BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Types of Visualization

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Types of Visualizations

Topics

- Textual
  - Simple Text
  - Tables
  - Heatmap
- Graphs
  - Points
  - Lines
  - Bars
  - Area
- Visuals be avoided

Text-based Visuals

Text-based Visuals

innovate achieve lead

Simple Text

- Suitable with just a number or two to be shown
- When we have a number or two to communicate, use the number solely and a few supporting words to make the point clear

96%

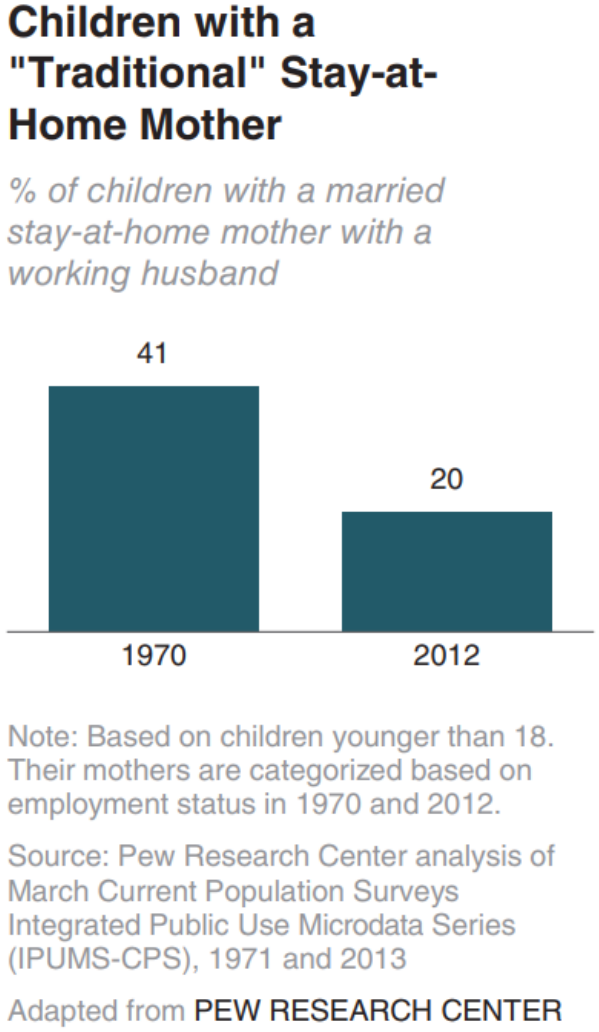
Source: Storytelling with Data by knaflic

Text-based Visuals

innovate achieve lead

Simple Text – Example

- This graph is from April 2014 Pew Research Center report on stay-at-home moms
- Quite a lot of text and space are used for just two numbers
- The graph doesn't help much in interpreting the numbers
- The positioning of the data labels outside of the bars also skews our visual perception of relative height
  - 20 is less than half of 41 doesn't really come across visually



Source: Storytelling with Data by knaflic

Text-based Visuals

innovate achieve lead

Simple Text – Example

- In this case, a simple sentence would suffice:
  - 20% of children had a traditional stay-at-home mom in 2012, compared to 41% in 1970
- Alternatively, a simple visual with a number and simple text could be used
- We might also want to show a different metric:
  - A percentage change
- Any time we reduce from multiple numbers down to a single one
  - Think about what context may be lost in doing so
- For example:
  - Instead of simply saying 50%, if we mention both (20% and 41%), it helps in interpreting and understanding the change

"The number of children having a traditional stay-at-home mom **decreased more than 50% between 1970 and 2012.**"

Source: Storytelling with Data by knaflic

Text-based Visuals

innovate achieve lead

Table

- If we need to communicate multiple different units of measure,
  - it would easier with a table than a graph
- Table structure uses *verbal system*
  - i.e. we read table across rows and down columns
- When you use a table, let data take the center stage
  - everything else should fade into the background
- Use light borders or white space to set apart elements
- Data should stand out, not the borders

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Source: Storytelling with Data by knaflic

Text-based Visuals

Impact of Table Borders

- Note how the data stands out more than the structural components of the table in the 2nd and 3rd cases (light borders, minimal borders)

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Source: Storytelling with Data by knaflic

Text-based Visuals

When to use tables?

- When we need to *compare* or look up *individual values*
- When we require *precise values*
- When the values involve *multiple units of measure*
- When the data has to communicate *quantitative information*, but not trends

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Source: Storytelling with Data by knaflic

Text-based Visuals

innovate achieve lead

Heatmaps

- A heatmap is a way to visualize data in tabular format
- It is used to mix data in a table with visual cues
- Useful to leverage colored cells to convey the *relative magnitude* of the numbers
- Excel's *conditional formatting* uses this feature
- Always include a legend to help the reader interpret the data
- Use *color saturation* to provide visual cues
  - Helps our eyes and brains more quickly target the potential points of interest.

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Source: Storytelling with Data by knaflic

Text-based Visuals

Heatmaps

- In the "Heatmap," higher the saturation of blue, the higher the number
- This makes picking out and processing the details easier compared to original table
  - The lowest number (11%) and highest number (58%)
- Plain table doesn't have any visual cues to help direct our attention

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

- Design Best Practices for Heat Map:
  - Use a basic and clear map outline to avoid distracting from the data
  - Use a single color in varying shades to show changes in data
    - Avoid using multiple patterns

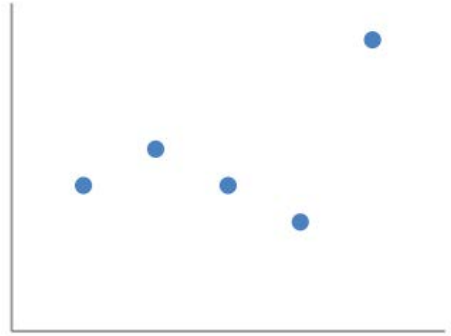
Source: Storytelling with Data by knaflic

Graph-based Visuals

Graph-based Visuals

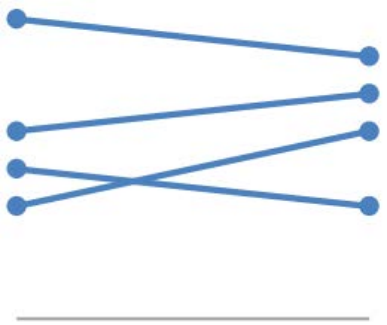
Graphs

- Graphs interact with our visual system
- A well-designed graph will typically get the information across more quickly than a well-designed table
- Graphs broadly fall into four categories:
  - Points, Lines, Bars, and Area

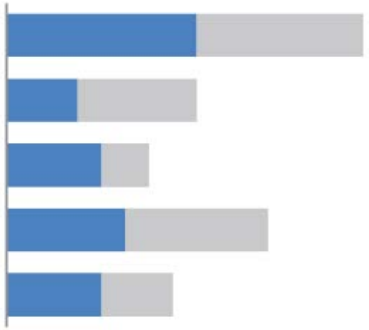


Scatterplot

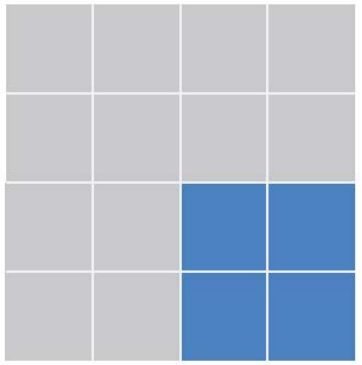
Line



Slopegraph



Stacked horizontal bar



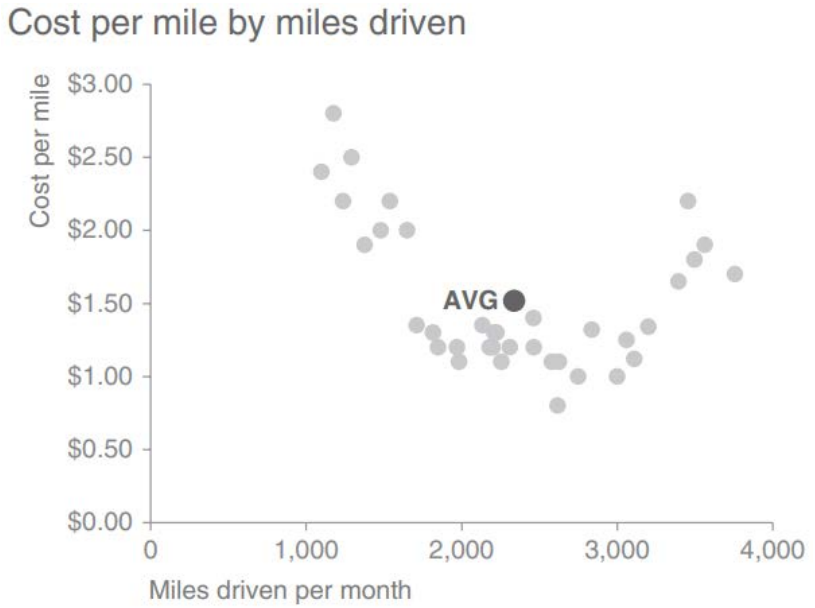
Square area

Source: Storytelling with Data by knaflic

Graph-based Visuals

Points (Scatterplot)

- Useful for showing the *relationship* between two things
- Allows us to encode data simultaneously on a horizontal x-axis and vertical y-axis
- Helps us see if any type of relationship exists between the two
- Typically used in scientific fields
- Figure shows the relationship between miles driven and cost per mile

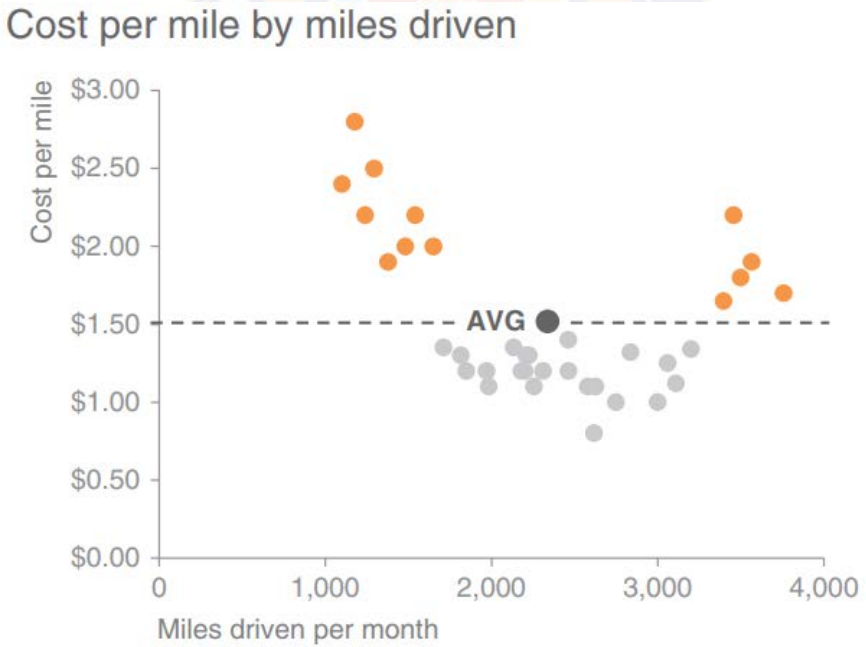


Source: Storytelling with Data by knaflic

Graph-based Visuals

Points (Scatterplot)

- A different color can be used to highlight those cases where cost per mile is above average

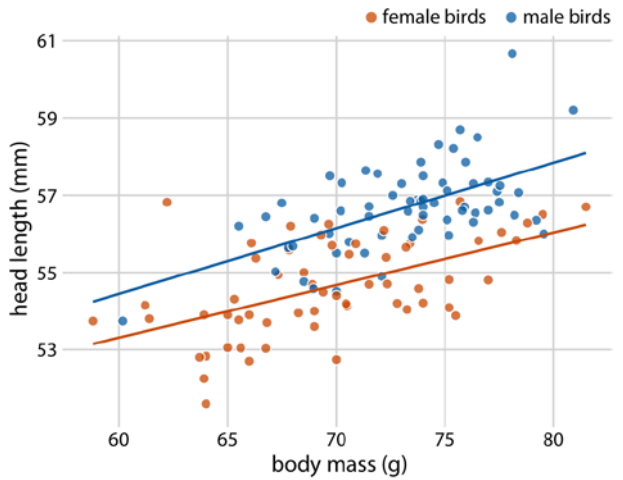


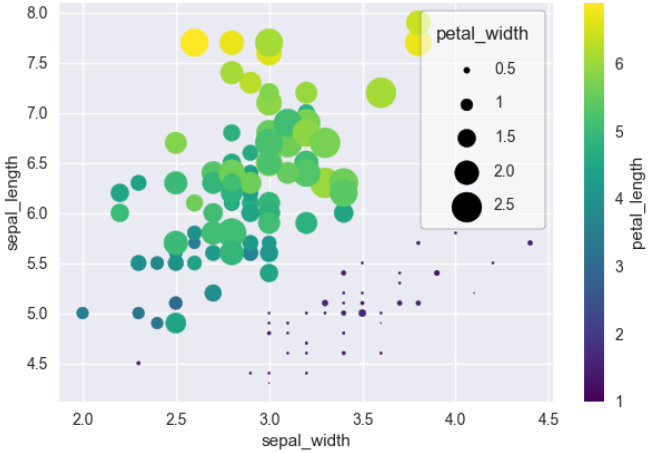
Source: Storytelling with Data by knaflic

Graph-based Visuals

Scatterplots: Design Best Practices

- Include more variables such as size and dot color to encode additional data variables
- Start y-axis at 0 to represent data accurately.
- Use trend lines
- When using trend lines, only use a maximum of two to make the plot easy to understand





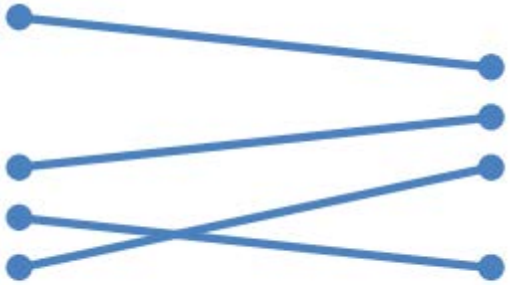
Source: Storytelling with Data by knaflic

Graph-based Visuals

Lines

- Commonly used for plotting *continuous data*
- Points are physically connected with a line
  - Connection between the points may not make sense for categorical data
- Horizontal axis usually *time based*:
  - days, months, quarters, or years
- Two types of line graphs:
  - standard line graph
  - slopegraph

Line



Slopegraph

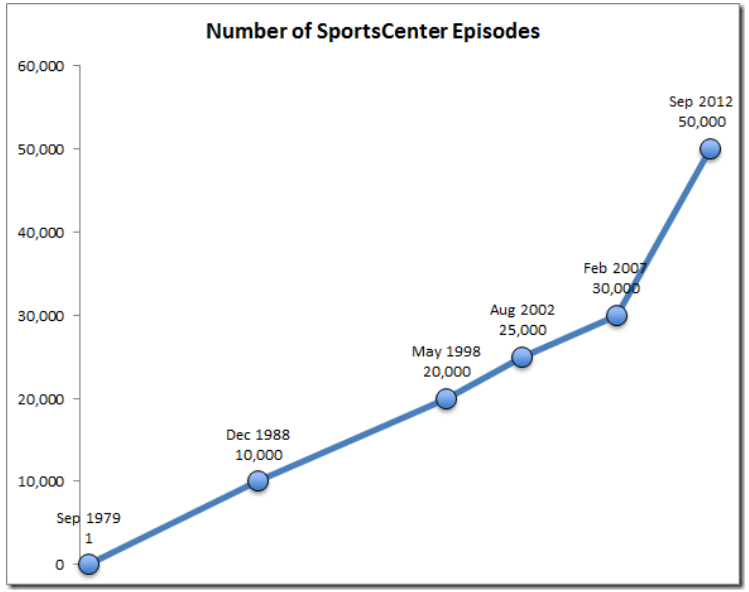
Source: Storytelling with Data by knaflic

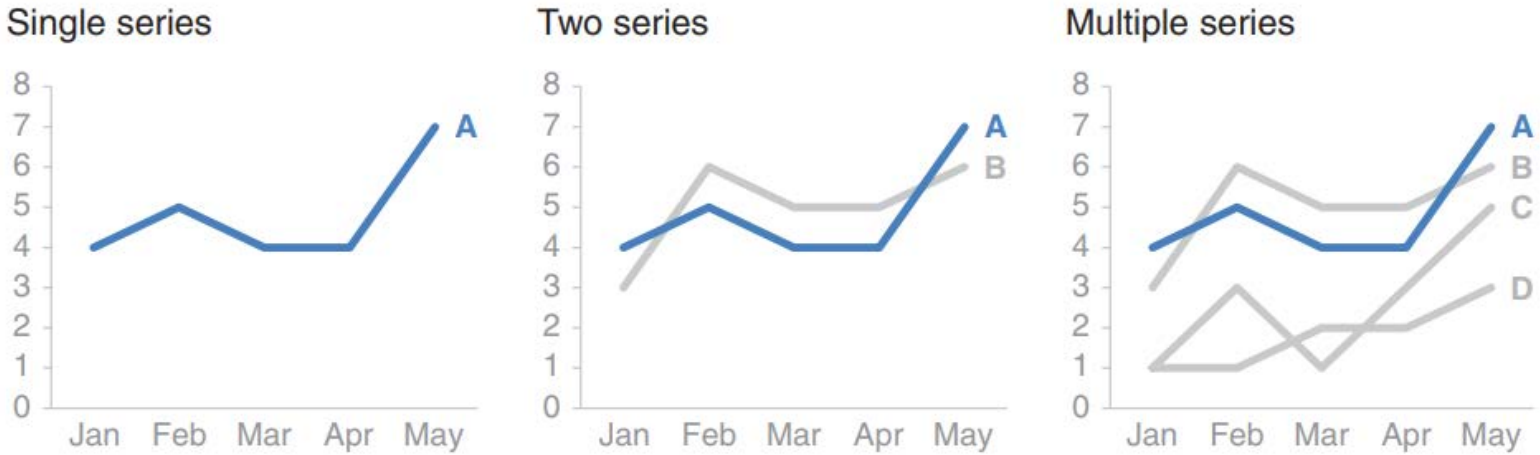
Graph-based Visuals

innovate achieve lead

Line Graph

- Shows a one or multiple series of data
- When using time on horizontal axis, data must be plotted on consistent intervals





Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Line Graphs

- Can be used to show summary statistic such as *average*, or the point estimate of a forecast
- A sense of range such as confidence level can be shown using line graph

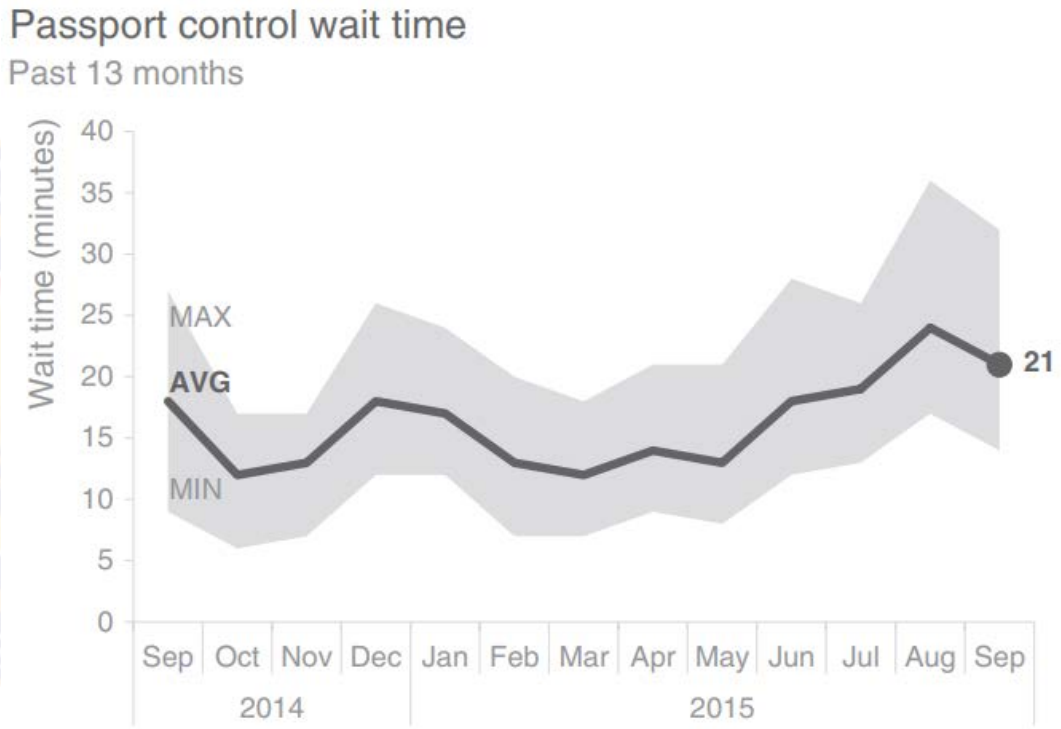


Figure shows minimum, average, and maximum wait times at passport control for an airport over a 13-month period

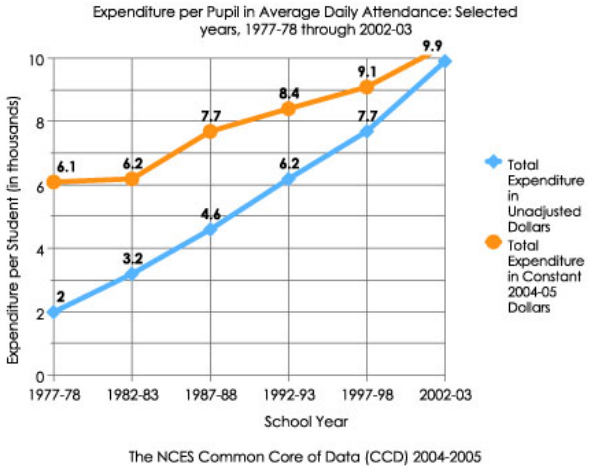
Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Line Graphs: Design Best Practices

- Use only solid lines
- Don't plot more than four lines to avoid visual distractions.
- Use the right height so the lines take up roughly 2/3 of the y-axis' height.



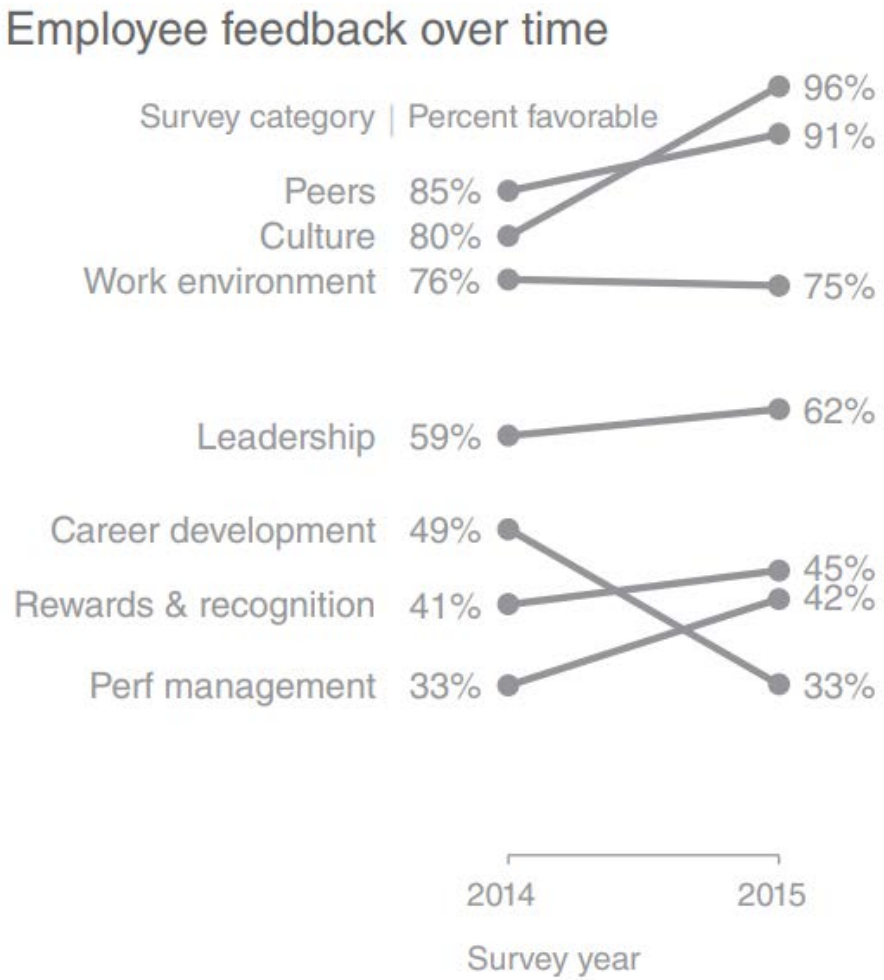
Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Slopegraph

- Useful when two time periods or points of comparison are present
- Can quickly show
  - relative increases and decreases or
  - differences across various categories between the two data points
- For Example:
  - To analyze data from an employee feedback survey, a slopegraph would be ideal to show the relative change in survey categories from 2014 to 2015



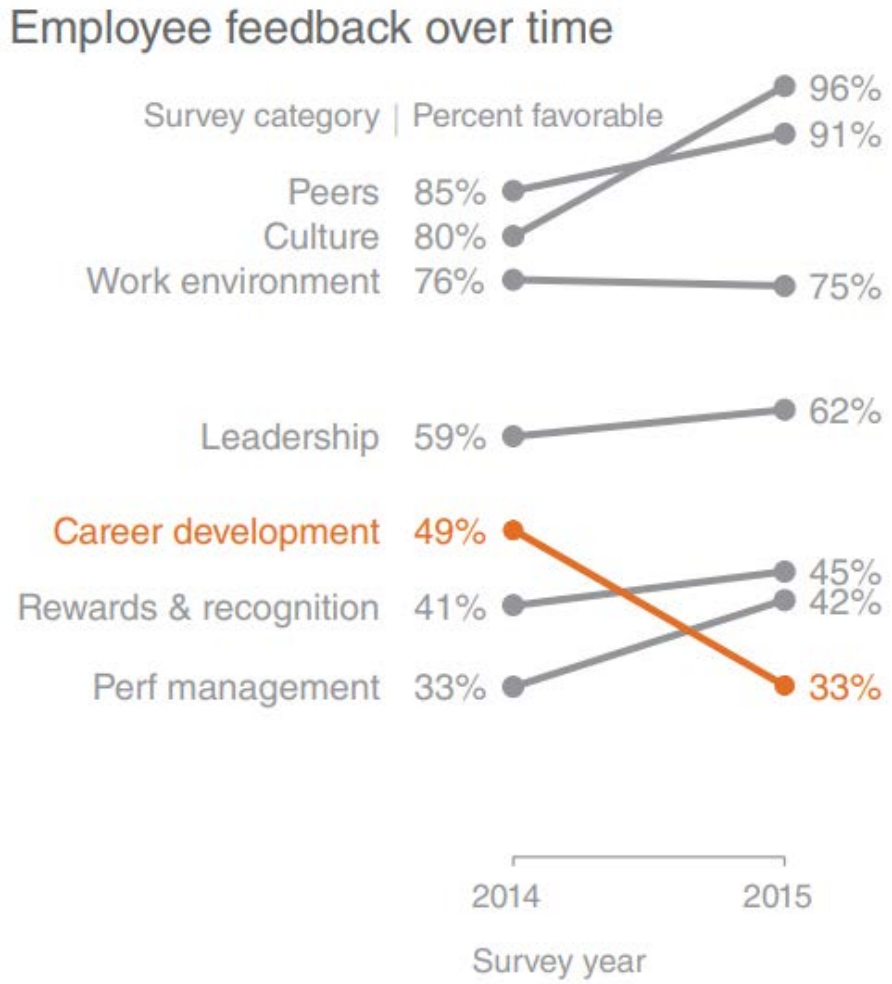
Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Slopegraph

- In addition to the absolute values (the points), the lines that connect them give a visual increase or decrease in rate of change
- A slopegraph may not work if many of the lines are overlapping
- In some cases, we can emphasize a single series at a time
- For example
  - We can draw attention to the single category that decreased over time with a different color

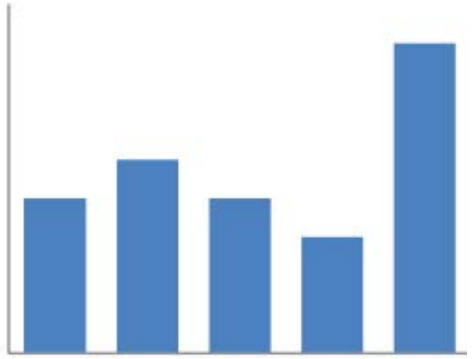


Source: Storytelling with Data by knaflic

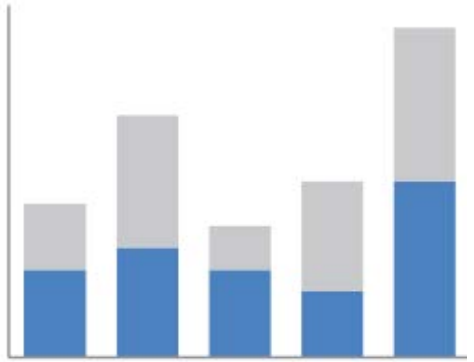
Graph-based Visuals

Bars

- Bar charts should be leveraged because they are common
  - This means less of a learning curve for our audience
  - They can spend time figuring out what information to take away from the visual
- Visually, bars are easy to read because our eyes compare the end points of the bars
  - It is easy to see which category is the biggest, the smallest, and also the incremental difference between categories
- Subtypes :
  - Vertical / Horizontal
  - Stacked
  - Waterfall



Vertical bar



Stacked vertical bar

Source: Storytelling with Data by knaflic

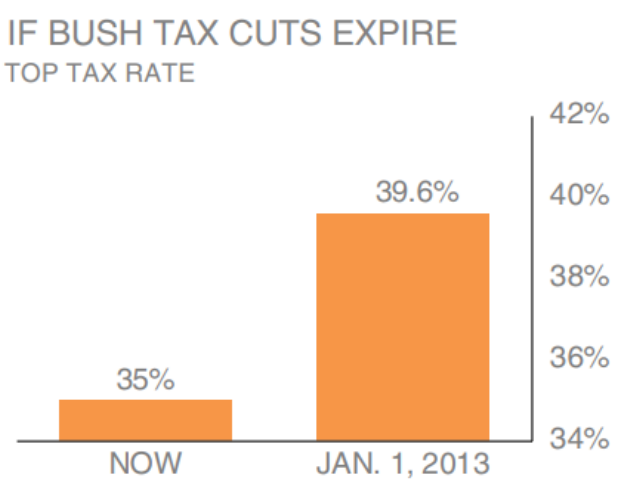
Graph-based Visuals

innovate achieve lead

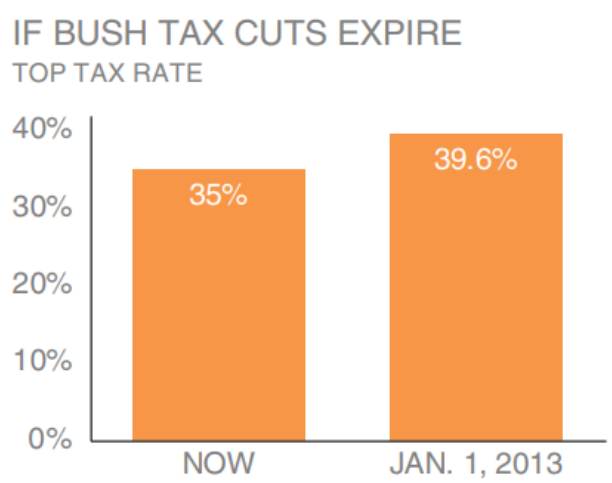
Bars

- It is important for bar charts to always have a zero baseline (where the x-axis crosses the y-axis at zero)
  - Because our eyes compare the relative end points of the bars
- Graph on the left side, the visual increase is 460%
  - The heights of the bars are 35 – 34 = 1 and 39.6 – 34 = 5.6
  - So (5.6 – 1) / 1 = 460%
- Graph on the right side, the visual increase is 13%
  - heights are accurately represented (35 and 39.6)
  - Actual visual increase = (39.6 – 35) / 35 = 13%

Non-zero baseline: as originally graphed



Zero baseline: as it should be graphed

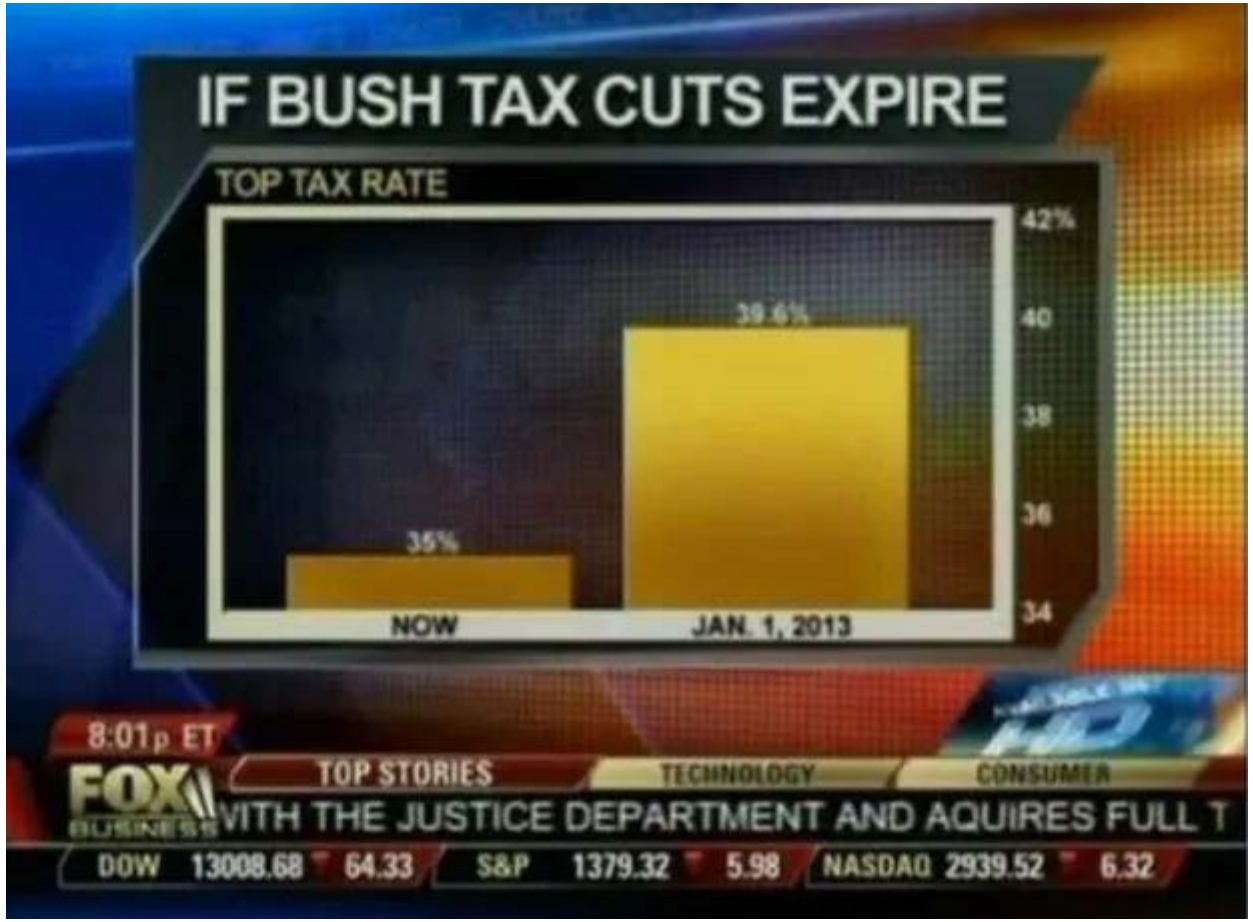


Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Bars – Real-life Example

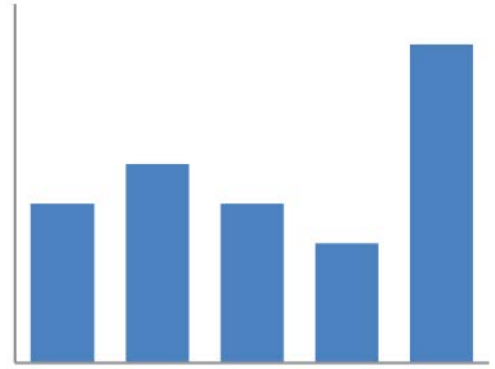


Source: Storytelling with Data by knaflic

Graph-based Visuals

Bars

- Vertical Bar Chart / Column Chart
  - Shows one or more than one series
  - Use multi series bars with care as more series clutters it
- Horizontal Bar chart
  - Single go-to graph for categorical data
  - Extremely easy to read
  - Useful when category names are long
  - Can show one or more than one series



Vertical bar

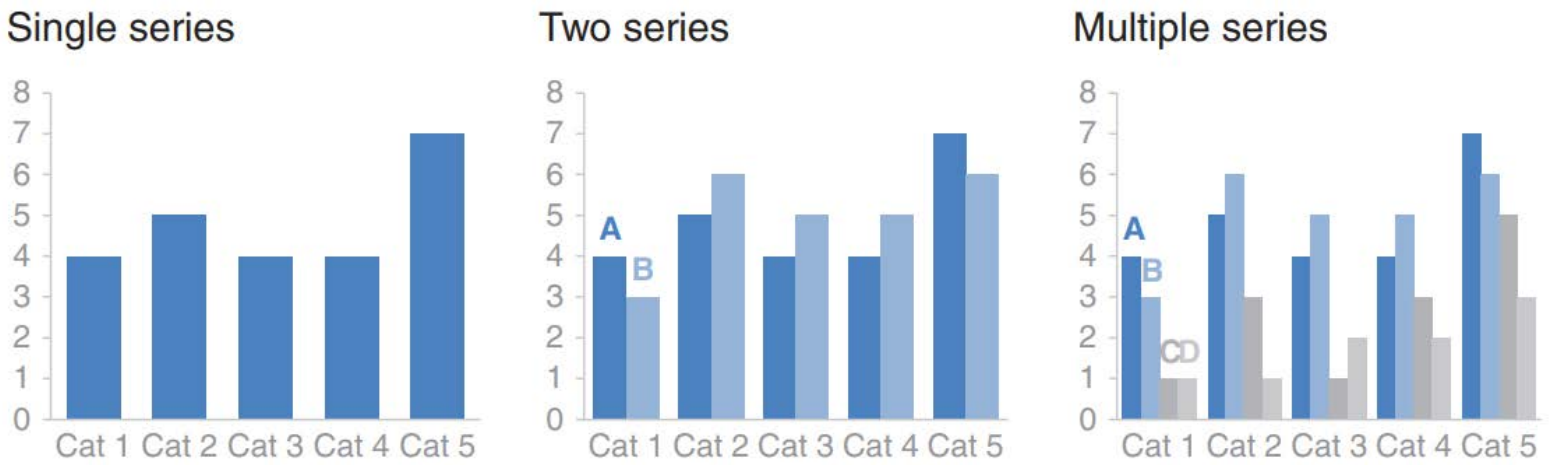
Horizontal bar

Source: Storytelling with Data by knaflic

Graph-based Visuals

Vertical Bar Charts

- Like line graphs, vertical bar charts can be single series, two series, or multiple series
- As more series of data added to the chart, it becomes more difficult to focus on one at a time
- There is visual grouping that happens as a result of the spacing in bar charts having more than one data series
  - This makes the relative order of the categorization important

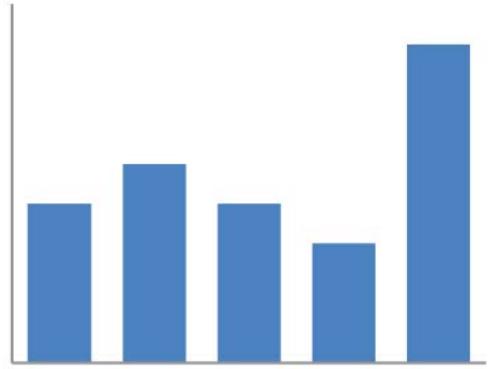


Source: Storytelling with Data by knaflic

Graph-based Visuals

Bar Charts – Design Best Practices:

- Use consistent colors throughout the chart
- Select accent colors to highlight meaningful data points or changes over time
  - Accent colors are those colors that are used for emphasizing in a color scheme.
- Use horizontal labels to improve readability
- Start the y-axis at 0 to appropriately reflect the values in the graph
- Make sure the bar width is wider than the white space between the bars



Vertical bar

Horizontal bar

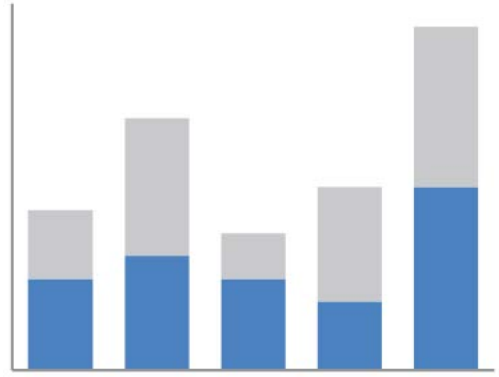
Source: Storytelling with Data by knaflic

Graph-based Visuals

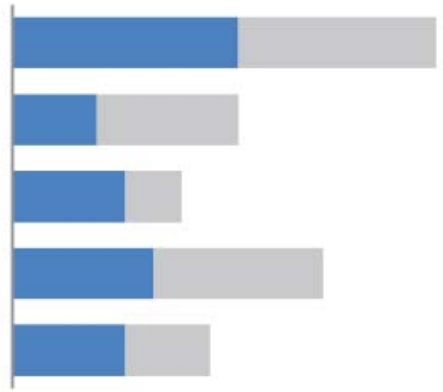
innovate achieve lead

Stacked Vertical/Horizontal Bar Chart

- Meant to allow comparison of totals across categories
- Also allows us to see the subcomponent pieces within a given category
- This can quickly become visually overwhelming
- Hard to compare subcomponents above the bottom series (the one directly next to the x-axis)
  - because there is no consistent baseline to compare



Stacked vertical bar



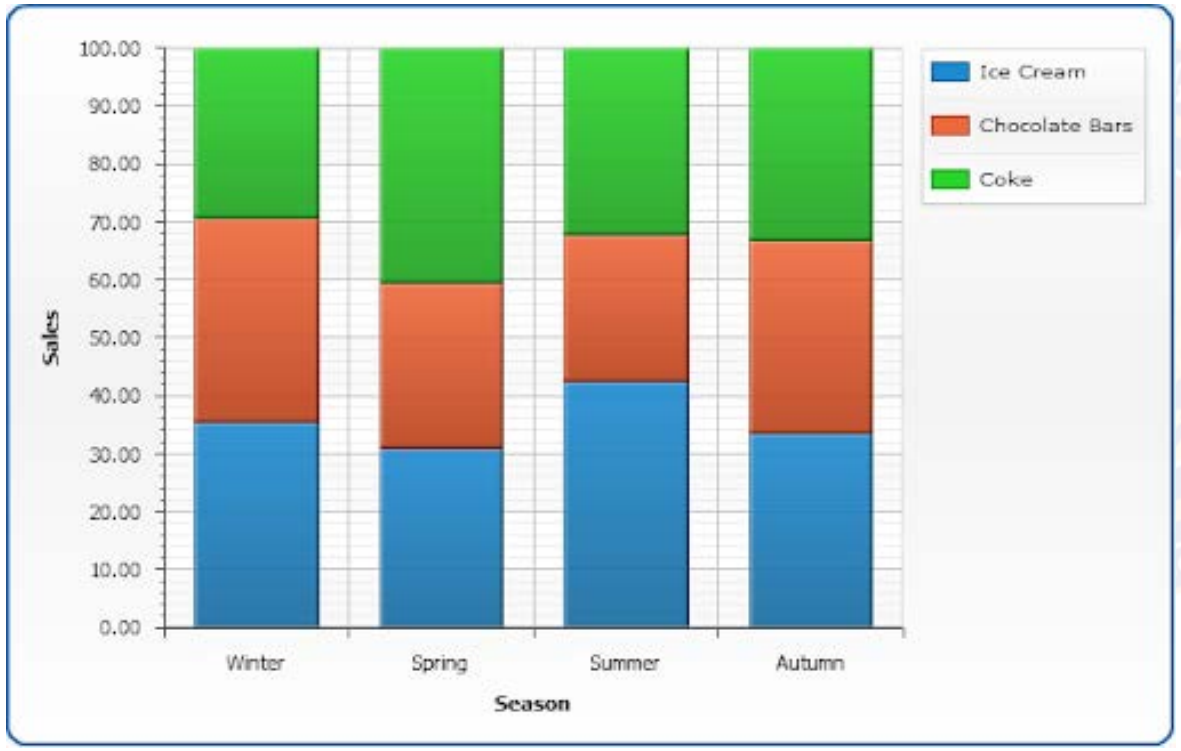
Stacked horizontal bar

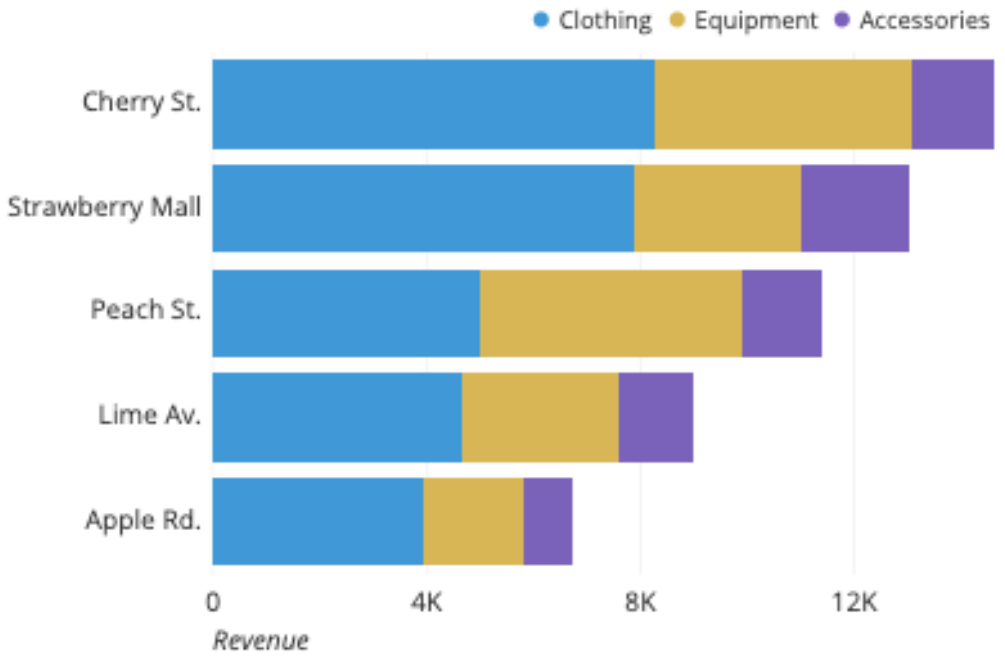
Source: Storytelling with Data by knaflic

Graph-based Visuals

innovate achieve lead

Stacked bar charts



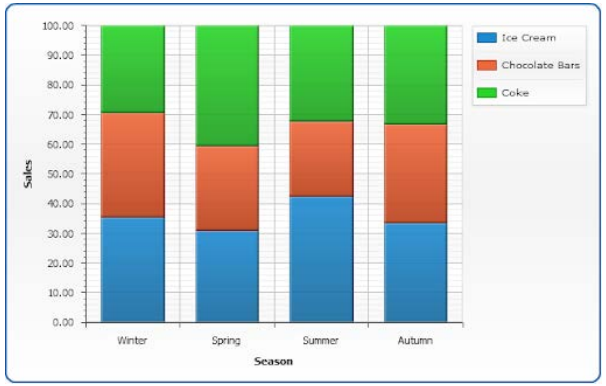


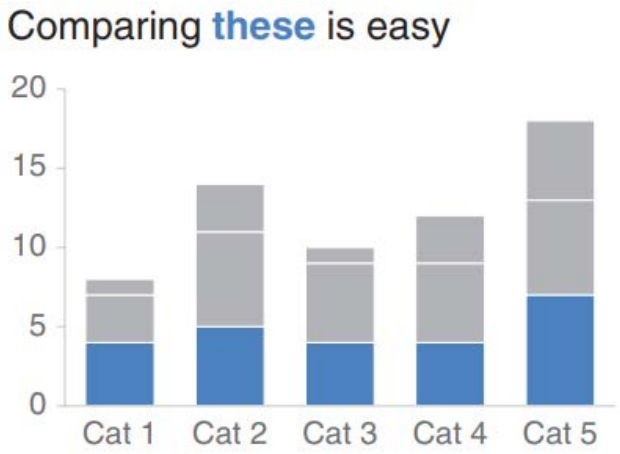
Source: Storytelling with Data by knaflic

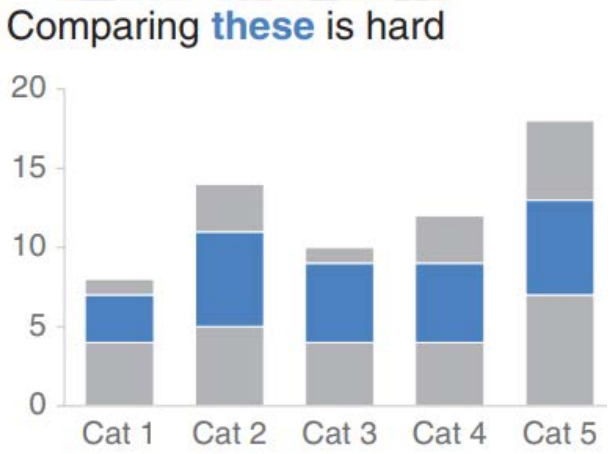
Graph-based Visuals

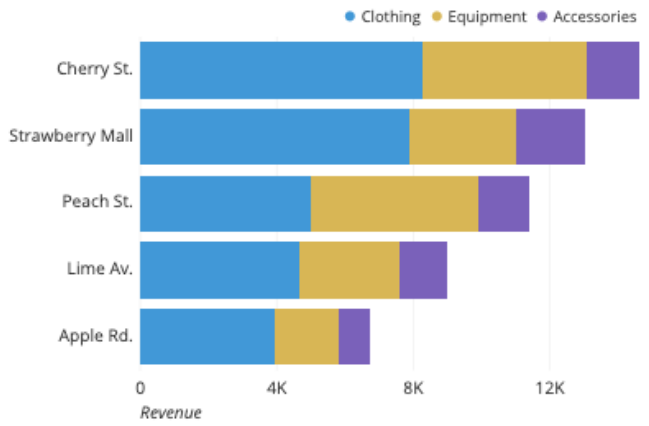
Stacked bar charts – Best Design Practices

- Best used to illustrate part-to-whole relationships
- Use contrasting colors for greater clarity
- Make chart scale large enough to view group sizes in relation to one another
- Show direct values or as percent









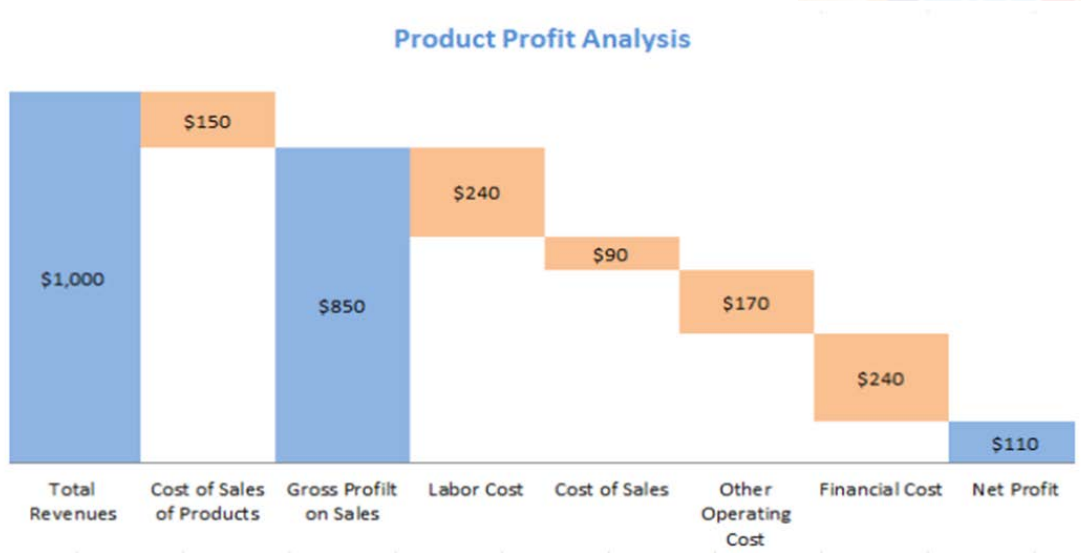
Source: Storytelling with Data by knaflic

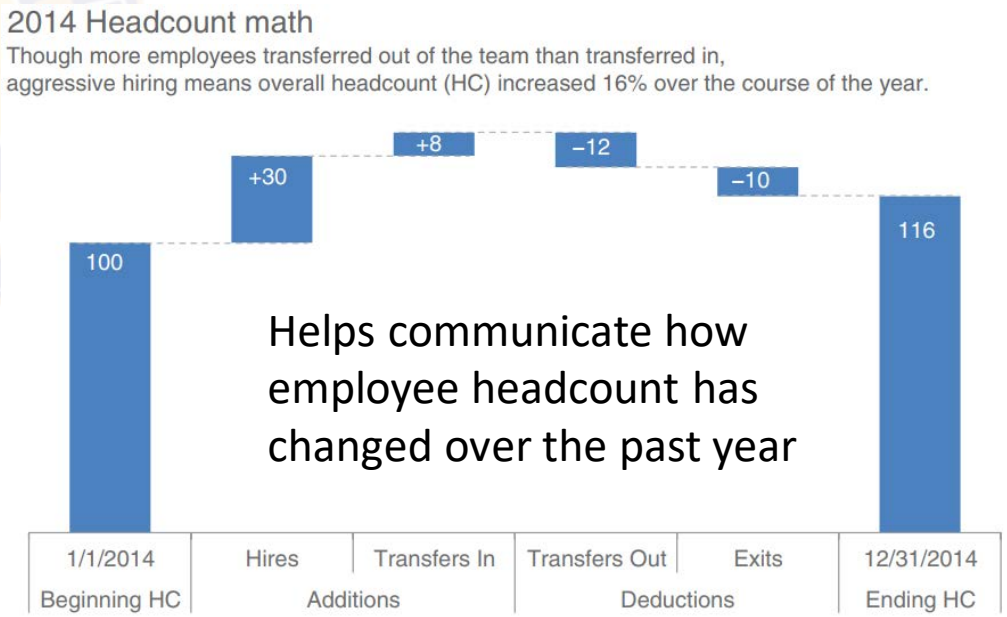
Graph-based Visuals

innovate achieve lead

Waterfall Chart

- Used to pull apart the pieces of a stacked bar chart to focus on one at a time
- Used to show a starting point, increases and decreases, and the resulting ending point
- Used to reveal the composition of a number





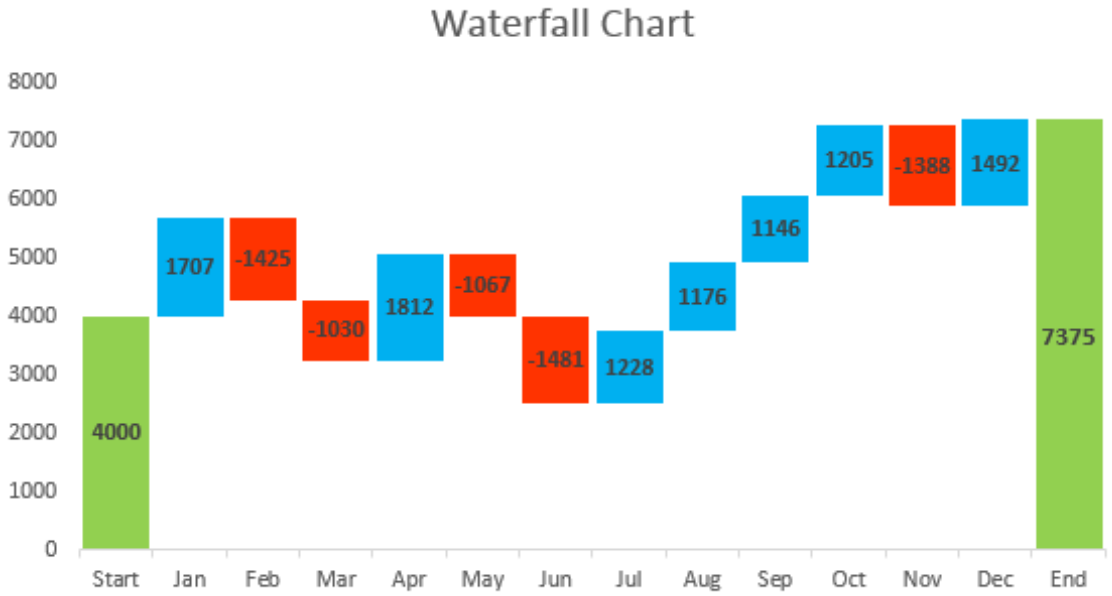
Helps communicate how employee headcount has changed over the past year

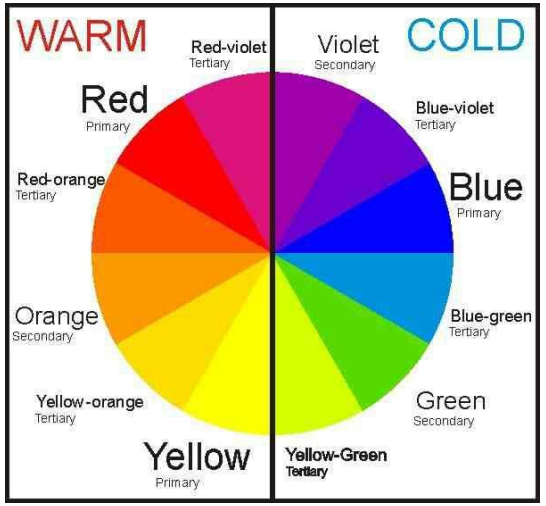
Source: Storytelling with Data by knaflic

Graph-based Visuals

Waterfall Chart – Best Design Practices

- Use contrasting colors to highlight differences in data sets.
- Choose warm colors to indicate increases and cool colors to indicate decreases





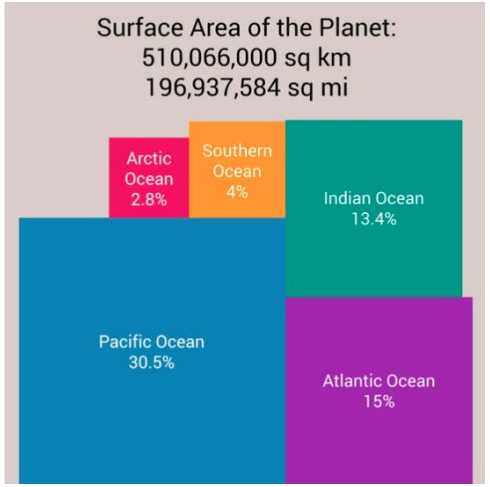
Source: Storytelling with Data by knaflic

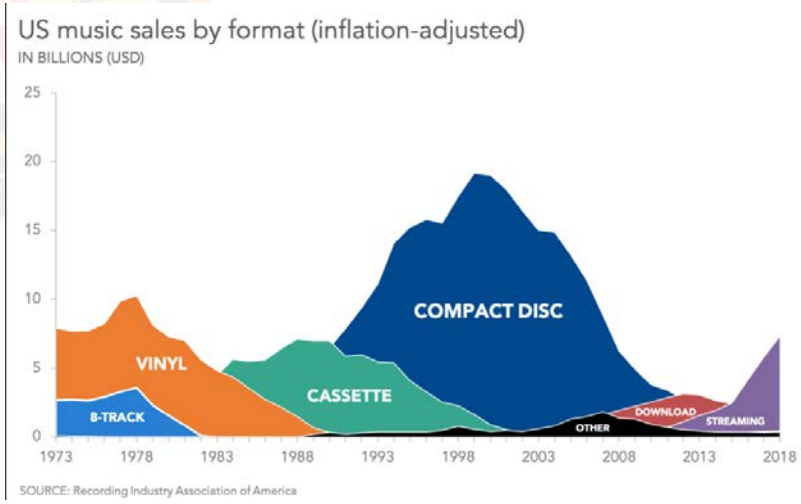
Graph-based Visuals

innovate achieve lead

Areas

- Basically a line chart, but the space between the x-axis and the line is filled with a color or pattern
- Useful for showing part-to-whole relations,
  - such as showing individual sales reps' contribution to total sales for a year
- Helps you analyze both overall and individual trend information





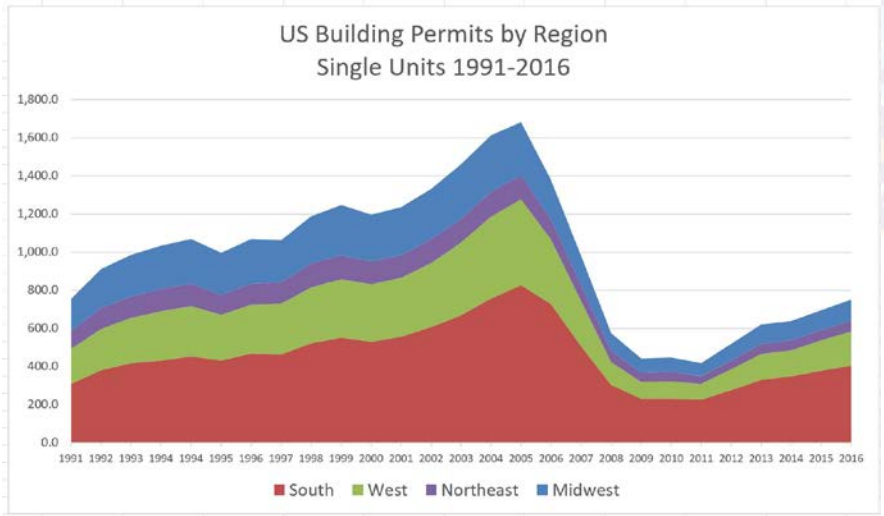
Source: Storytelling with Data by knaflic

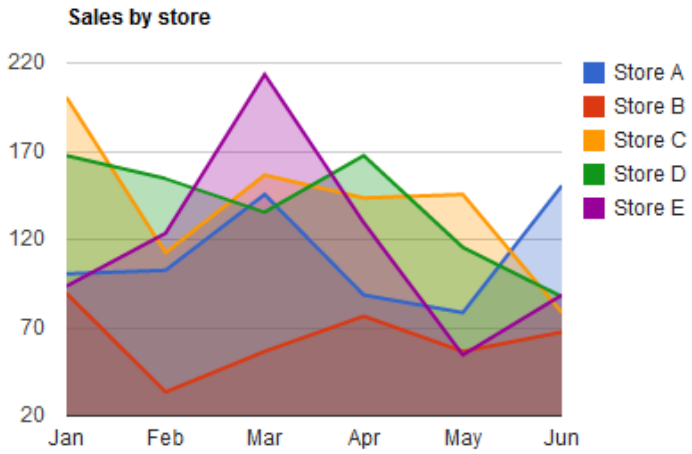
Graph-based Visuals

innovate achieve lead

Areas – Best Design Practices

- Use transparent colors so information isn't obscured in the background
  - E.g., highlighter color that lets the light pass through through
- Don't display more than four categories to avoid clutter.
- Organize highly variable data at the top of the chart to make it easy to read





Source: Storytelling with Data by knaflic

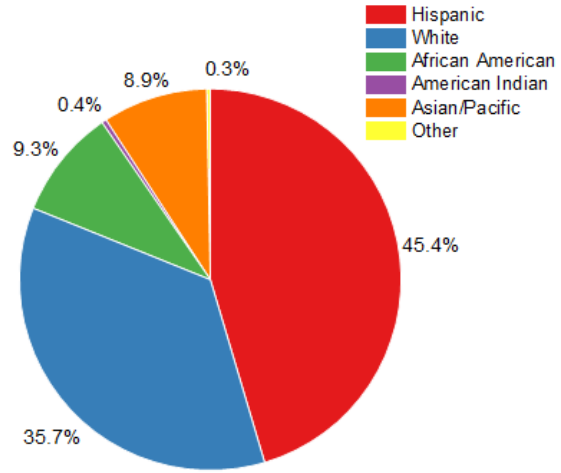
Graphs to be avoided

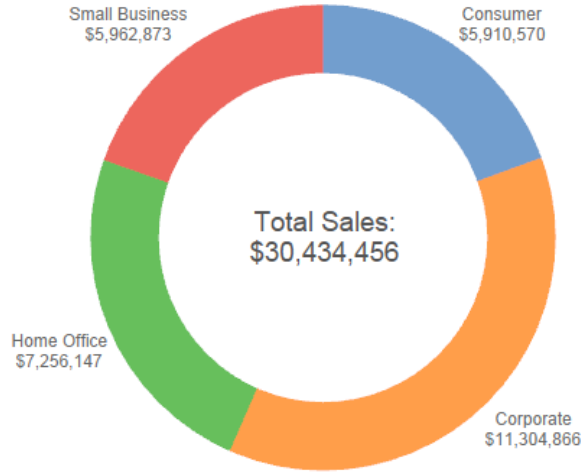
Graphs to be Avoided

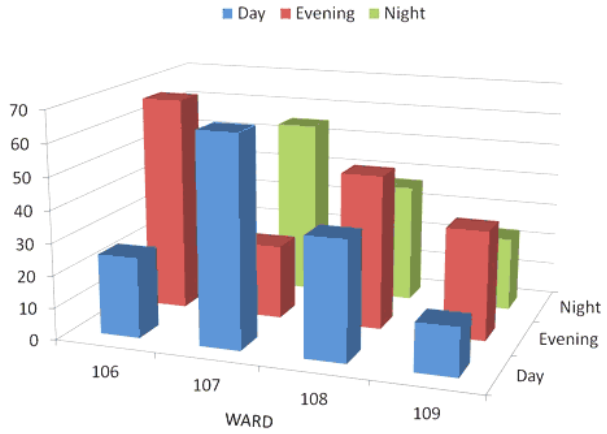
innovate achieve lead

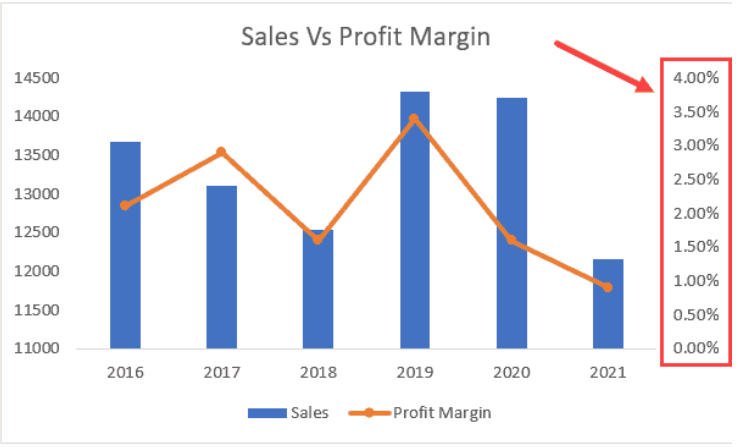
Four Types of Graphs

- Pie charts
- Donut charts
- 3D charts
- Dual Axis charts







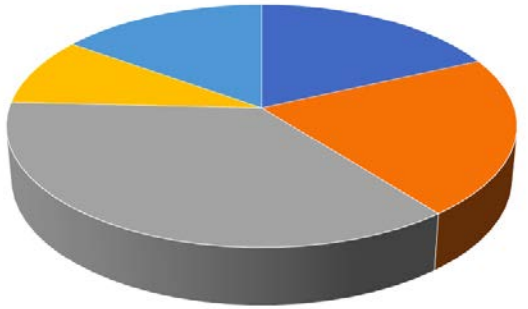


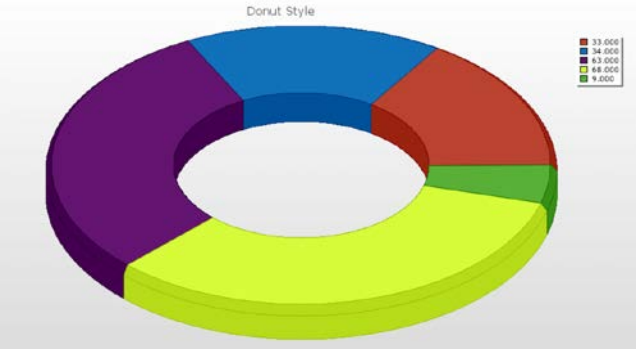
Source: Storytelling with Data by knaflic

Graphs to be Avoided

Pie & Donut Charts

- Most widely used (and misused)
- Used to visualize a part to whole relationship or a composition
- Typically represents numbers in percentages,
- The human mind thinks linearly
  - but, when it comes to angles and areas, most of us can't judge them well.





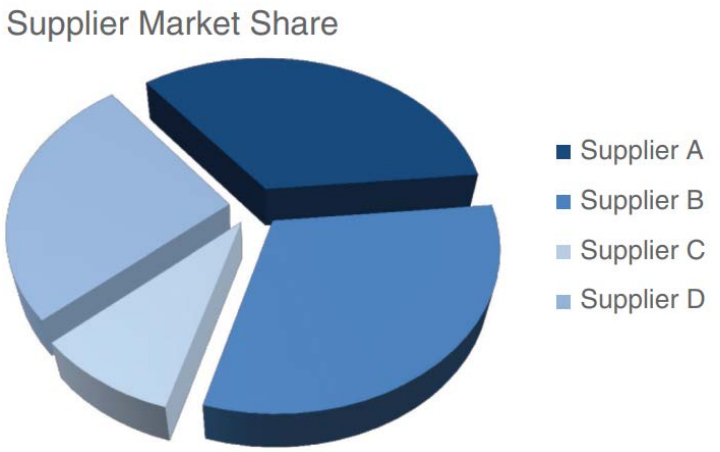
Source: Storytelling with Data by knaflic

Graphs to be Avoided

innovate achieve lead

Pie Charts

- The pie chart shown in the figure shows market share across four suppliers: A, B, C, and D
- Which supplier is the largest based on this visual?
  - Most people will agree that "Supplier B," (medium blue at the bottom right) appears to be the largest
- What proportion supplier B makes up of the overall market?
  - We might estimate 35% or 40%



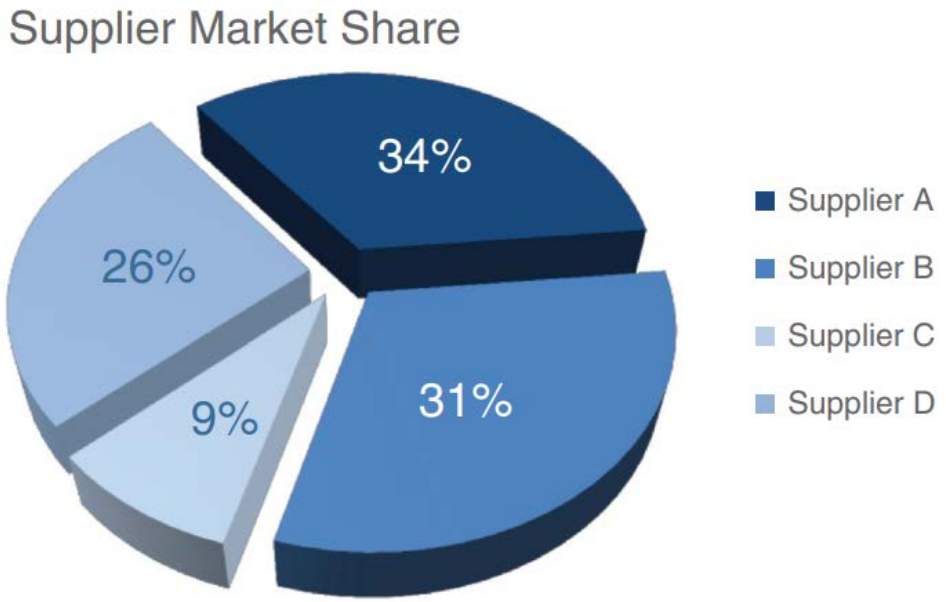
Source: Storytelling with Data by knaflic

Graphs to be Avoided

innovate achieve lead

Pie Charts

- 3D tilted view distorts the figure and makes the pieces at the top appear farther away
  - This makes them to appear smaller than they are
- Data visualization rule: don't use 3D!
  - It skews the visual perception of the numbers
- Even if we take 3D away
  - It's difficult to tell judge the size of segments when they are closer

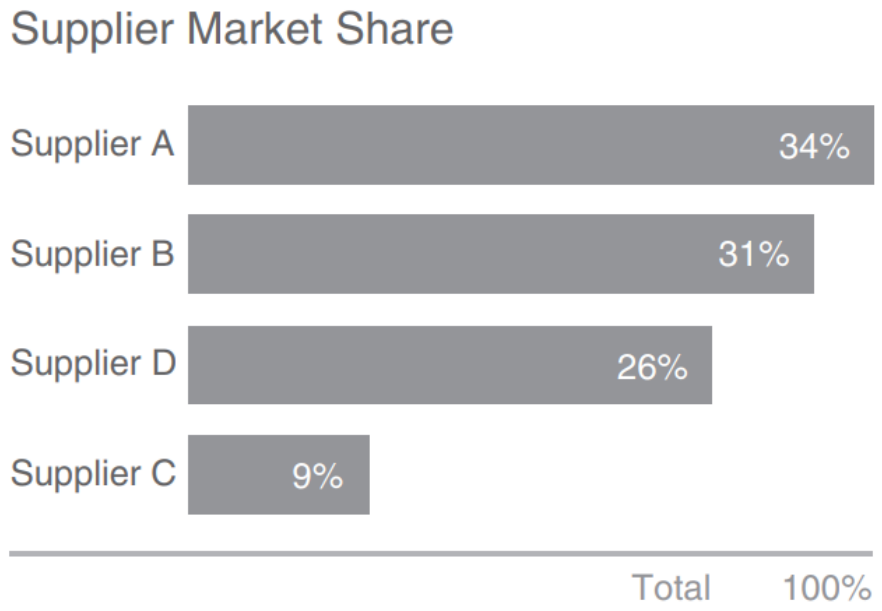


Source: Storytelling with Data by knaflic

Graphs to be Avoided

Pie Charts - Alternative

- One approach is to replace the pie chart with a horizontal bar chart
- We can order them from greatest to least or vice versa (unless there is some other natural ordering to be followed)
- As the bar charts are aligned at a common baseline, it is easy to assess relative size
- This makes it easy to see the size of the segments and the incremental difference between the segments

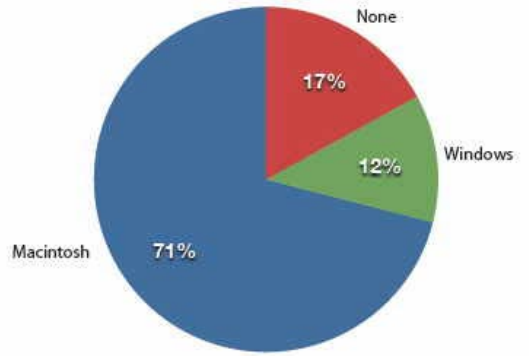


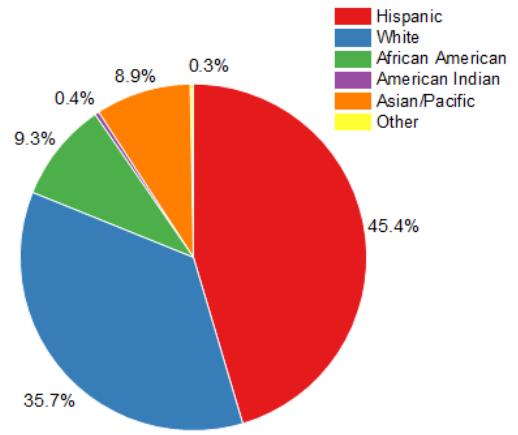
Source: Storytelling with Data by knaflic

Graphs to be Avoided

Pie Charts – Dos and Don'ts

- Make sure that the total sum of all segments equals 100 percent
- Use pie charts only if you have less than six categories, unless there's a clear winner you want to focus on
- Ideally, there should be only two categories, such as
  - men and women visiting a website
  - market share of your company, compared to the whole market
- Don't use a pie chart if the category values are almost identical
- You could add labels, but that's a patch, not an improvement
- Don't use 3D or blow apart effects
  - they reduce comprehension and show incorrect proportions



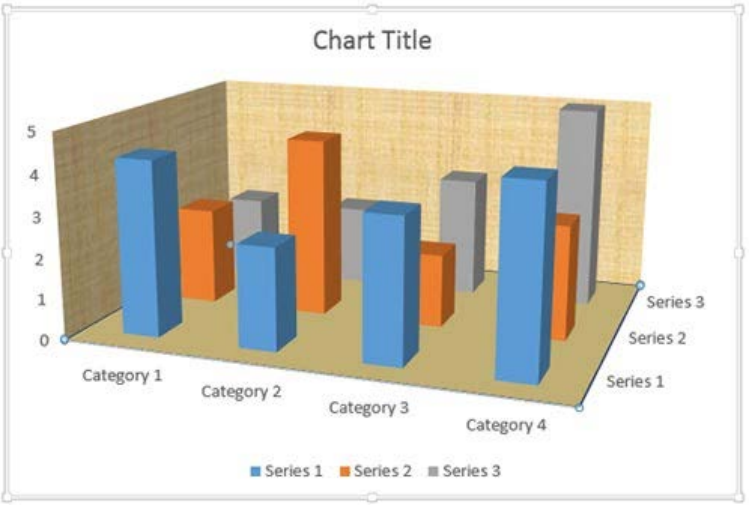


Source: Storytelling with Data by knaflic

Graphs to be Avoided

3D Charts

- Golden Rule: Never use 3D
  - Only use when really third dimension is needed
- Never use 3D to plot to show one or two dimensions
- 3D skews our numbers
  - Makes difficult interpret numbers
  - Makes comparisons challenging

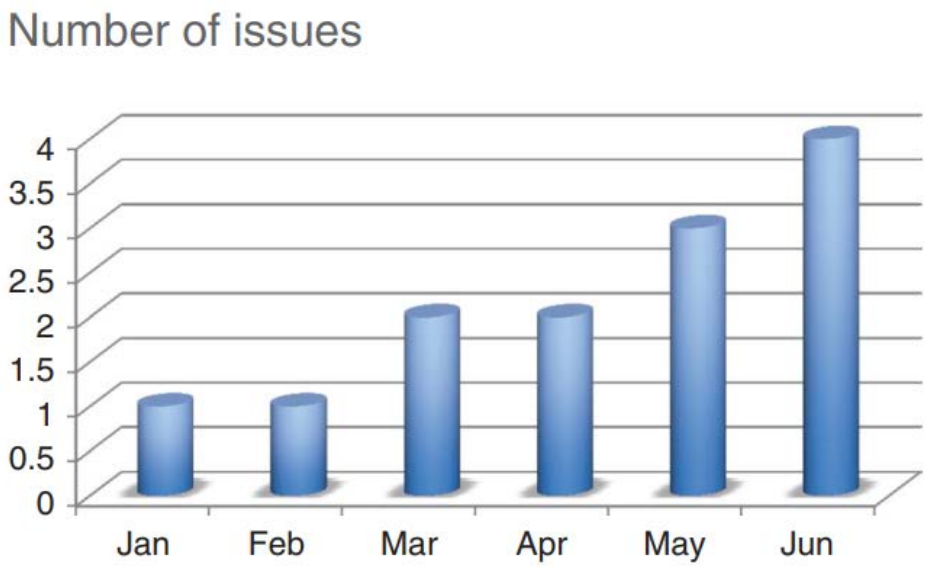


Source: Storytelling with Data by knaflic

Graphs to be Avoided

3D Charts

- 3D graphs introduces unnecessary chart elements like side and floor panels
- Makes the bar height to be determined by an invisible tangent plane intersecting the corresponding height on the y-axis
- How many issues were there in January and February?
  - Actually, there is only single issue in each of these months
  - Visually, we may estimate a value of 0.8



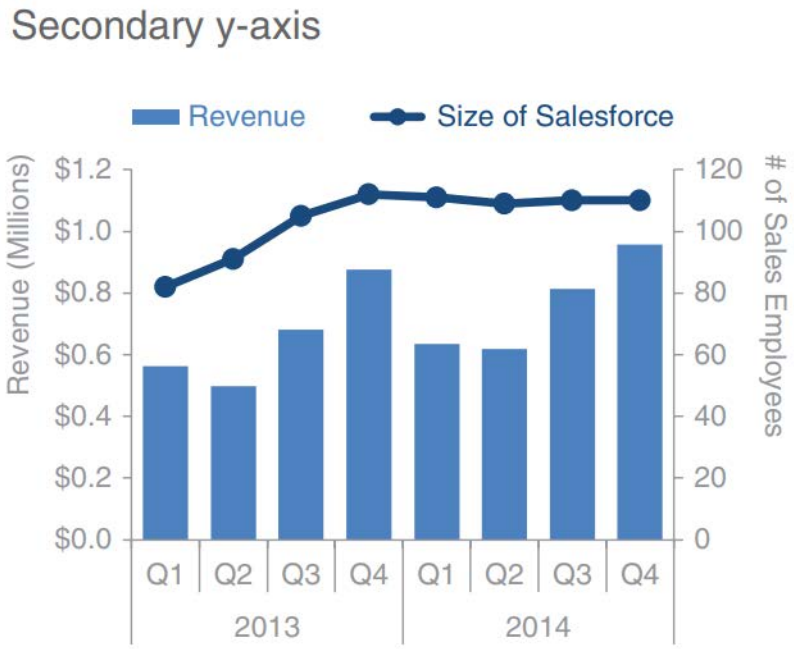
Source: Storytelling with Data by knaflic

Graphs to be Avoided

innovate achieve lead

Dual Axis Charts

- Sometimes it's useful to plot data that is in entirely different units against the same x-axis
- This often gives rise to the secondary y-axis
  - another vertical axis on the right-hand side of the graph.
- Used to show relationships and compare variables on vastly different scales
- Much more difficult to read and understand



Source: Storytelling with Data by knaflic

Graphs to be Avoided

innovate achieve lead

Dual Axis Charts - Alternative

- Don’t show the second y-axis
  - Instead, label the data points that belong on this axis directly

Alternative 1: **label directly**

Alternative 2: **pull apart vertically**

- Pull the graphs apart vertically
  - Have a separate y-axis for each (both along the left)
  - leverage the same x-axis across both.

Source: Storytelling with Data by knaflic

Types of Visualization

Recap

- Text Based
  - Simple text
  - Table
  - Heatmap
- Graphs
  - Points
  - Lines
  - Bars
  - Areas
- To be avoided
  - Pie and Donut
  - 3D
  - Dual Axis

Source: Storytelling with Data by knaflic

Thank You!