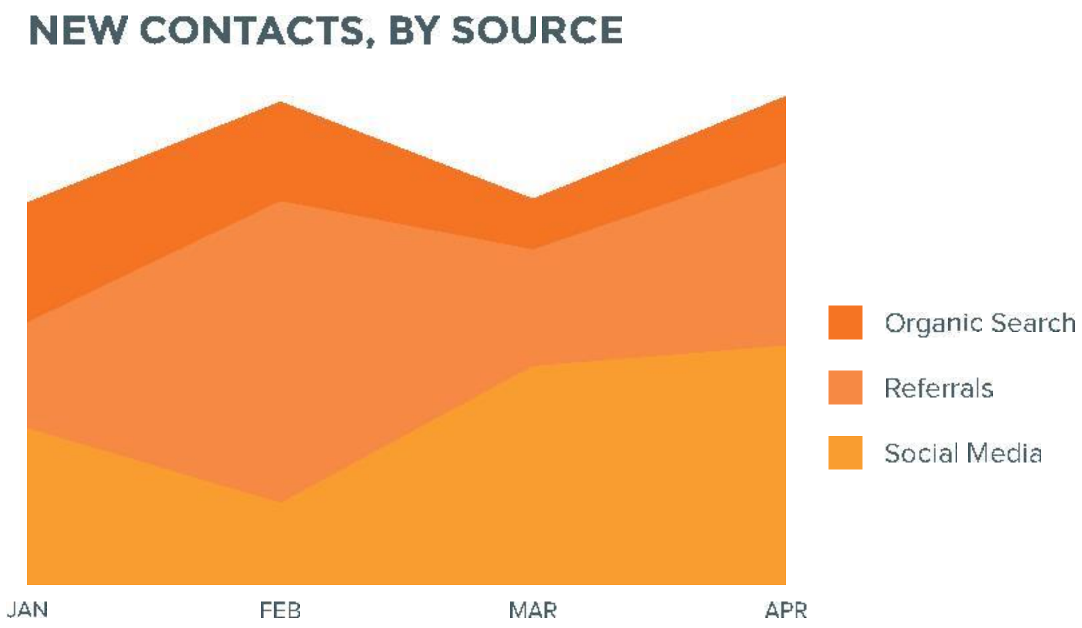
Plot all data points so that the line chart takes up approximately two-thirds of the y-axis' total scale.

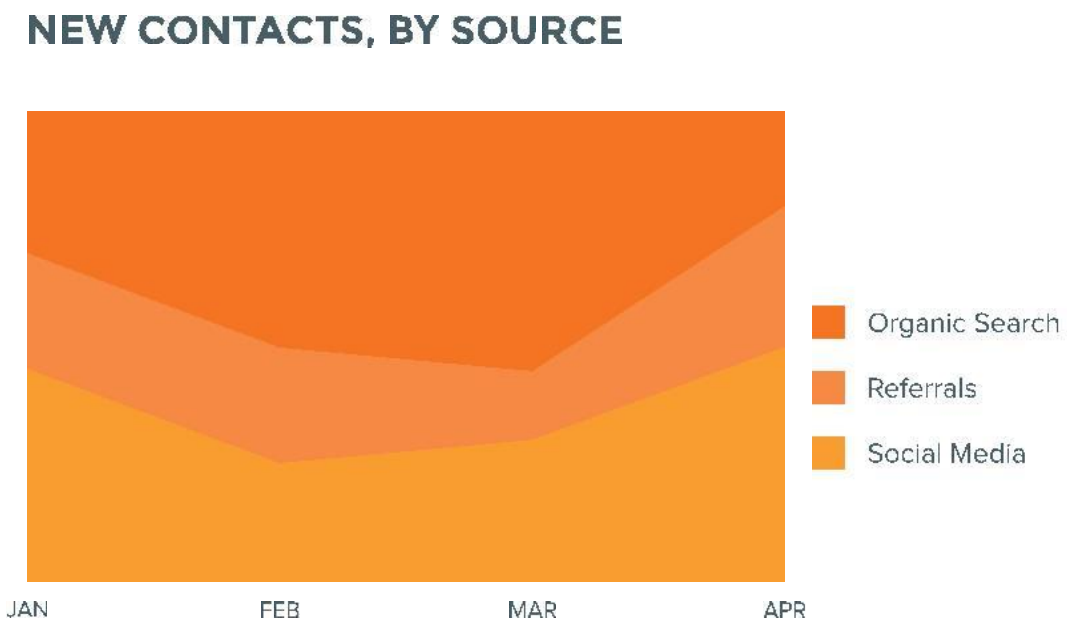
AREA CHART

Area charts depict a time-series relationship, but they are different than line charts in that they can represent volume.

VARIATIONS OF AREA CHARTS







AREA CHART

Best used to show or compare a quantitative progression over time.

STACKED AREA

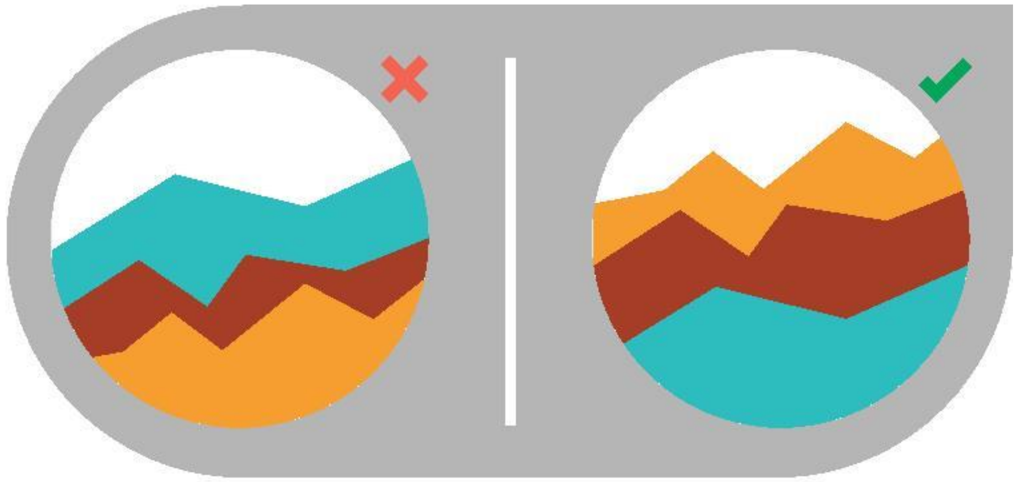
Best used to visualize part-to-whole relationships, helping show how each category contributes to the cumulative total.

100% STACKED AREA

Best used to show distribution of categories as part of a whole, where the cumulative total is unimportant.

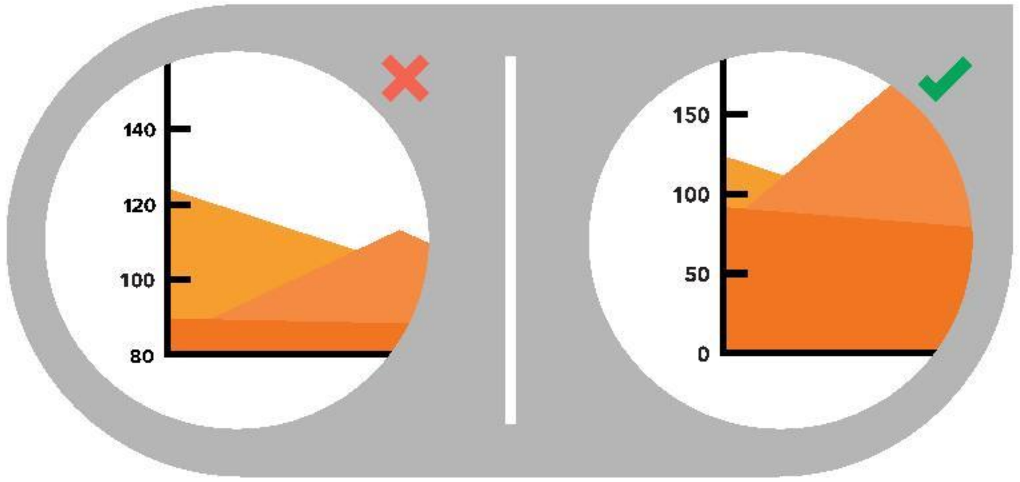
AREA CHART

DESIGN BEST PRACTICES



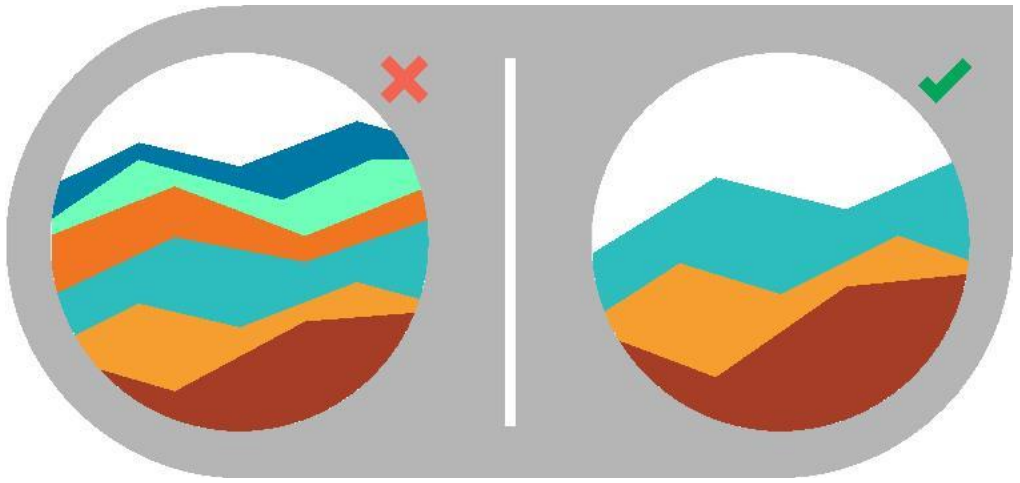
MAKE IT EASY TO READ

In stacked area charts, arrange data to position categories with highly variable data on the top of the chart and low variability on the bottom.



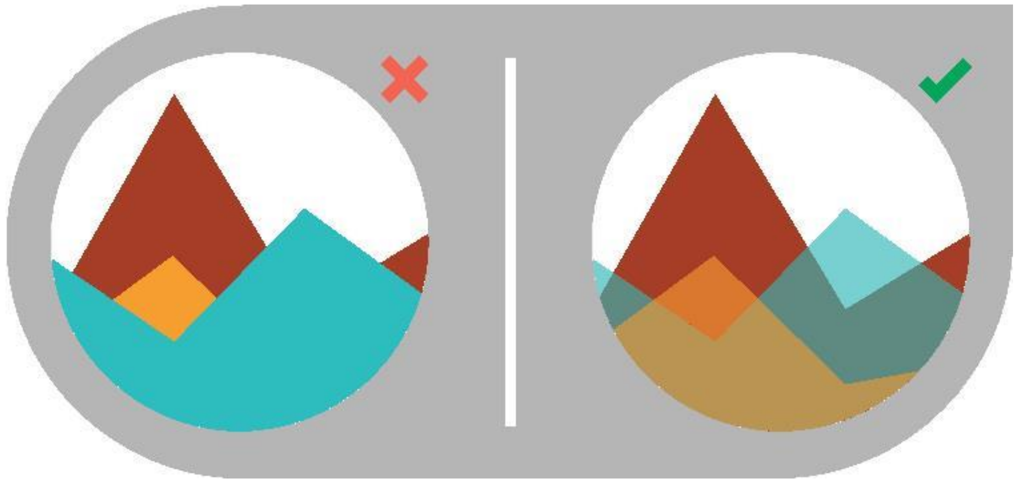
START Y-AXIS VALUE AT 0

Starting the axis above zero truncates the visualization of values.



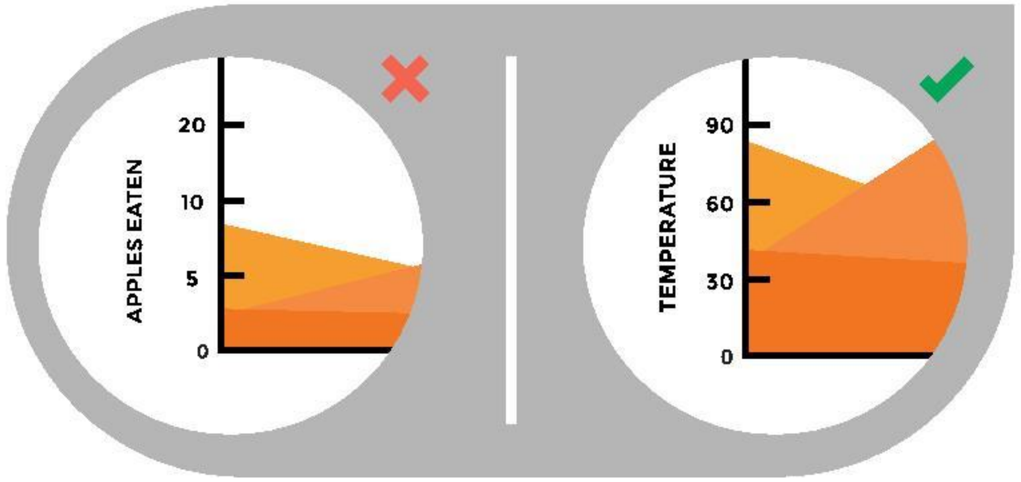
DON'T DISPLAY MORE THAN 4 DATA CATEGORIES

Too many will result in a cluttered visual that is difficult to decipher.



USE TRANSPARENT COLORS

In standard area charts, ensure data isn't obscured in the background by ordering thoughtfully and using transparency.



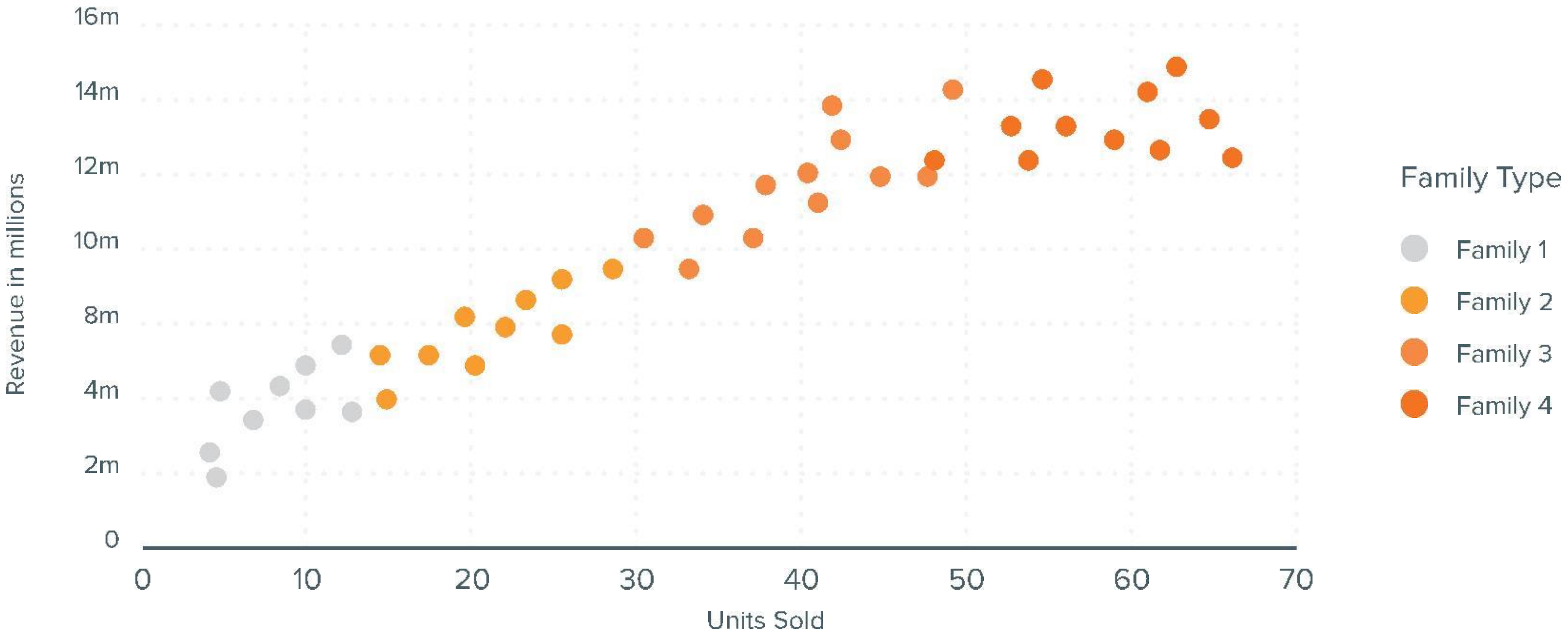
DON'T USE AREA CHARTS TO DISPLAY DISCRETE DATA

The connected lines imply intermediate values, which only exist with continuous data.

SCATTER PLOT

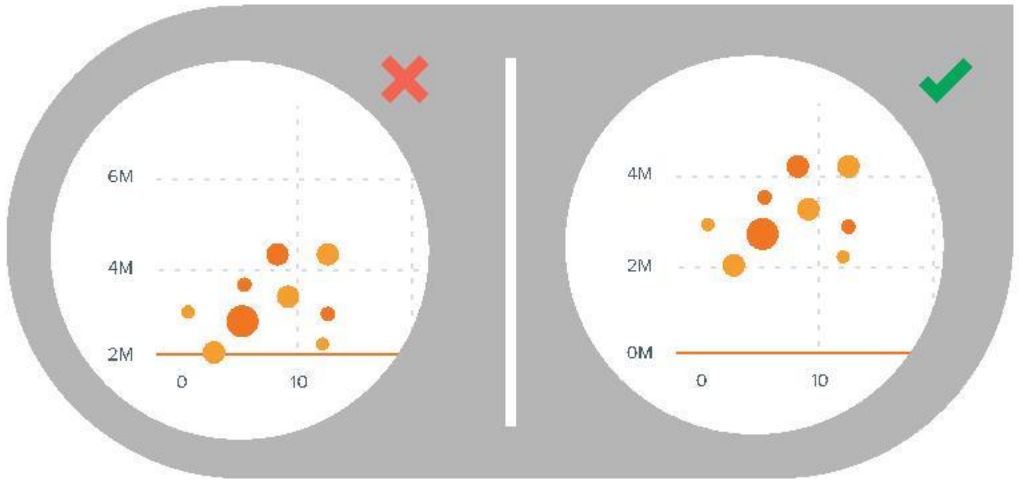
Scatter plots show the relationship between items based on two sets of variables. They are best used to show correlation in a large amount of data.

REVENUE, BY PRODUCT FAMILY

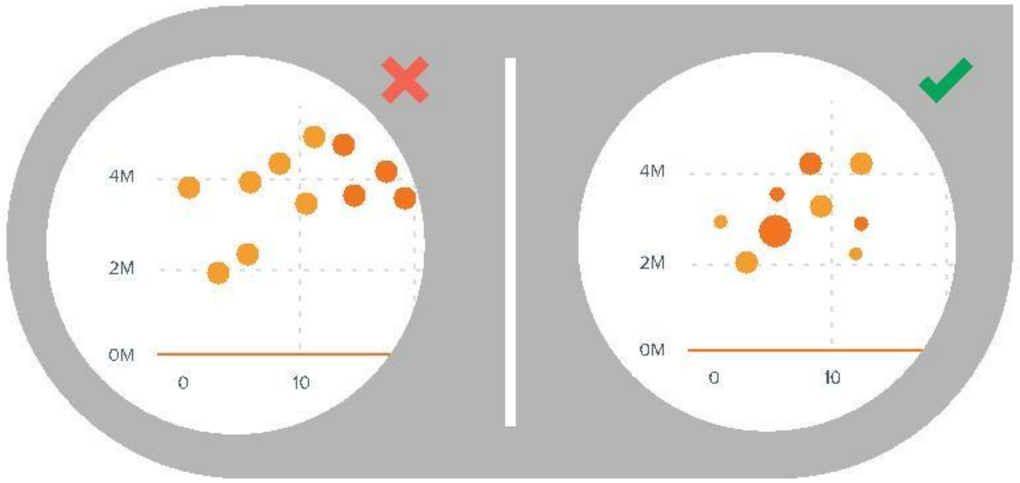


SCATTER PLOT

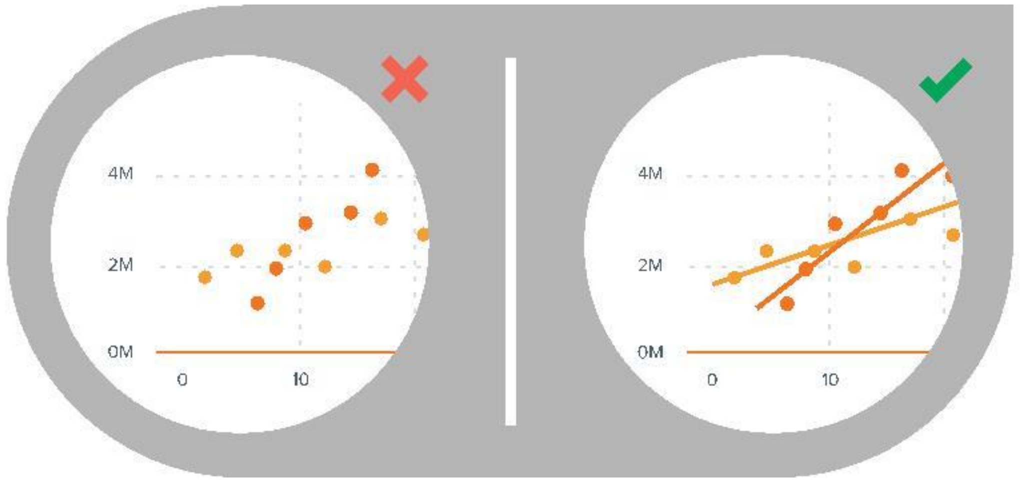
DESIGN BEST PRACTICES



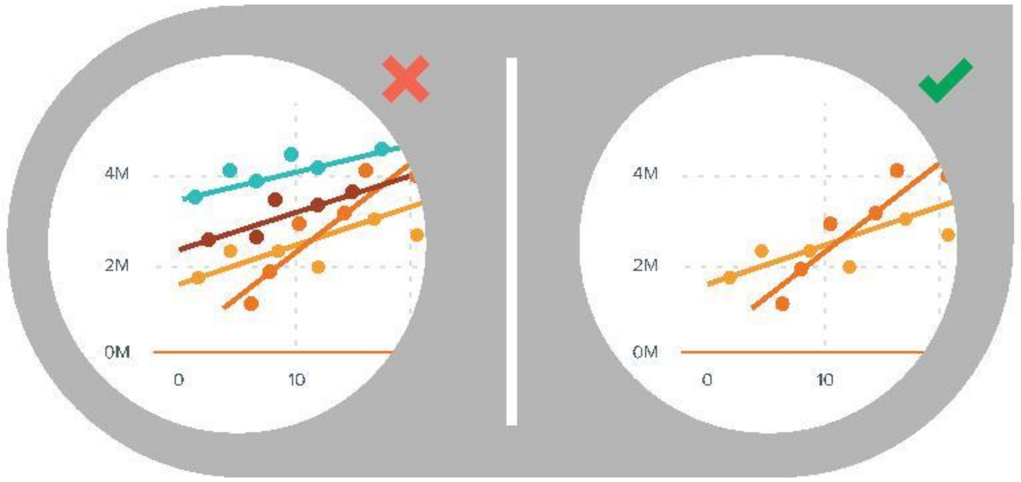
START Y-AXIS VALUE AT 0
Starting the axis above zero truncates the visualization of values.



INCLUDE MORE VARIABLES
Use size and dot color to encode additional data variables.



USE TREND LINES
These help draw correlation between the variables to show trends.



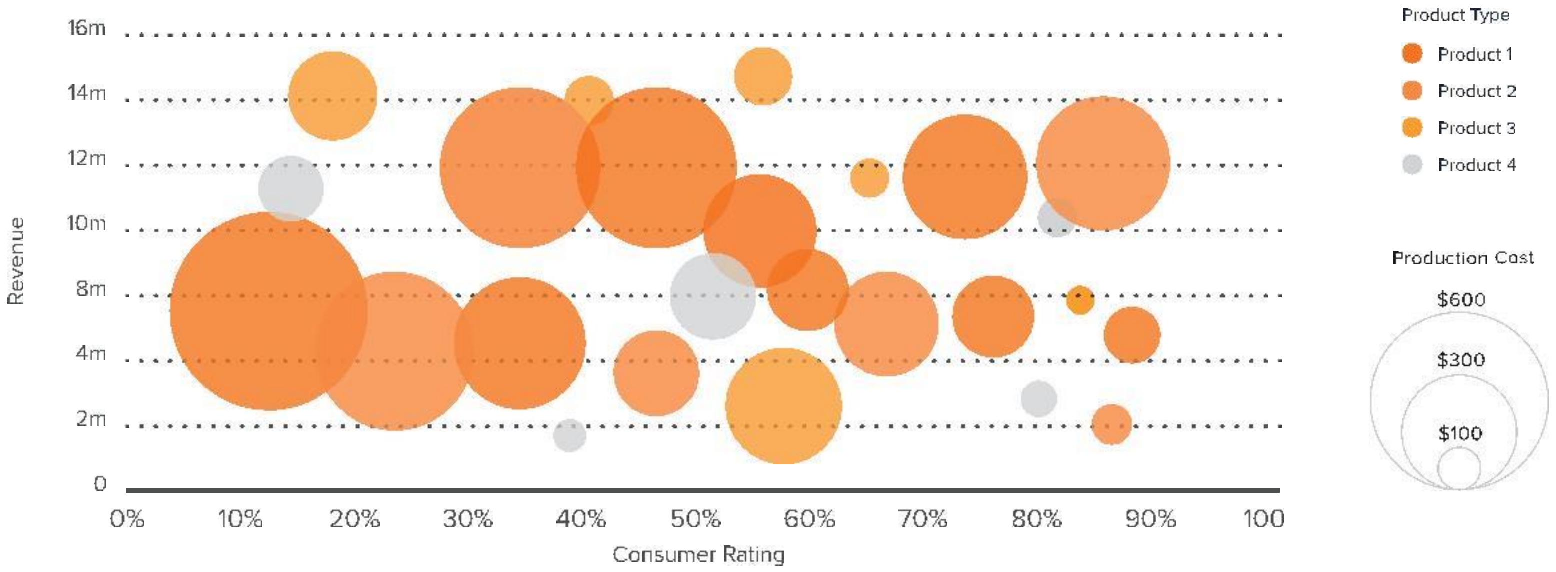
DON'T COMPARE MORE THAN 2 TREND LINES
Too many lines make data difficult to interpret.

BUBBLE CHART

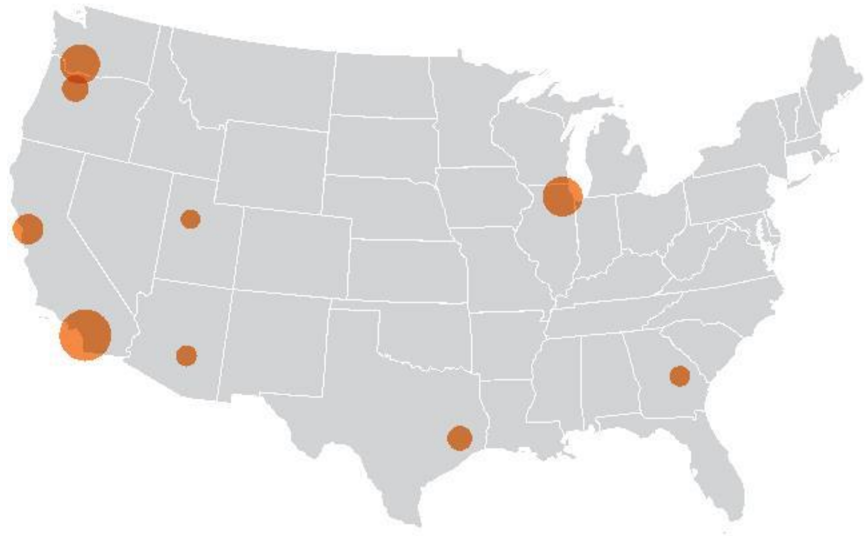
Bubble charts are good for displaying nominal comparisons or ranking relationships.

VARIATIONS OF BUBBLE CHARTS

REVENUE VS. RATING



BIGGEST SALES INCREASE

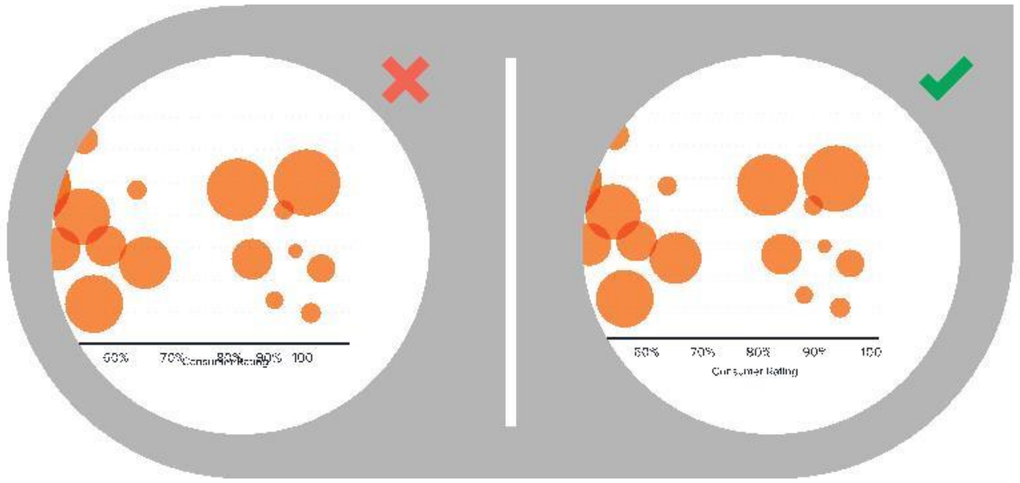


BUBBLE PLOT

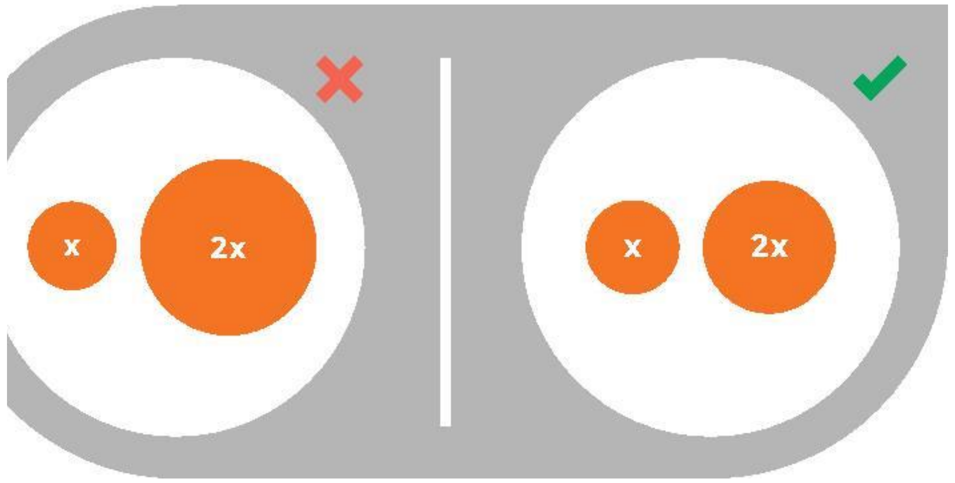
This is a scatter plot with bubbles, best used to display an additional variable.

BUBBLE MAP

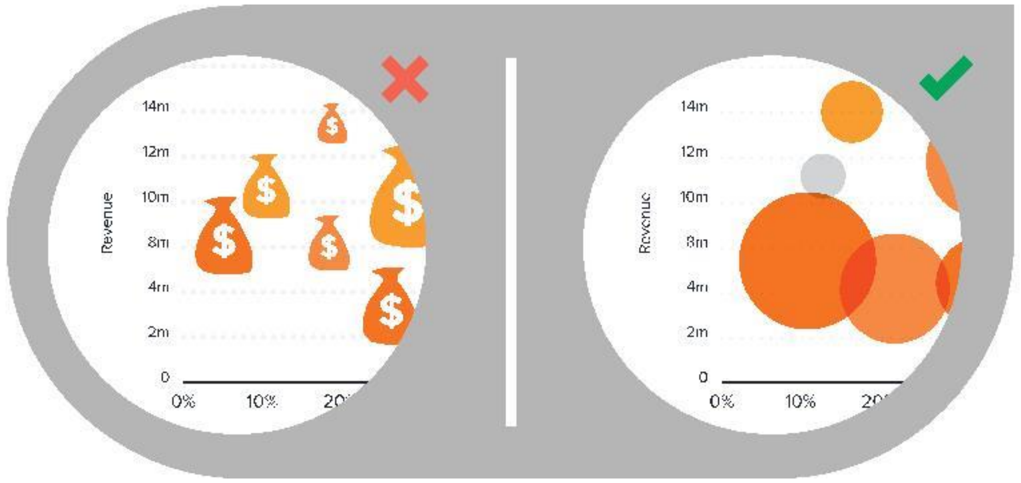
Best used for visualizing values for specific geographic regions.



MAKE SURE LABELS ARE VISIBLE
All labels should be unobstructed and easily identified with the corresponding bubble.



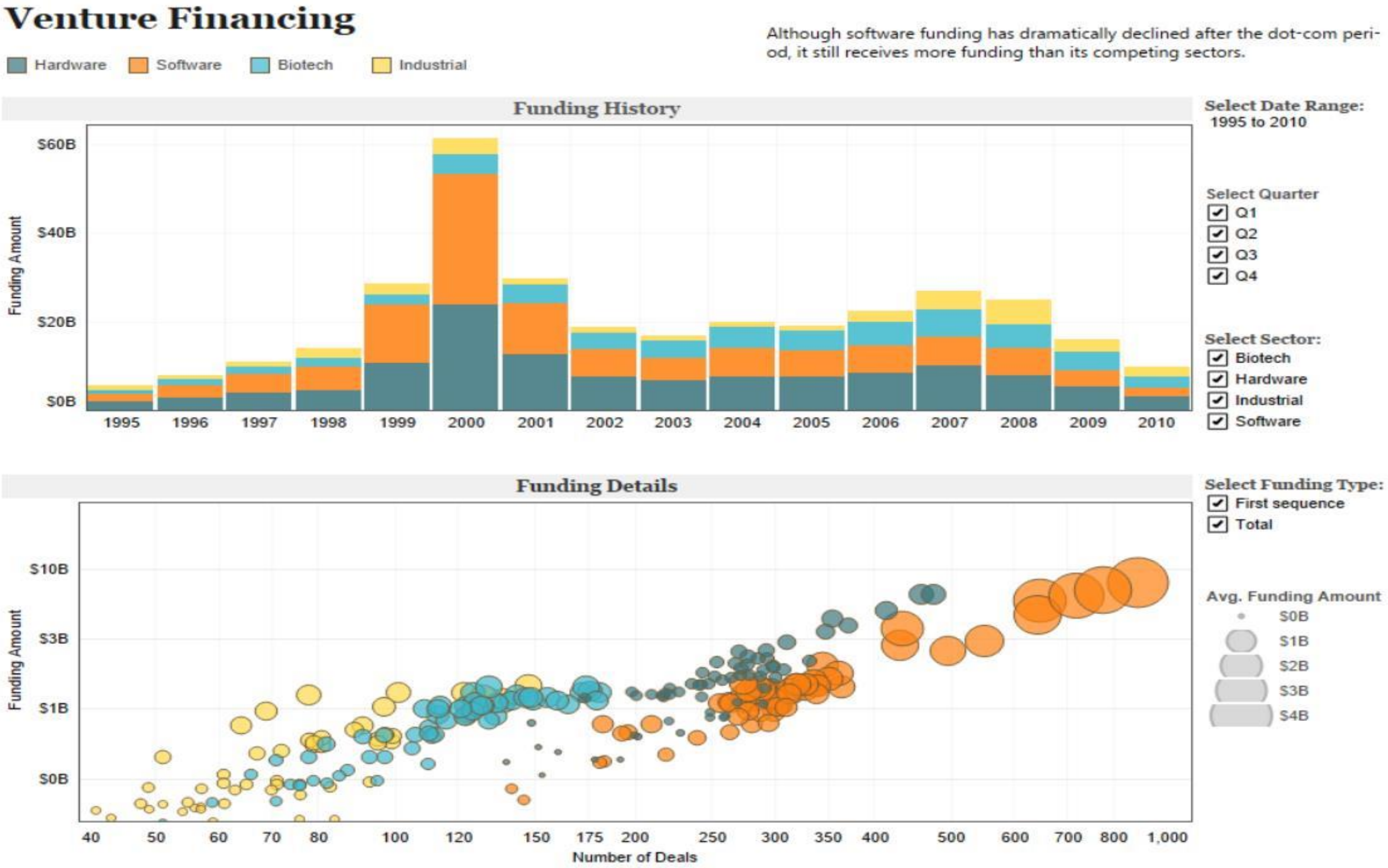
SIZE BUBBLES APPROPRIATELY
Bubbles should be scaled according to area, not diameter.



DON’T USE ODD SHAPES
Avoid adding too much detail or using shapes that are not entirely circular; this can lead to inaccuracies.

BUBBLE CHART

DESIGN BEST PRACTICES



Venture Financing

■ Hardware ■ Software ■ Biotech ■ Industrial

Although software funding has dramatically declined after the dot-com period, it still receives more funding than its competing sectors.

Funding History

Select Date Range:
1995 to 2010

Select Quarter
☑ Q1
☑ Q2
☑ Q3
☑ Q4

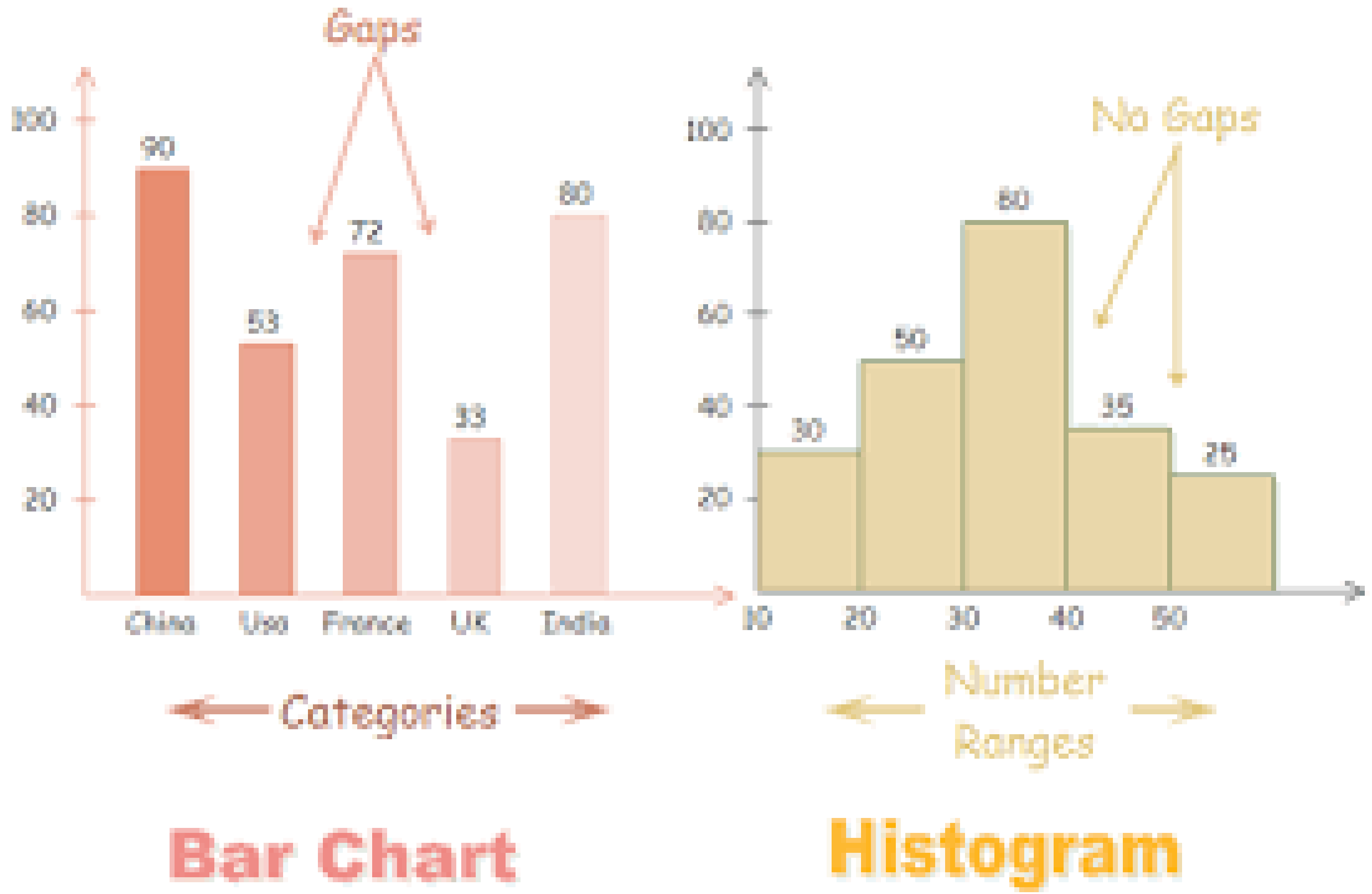
Select Sector:
☑ Biotech
☑ Hardware
☑ Industrial
☑ Software

Funding Details

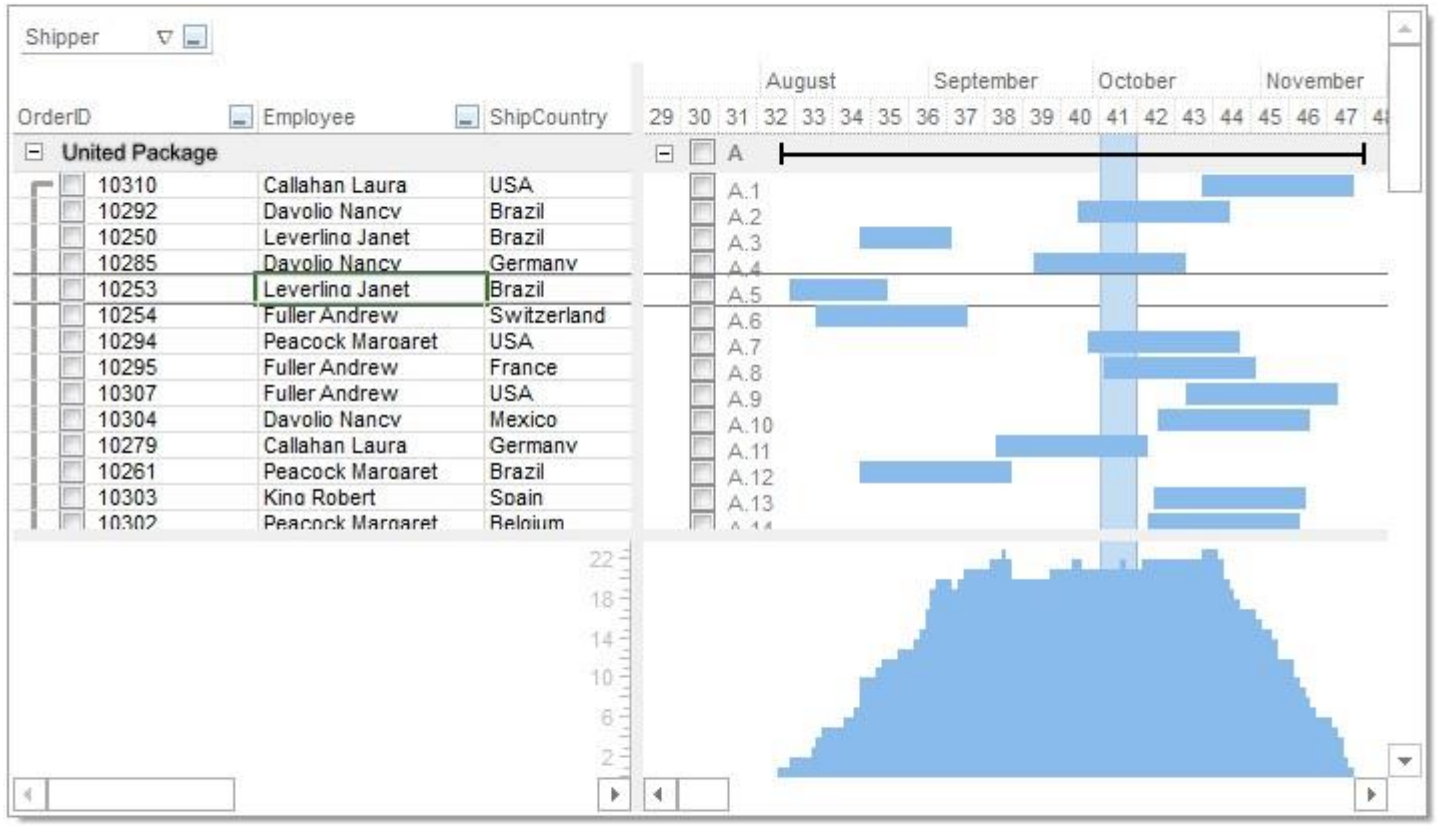
Select Funding Type:
☑ First sequence
☑ Total

Avg. Funding Amount
● $0B
● $1B
● $2B
● $3B
● $4B

Histogram



Gantt Chart

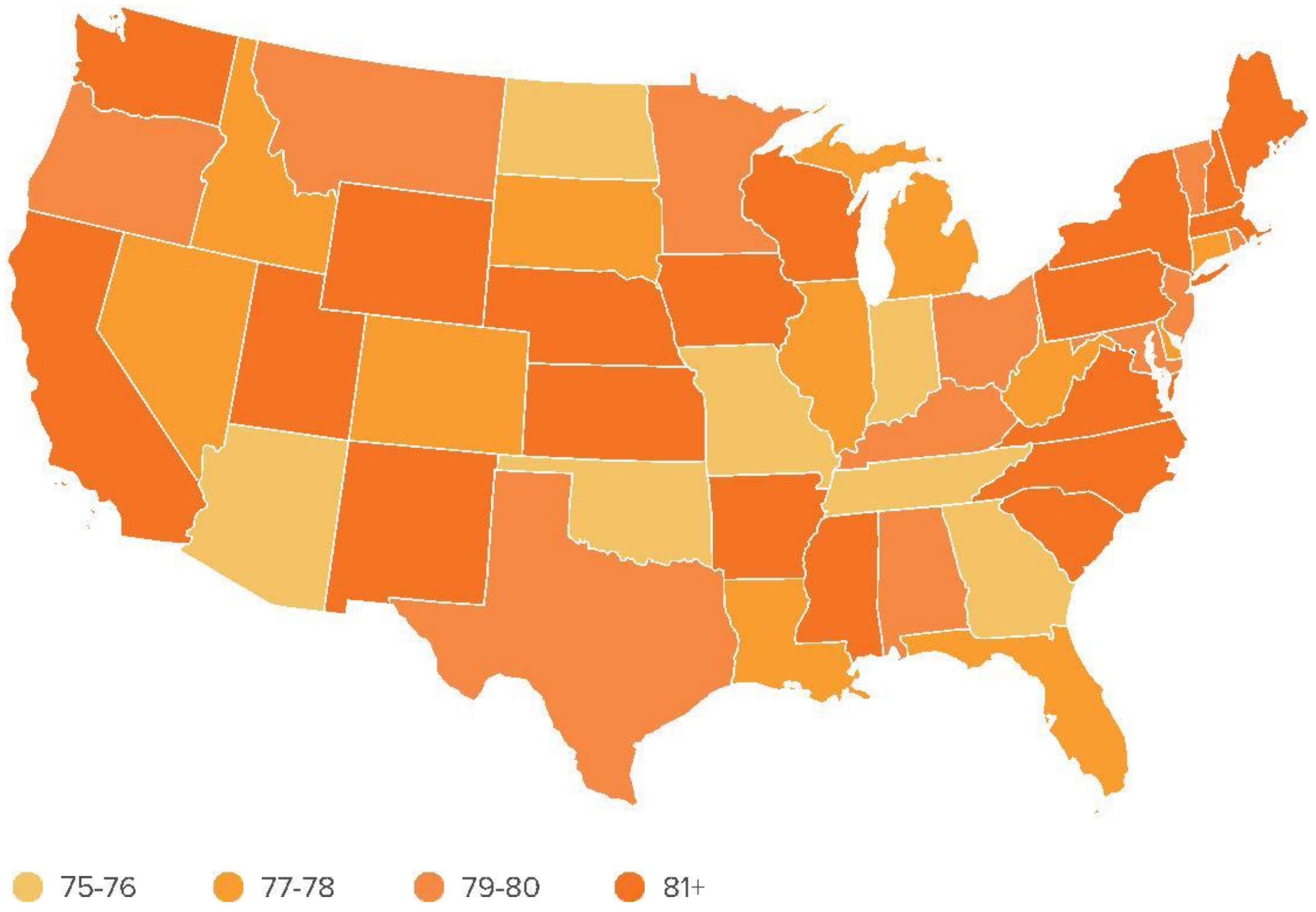


OrderID	Employee	ShipCountry
United Package		
10310	Callahan Laura	USA
10292	Davolio Nancy	Brazil
10250	Leverling Janet	Brazil
10285	Davolio Nancy	Germany
10253	Leverling Janet	Brazil
10254	Fuller Andrew	Switzerland
10294	Peacock Margaret	USA
10295	Fuller Andrew	France
10307	Fuller Andrew	USA
10304	Davolio Nancy	Mexico
10279	Callahan Laura	Germany
10261	Peacock Margaret	Brazil
10303	King Robert	Spain
10302	Peacock Margaret	Belgium

HEAT MAP

Heat maps display categorical data, using intensity of color to represent values of geographic areas or data tables.

STATES WITH NEW SERVICE CONTRACTS

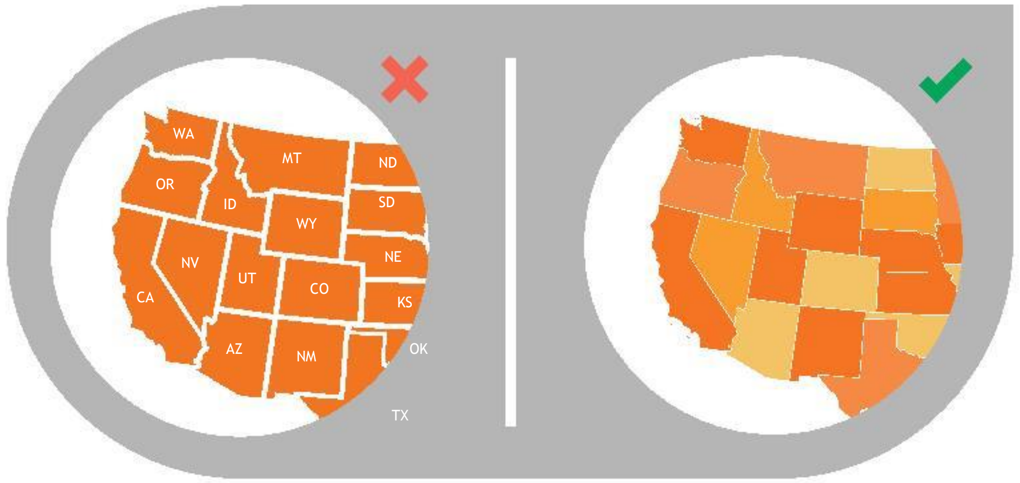


Heat map is a type of visualization tool that is very apt to compare different categories. It helps to visualize measures against dimensions with the help of colors and size to compare one or more dimensions & up to two measures. The layout is similar to a text table with variations in values encoded as colors. In heat map, you can quickly see a wide array of information.

In a heat map, one measure can be assigned to the color and another measure can be assigned to the size.

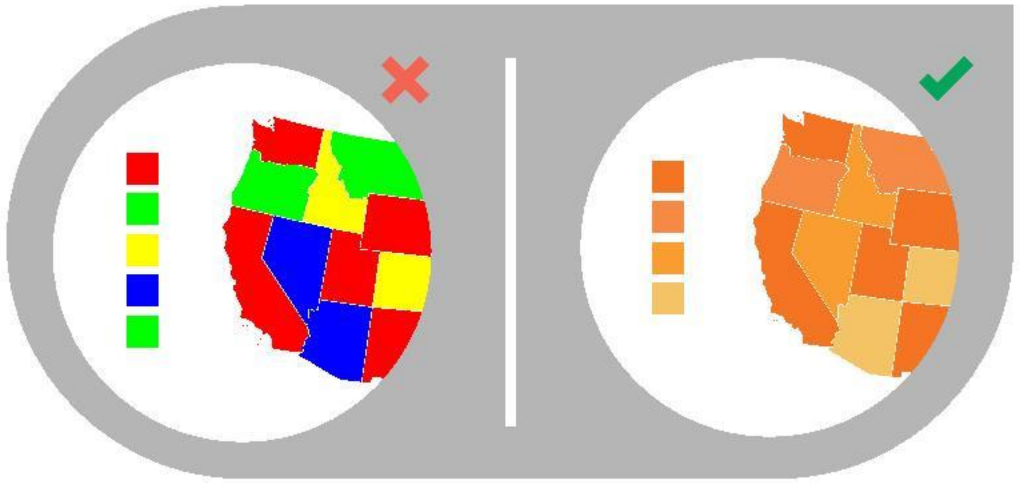
HEAT MAP

DESIGN BEST PRACTICES



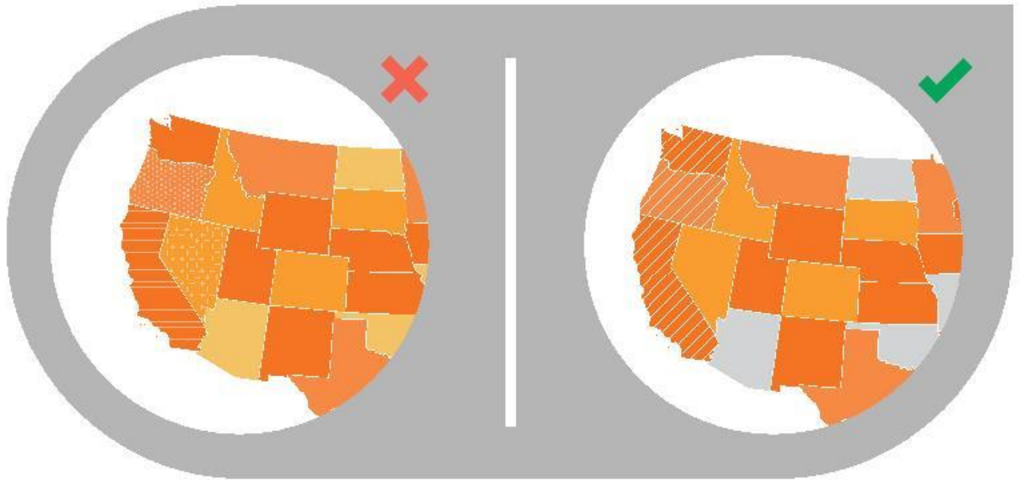
USE A SIMPLE MAP OUTLINE

These lines are meant to frame the data, not distract.



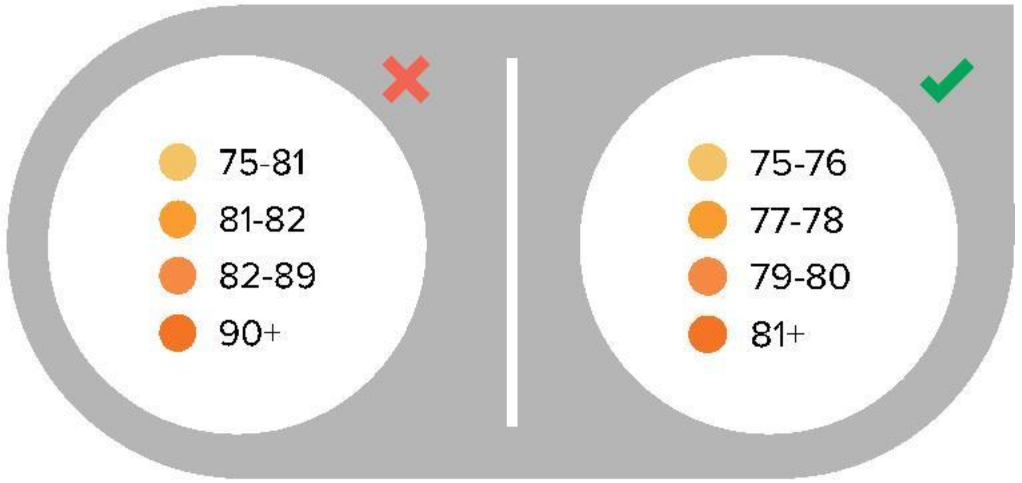
SELECT COLORS APPROPRIATELY

Some colors stand out more than others, giving unnecessary weight to that data. Instead, use a single color with varying shade or a spectrum between two analogous colors to show intensity. Also remember to intuitively code color intensity according to values.



USE PATTERNS SPARINGLY

A pattern overlay that indicates a second variable is acceptable, but using multiple is overwhelming and distracting.



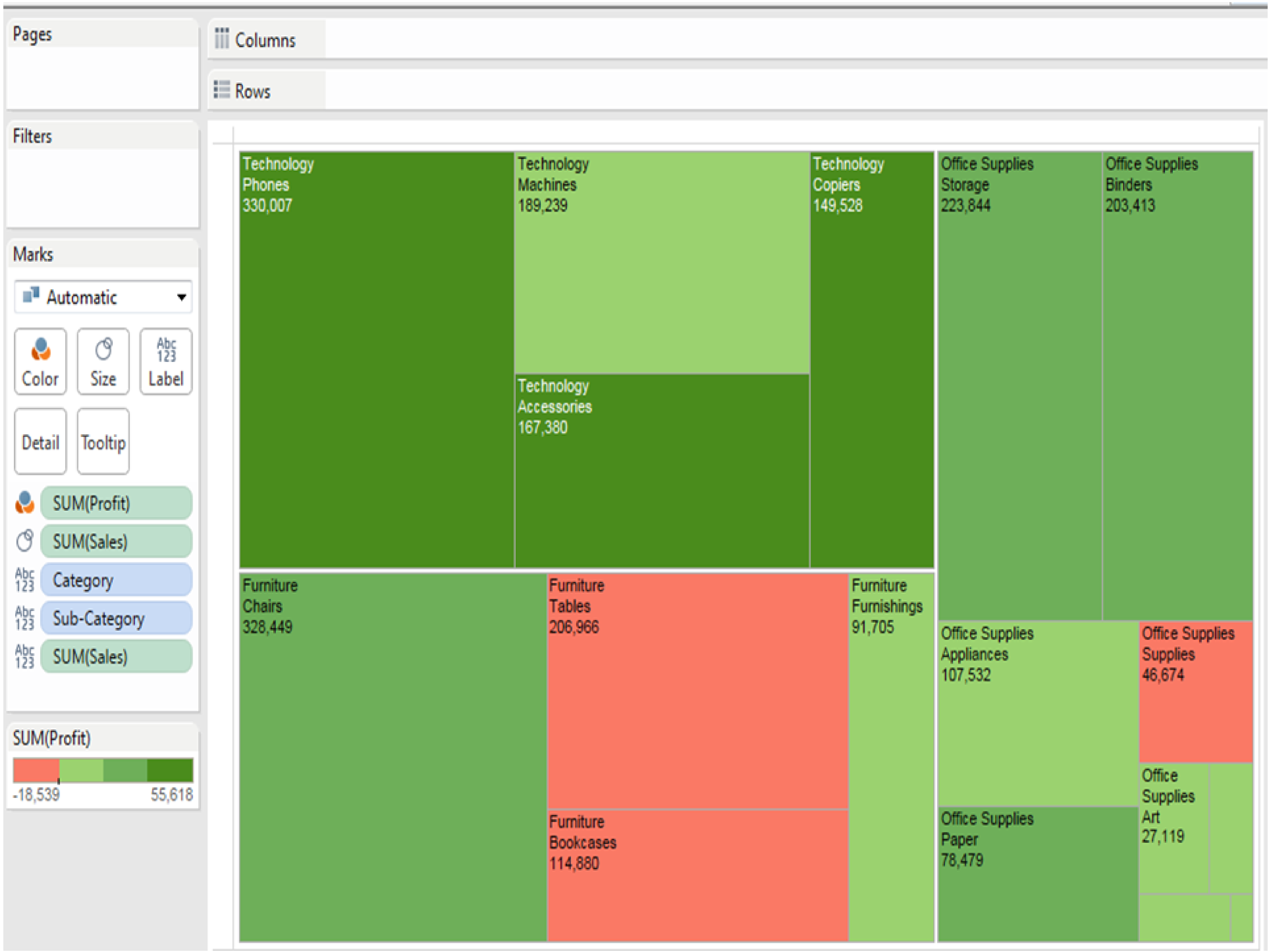
CHOOSE APPROPRIATE DATA RANGES

Select 3-5 numerical ranges that enable fairly even distribution of data between them. Use +/- signs to extend high and low ranges.

What are tree maps?

The ‘tree map’ is a chart type that displays hierarchical or part-to-whole relationships via rectangles.
In case of hierarchical (tree-structured) data these rectangles are nested. The space in the view is divided into rectangles that are sized and ordered by a measure. Nested rectangles mean that hierarchy levels in the data are expressed by larger rectangles (above in the hierarchy) containing smaller ones (below in the hierarchy). The rectangles in the tree map range in size from the top left corner of the chart to the bottom right corner, with the largest rectangle positioned in the top left corner and the smallest rectangle in the bottom right corner.

In a tree map 1 or more dimensions & up to 2 measures are used to create such a map.



Marks: Automatic

Color: SUM(Profit)

Size: SUM(Sales)

Label: Category

Label: Sub-Category

Label: SUM(Sales)

SUM(Profit): -18,539 to 55,618

Pages
Columns
Rows
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Technology Phones 330,007
Technology Machines 189,239
Technology Accessories 167,380
Technology Copiers 149,528
Office Supplies Storage 223,844
Office Supplies Binders 203,413
Furniture Chairs 328,449
Furniture Tables 206,966
Furniture Bookcases 114,880
Furniture Furnishings 91,705
Office Supplies Appliances 107,532
Office Supplies Supplies 46,674
Office Supplies Paper 78,479
Office Supplies Art 27,119

10 DATA DESIGN DOS AND DON'TS

Designing your data doesn't have to be overwhelming. With a basic understanding of how different data sets should be visualized, along with a few fundamental design tips and best practices, you can create more accurate, more effective data visualizations. Follow these 10 tips to ensure your design does your data justice.

1 | DO USE ONE COLOR TO REPRESENT EACH CATEGORY.

2 | DO ORDER DATA SETS USING LOGICAL HEIRARCHY.

3 | DO USE CALLOUTS TO HIGHLIGHT IMPORTANT OR INTERESTING INFORMATION.

4 | DO VISUALIZE DATA IN A WAY THAT IS EASY FOR READERS TO COMPARE VALUES.

5 | DO USE ICONS TO ENHANCE COMPREHENSION AND REDUCE UNNECESSARY LABELING.

6 | DON'T USE HIGH CONTRAST COLOR COMBINATIONS SUCH AS RED/GREEN OR BLUE/YELLOW.

7 | DON'T USE 3D CHARTS. THEY CAN SKEW PERCEPTION OF THE VISUALIZATION.

8 | DON'T ADD CHART JUNK. UNNECESSARY ILLUSTRATIONS, DROP SHADOWS, OR ORNAMENTATIONS DISTRACT FROM THE DATA.

9 | DON'T USE MORE THAN 6 COLORS IN A SINGLE LAYOUT.

10 | DON'T USE DISTRACTING FONTS OR ELEMENTS (SUCH AS BOLD, ITALIC, OR UNDERLINED TEXT).

References

Visual display of Quantitative Information: Edward Tufte http://goo.gl/qb5ej

Exploratory Data Analysis: John Tukey http://goo.gl/tV57HP

Data Science Life cycle : Maloy Manna

http://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle

Selecting right graph for your message: Stephen Few

www.perceptualedge.com/articles/ie/the_right_graph.pdf

Practical rules for using color in charts: Stephen Few

www.perceptualedge.com/articles/visual.../rules_for_using_color.pdf

OpenIntro Statistics: https://www.openintro.org/stat/

Misleading with statistics: Eric Portelance

https://medium.com/i-data/misleading-with-statistics-c63780efa928

Computational Information Design: Ben Fry

http://benfry.com/phd/dissertation-050312b-acrobat.pdf

Statistical Persuasion: How to Collect, Analyze, and Present Data By Robert W. Pearson

Color Matters By Lloyd Treinish, IBM Research, http://www.research.ibm.com/people/l/lloydt/