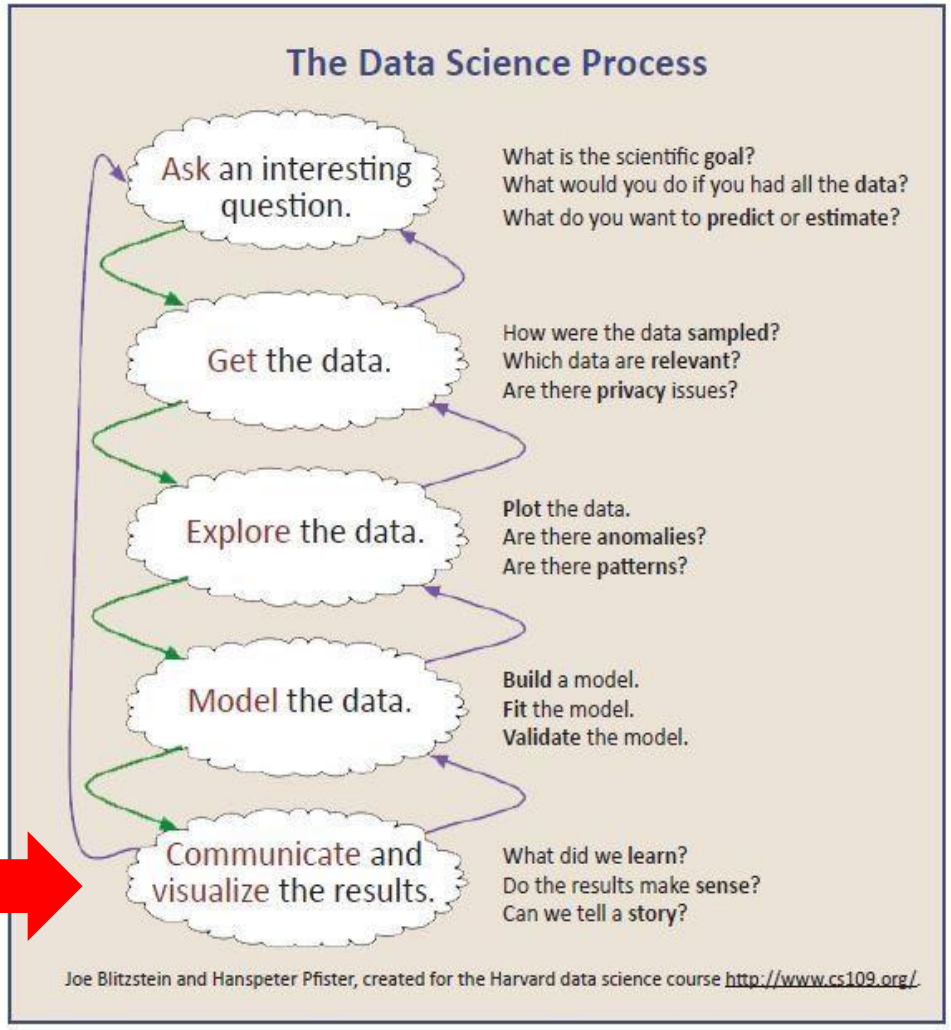
CDS6214
Data Science Fundamentals

Lecture 7
Data Visualization

Outline

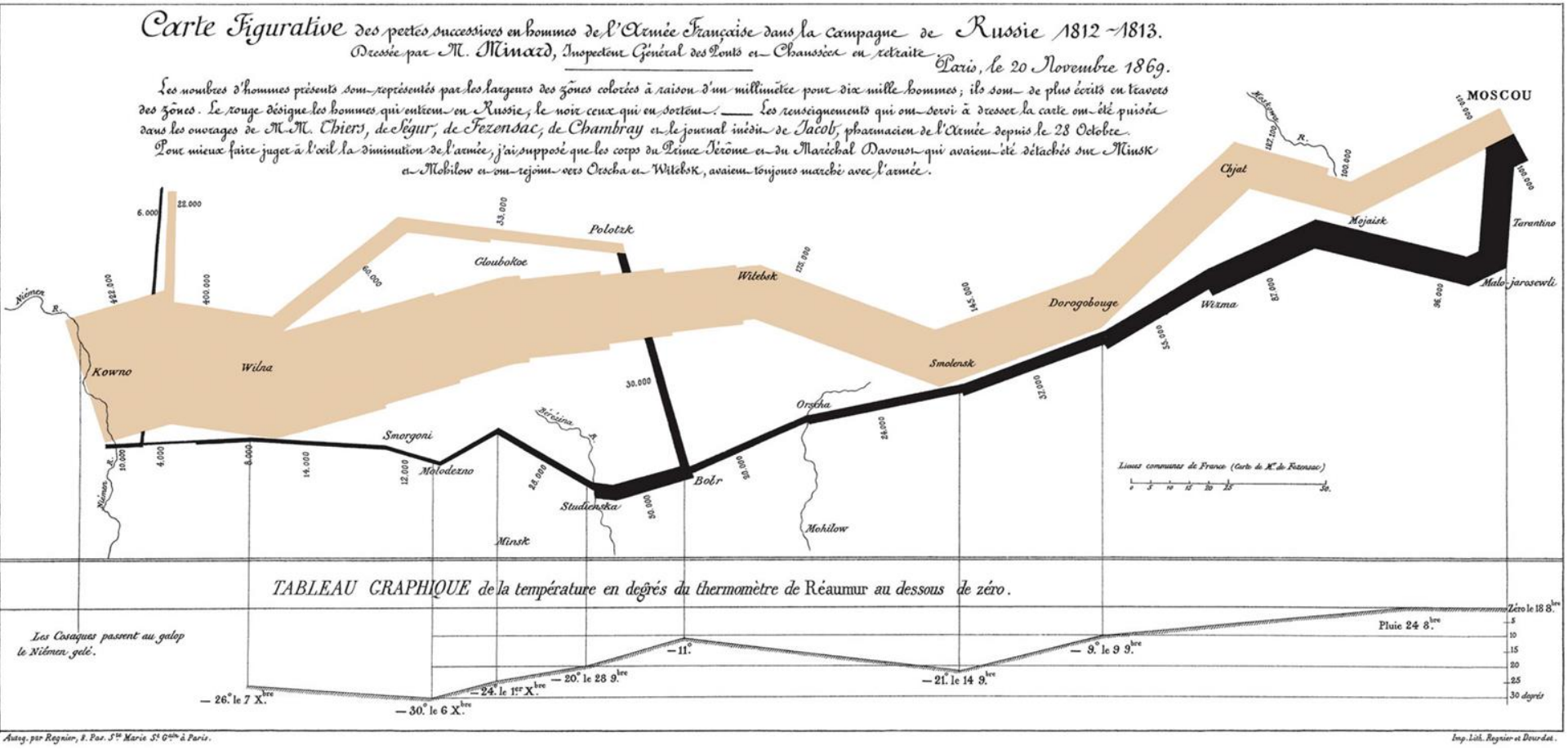
- Why Visualize?
- Storytelling with Data
- Types of Visualizations
- Best Practices for Data Visualization
- Visualization Dashboards
- Law of Simplicity

Data Science Process



What should we do next?

- Identify the question
- Collect and pre-process the data
- Explore and analyze the data
- Model the data
- Infer and visualize results



Napoleon's March

Why Visualize?

- To explore the unknown
- To analyze a hypotheses
- To present “everything” known about the data, and communicate the results

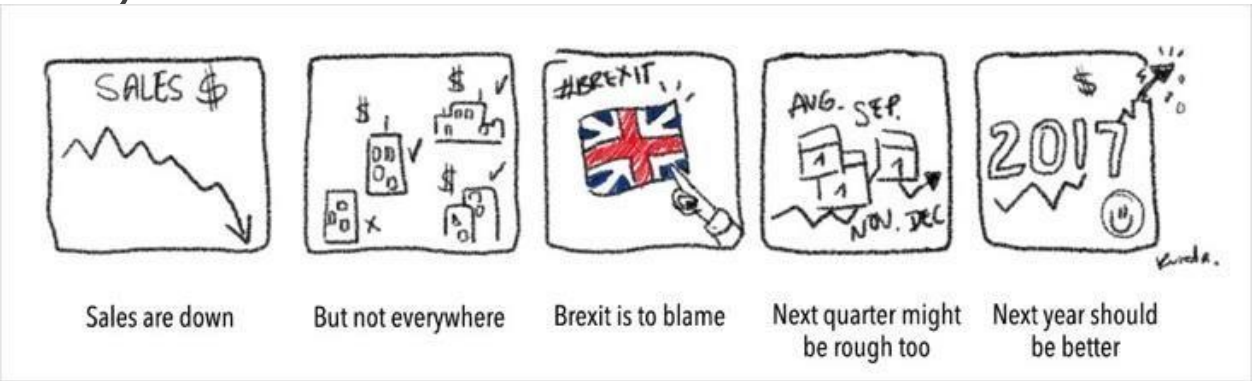
Storytelling with Data

- It is important to have a cohesive narrative. You must have a storyboard.
- Before creating a storyboard, preliminary exploration of the data is required.
- You will then need to (temporarily ignoring the numbers), focus on the flow of arguments..

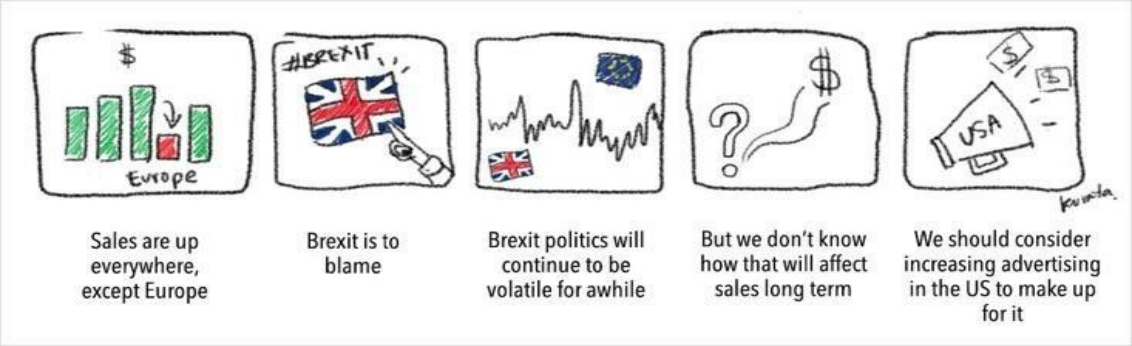
- Once a reasonable narrative is in place, drill down into the data to extract and generate the information required.
- Know what you're trying to say, inject emotion and suspense and interest into your narrative whenever possible

Varying stories for different audience

- Storyboard 1



- Storyboard 2



Good data visualization should…

- Help the audience think about the the important message(s) from the data, rather than about methodology (graphic design, the technology of graphic production etc), or something else
- Avoid distorting what the data have to say – report the truth
- Present many numbers in a small space - but also emphasize the important numbers
- Make large data sets coherent, and encourage the audience to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure

Design for your audience

- Test your visualisation with the key audience
- Limit the number of categories shown in a visualisation – be selective in what you present and emphasise the key message(s)
- Handy considerations:
  - Know when to use charts, and when to use tables*
  - Try to avoid using pie charts (unless your audience really does not like bar-charts!)*

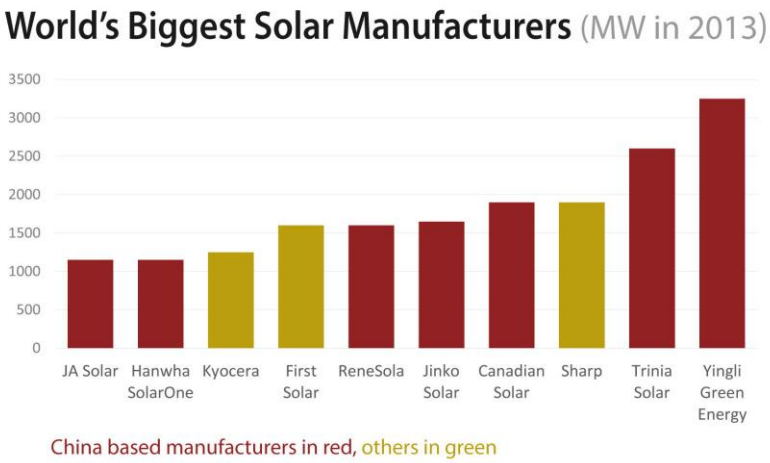
Tables

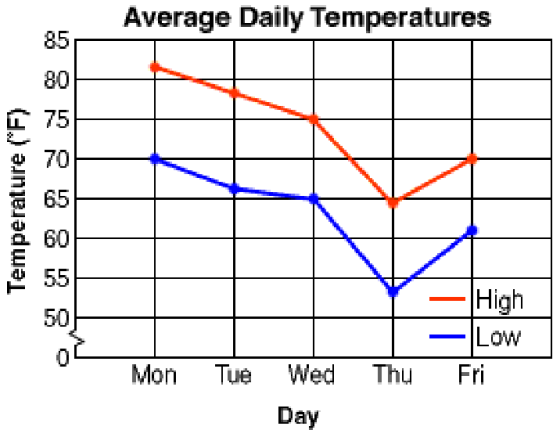
- **Tables** are useful when:
  - You want to show individual data values, and allow audience to compare between individual values
  - Precise values are required
  - The quantitative information to be communicated involves more than one unit of measure

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Graphs

- **Graphs** are useful when:
  - You want to reveal relationships between multiple values, for example broad comparison of trends over time or differences between areas
  - General patterns are the key point you want to present, rather than the exact data values





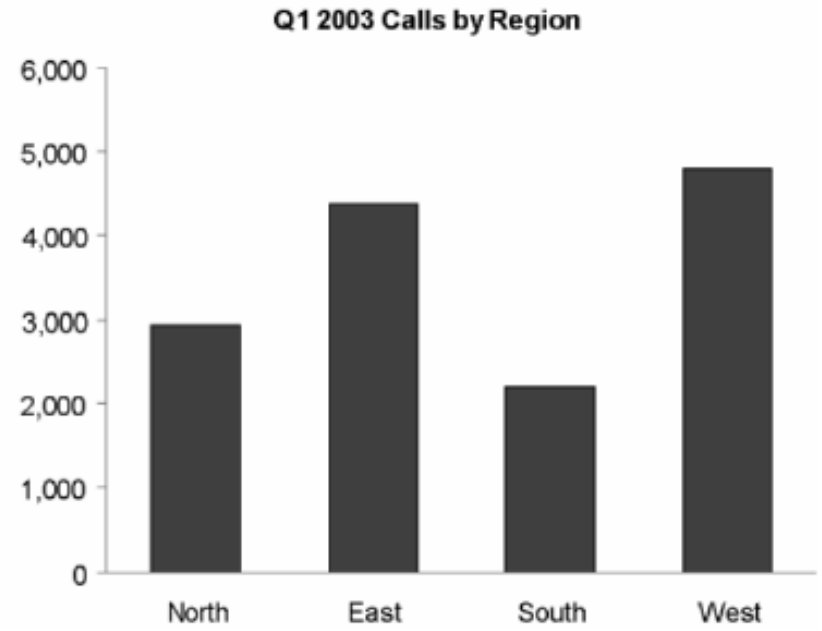
Source: https://solarlove.org/rooftop-solar-explained-graphs/

Seven Common Relationships in Graphs

- Nominal Comparison
- Time-series
- Ranking
- Part-to-Whole
- Deviation
- Frequency Distribution
- Correlation

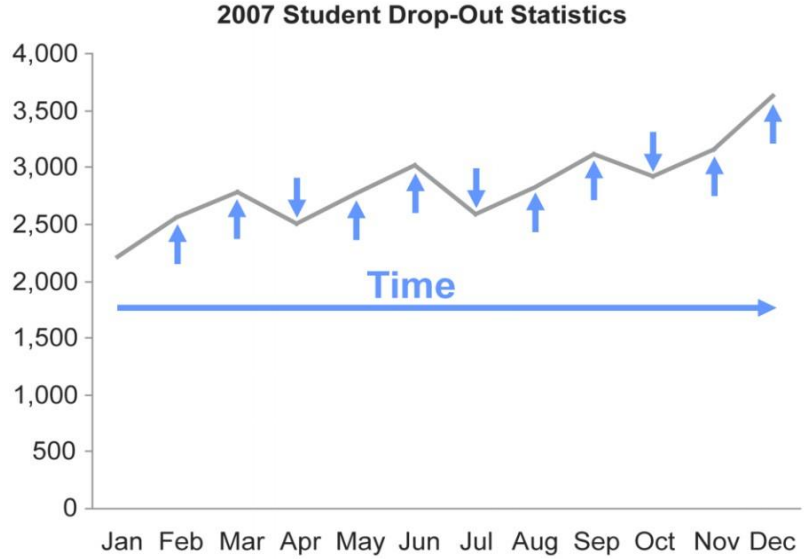
Nominal Comparison

Description	Methods
A simple comparison of the categorical subdivisions of one or more measures in no particular order	Bars only (horizontal or vertical)



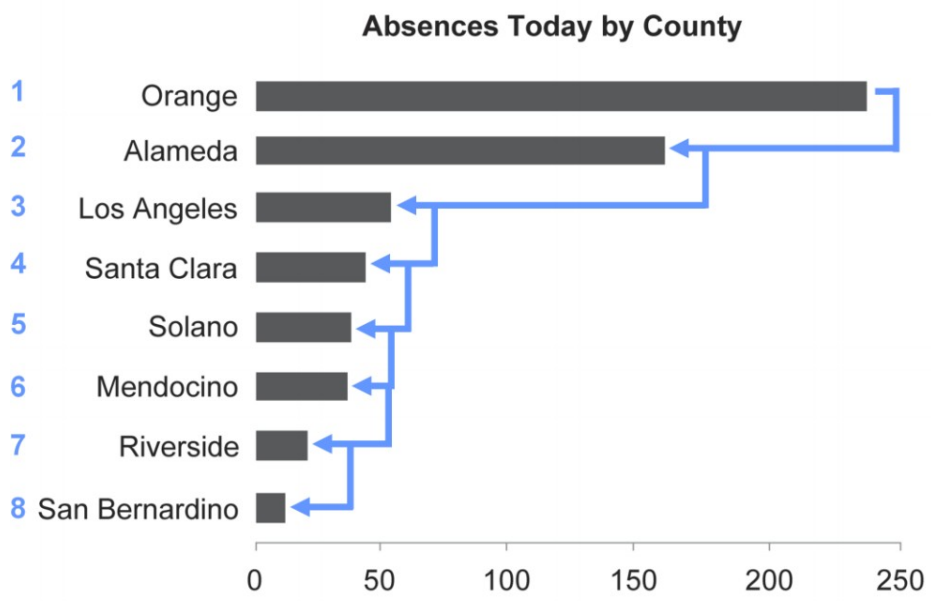
Time-Series

Description	Methods
Multiple instances of one or more measures taken at equidistant points in time	• Lines to emphasize overall pattern • Bars to emphasize individual values • Points connected by lines to slightly emphasize individual values while still highlighting the overall pattern • Always place time on the horizontal axis



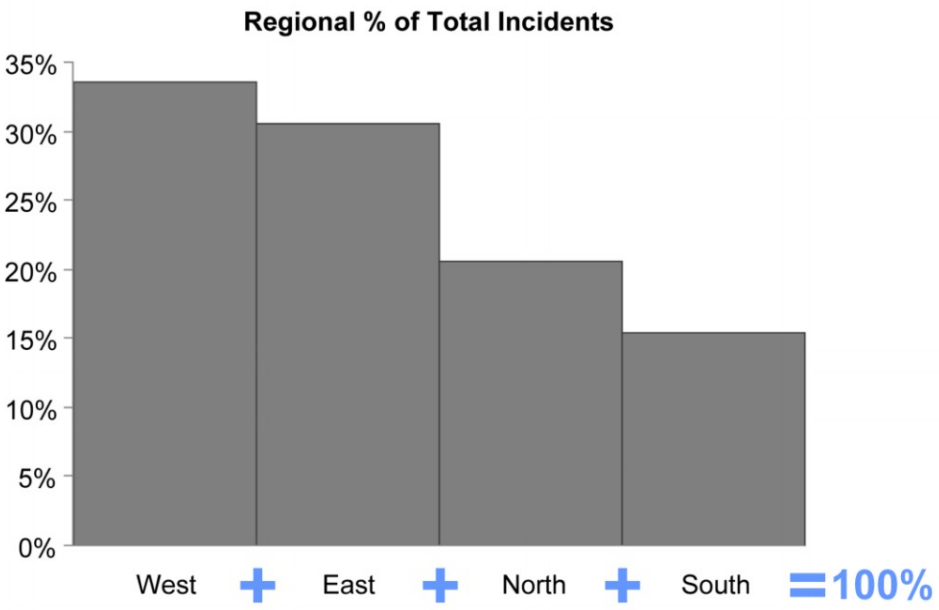
Ranking

Description	Methods
Categorical subdivisions of a measure ordered by size (either descending or ascending)	• Bars only (horizontal or vertical) • To highlight high values, sort in descending order • To highlight low values, sort in ascending order



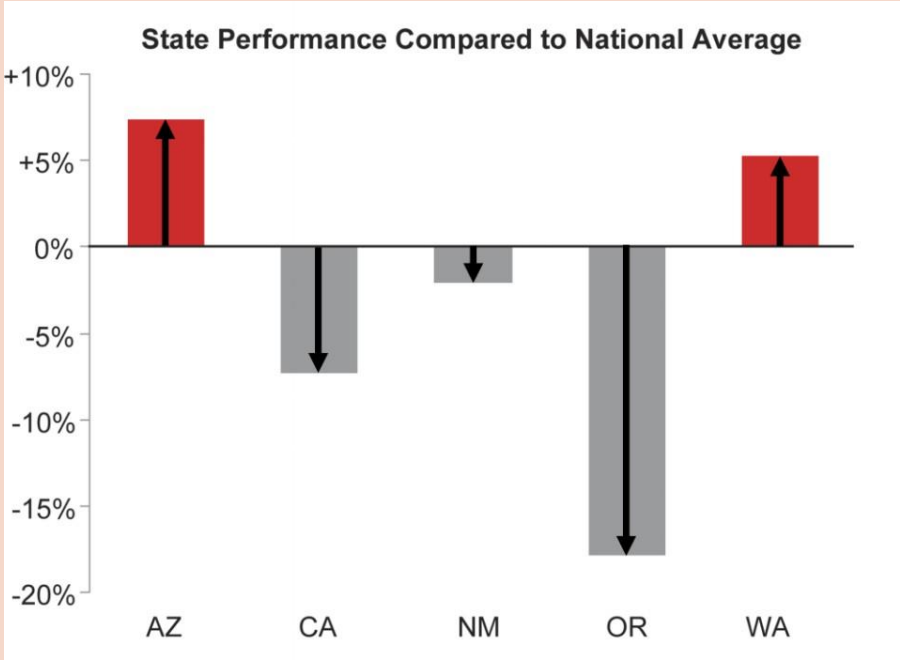
Part-to-Whole

Description	Methods
Measures of individual categorical subdivisions as ratios to the whole	• Bars only (horizontal or vertical) • Use stacked bars only when you must display measures of the whole as well as the parts



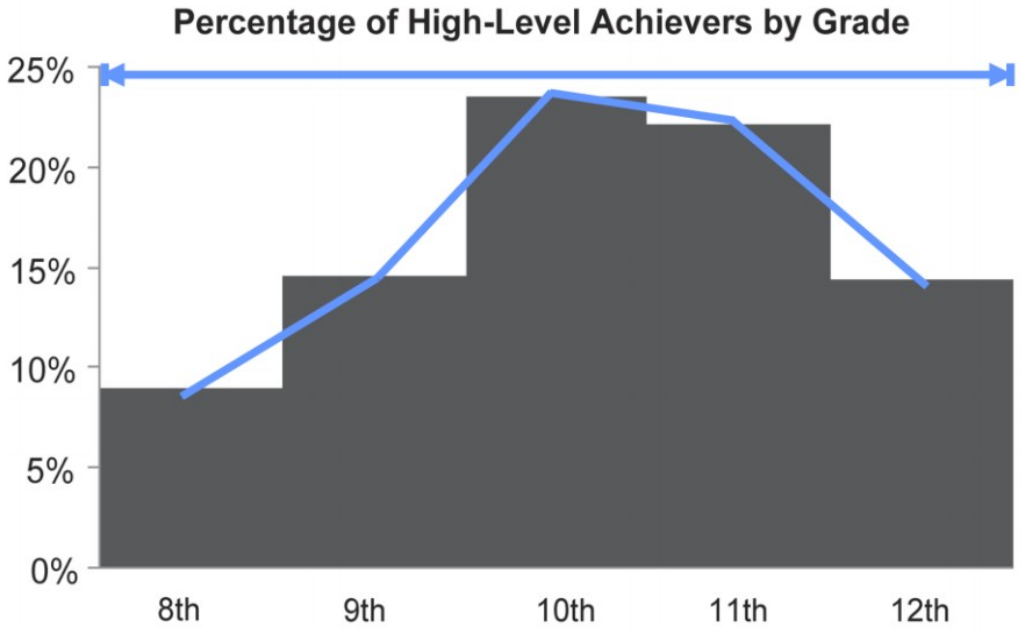
Deviation

Description	Methods
Categorical subdivisions of a measure compared to a reference measure, expressed as the differences between them	• Lines to emphasize the overall pattern only when displaying deviation and time-series relationships together • Points connected by lines: slightly emphasize individual data points while also highlighting the overall pattern when displaying deviation and time-series relationships together • Bars: Emphasize individual values, but limit to vertical bars when a time-series relationship is included • Always include a reference line to compare the measures of deviation



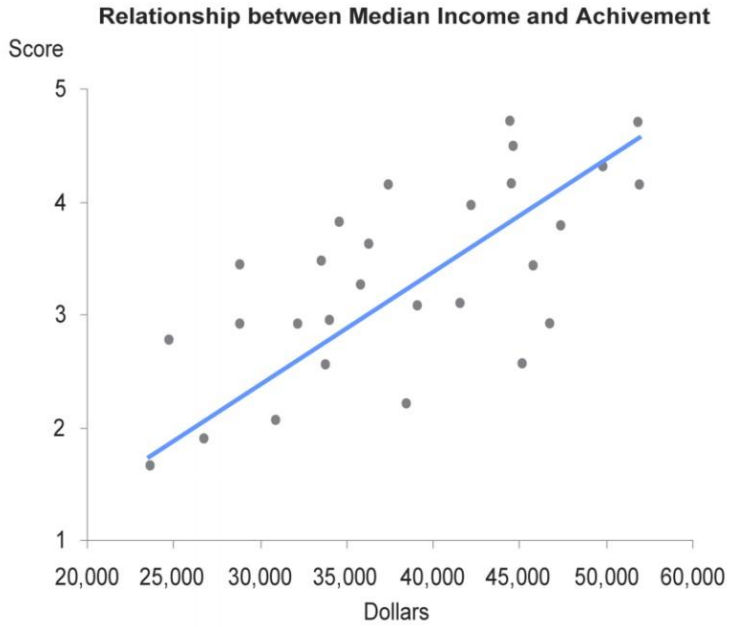
Frequency Distribution

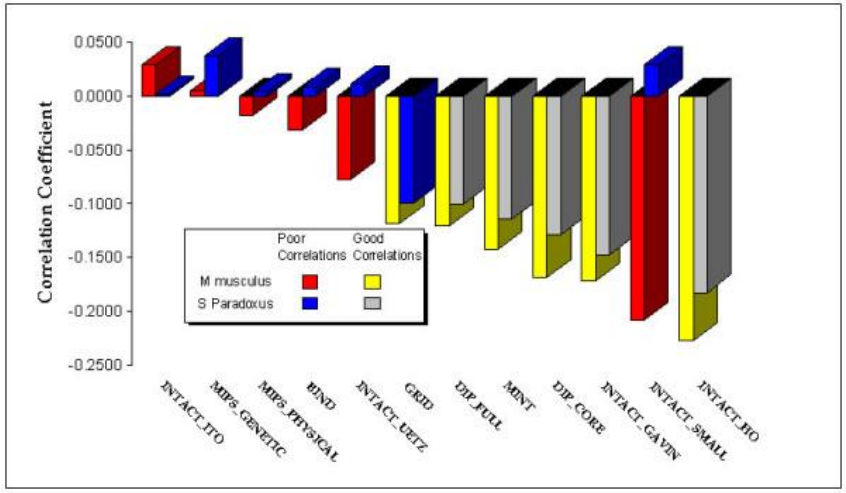
Description	Methods
Counts of something per categorical subdivisions (intervals) of a quantitative range	• Vertical bars to emphasize individual values (called a histogram) • Lines to emphasize the overall pattern (called a frequency polygon)



Correlation

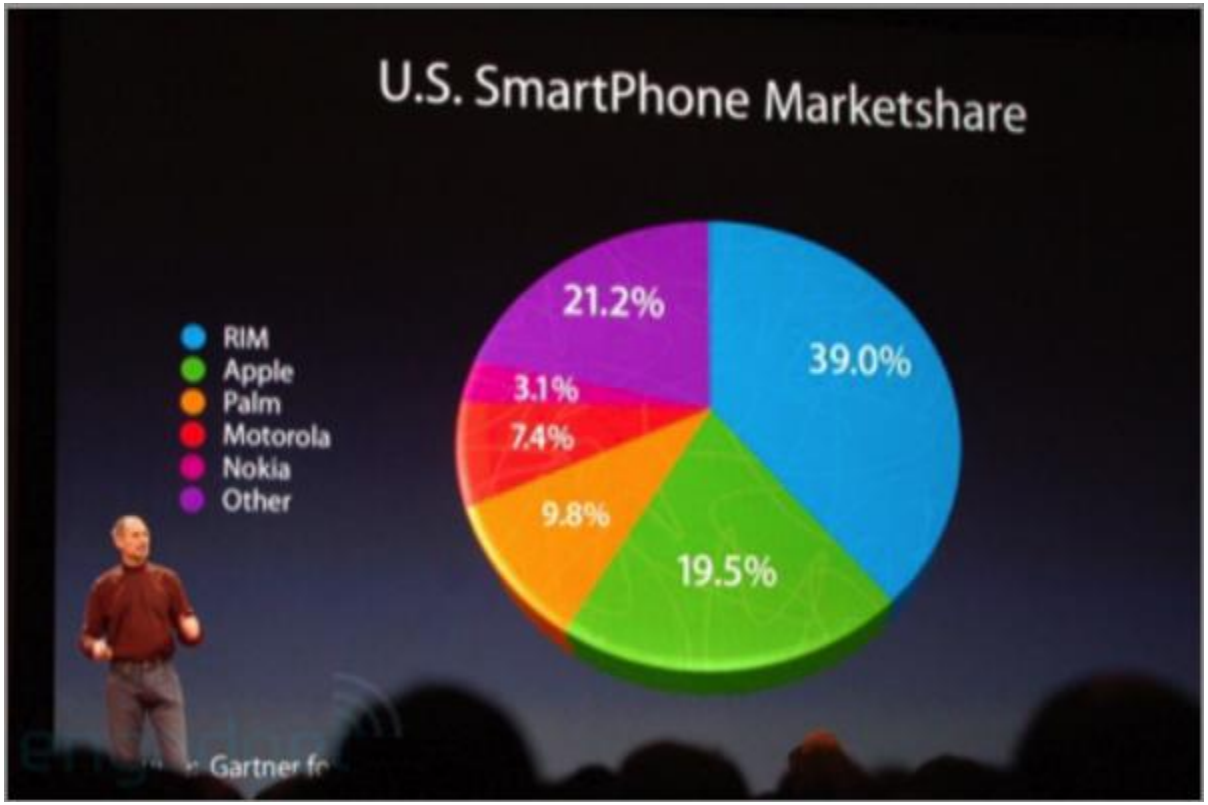
Description	Methods
Comparisons of two paired sets of measures to determine if one set goes up, the other set goes either up or down in a corresponding manner, and if so, how strongly	• Points and a trend line in the form of a scatter plot • Bars may be used, arranged as a paired bar graph or a correlation bar graph, if scatter plots are unfamiliar





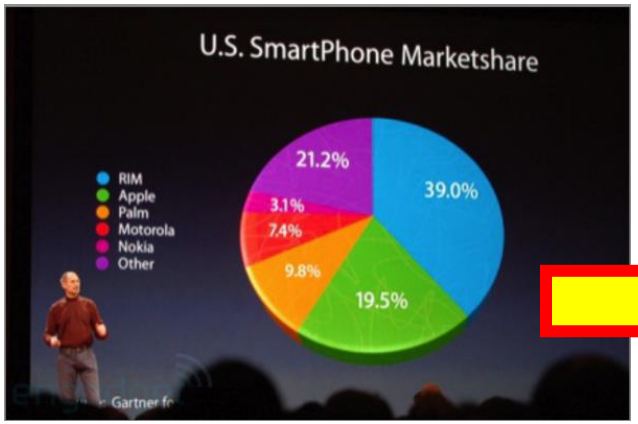
Correlation bar graph

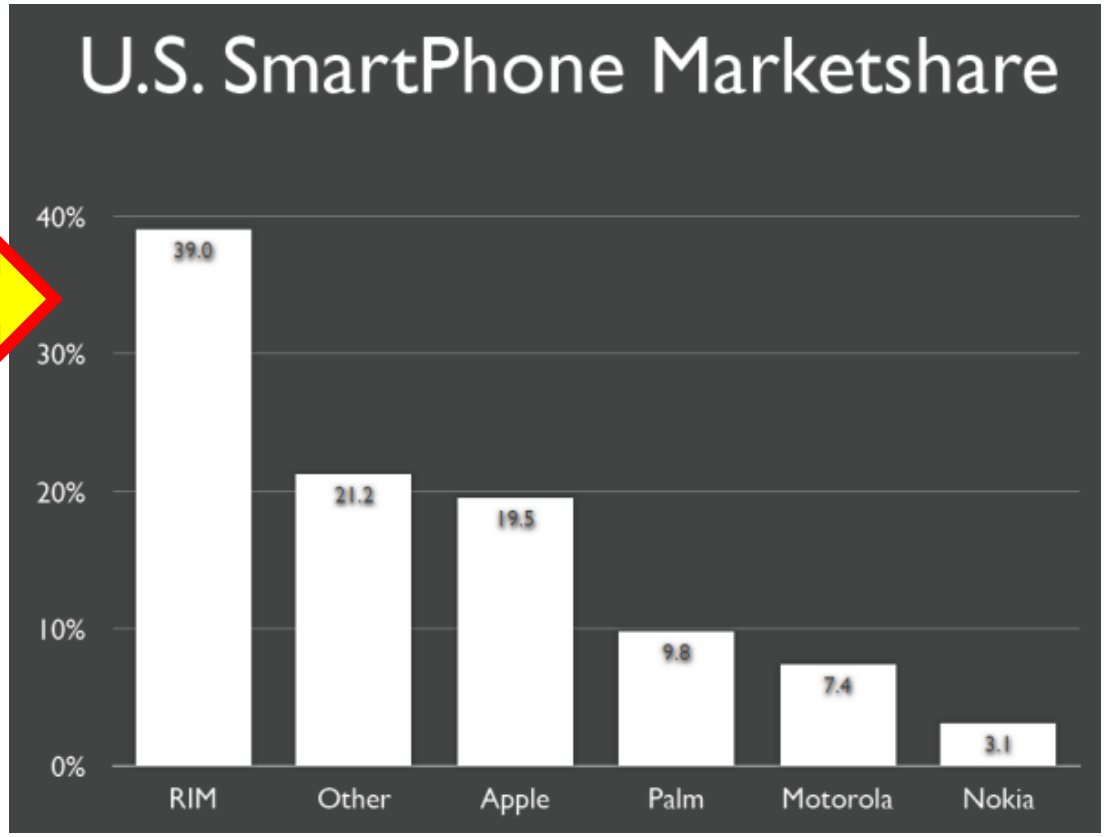
Pie-Charts: Dubious Message



What "trick" is Steve Jobs using here to convey what message?
What is the weakness of a pie chart in this instance at communicating accurate data?

Pie-Charts: Dubious Message



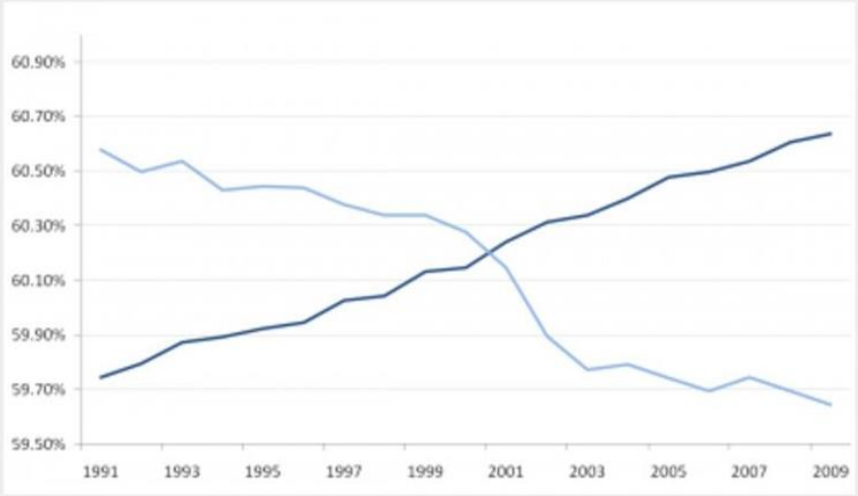
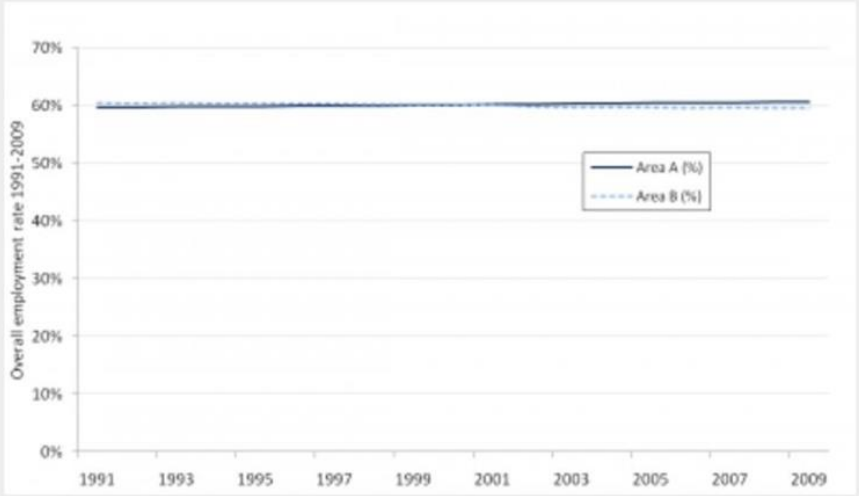


- The pie chart distorts the data in Steve Jobs' favour
- The pie chart makes it difficult to **visually** compare--in an instant--the relative rankings of Smartphone market share

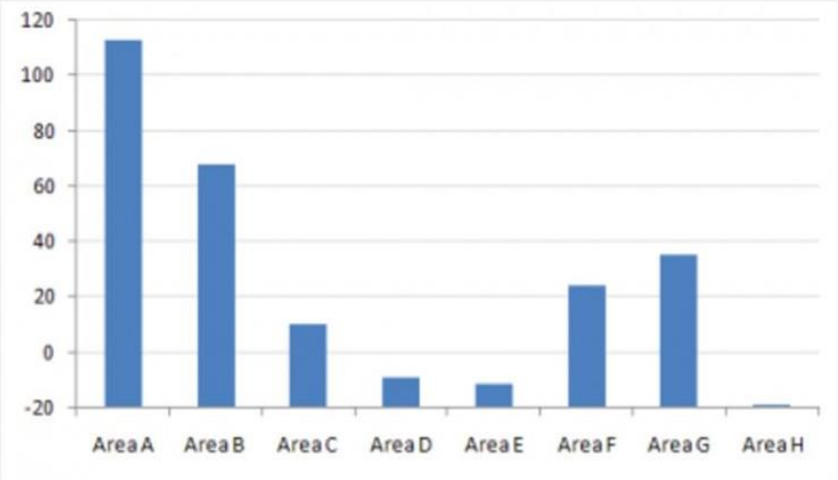
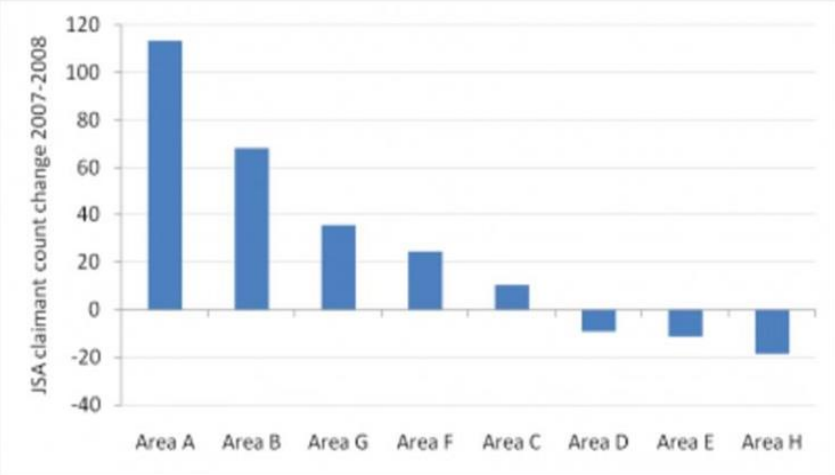
Accurately Represent the Data

- Keep the zero on the axis scale*
- For bar charts, set the base of the bars to zero (not the lowest value)*
- Avoid varying the size of objects in graphs, except to convey difference in values
- Avoid using line charts where data is only available for a small number of time points*

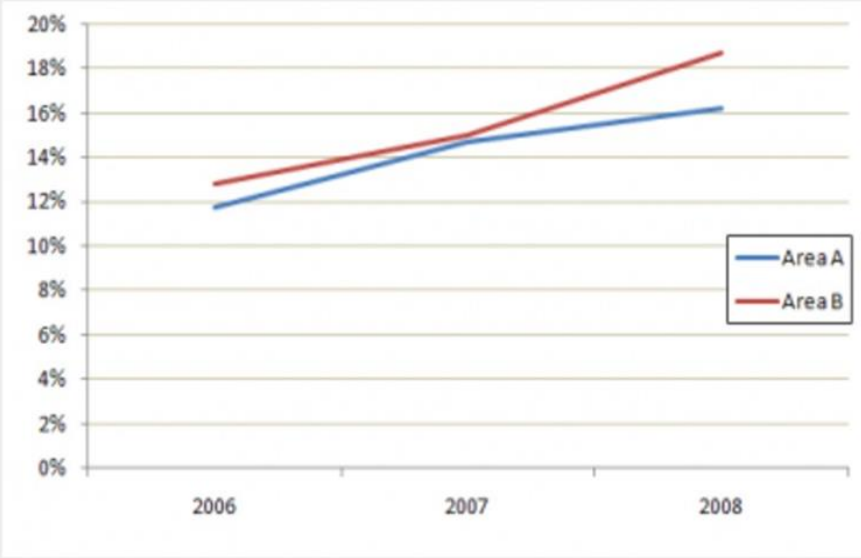
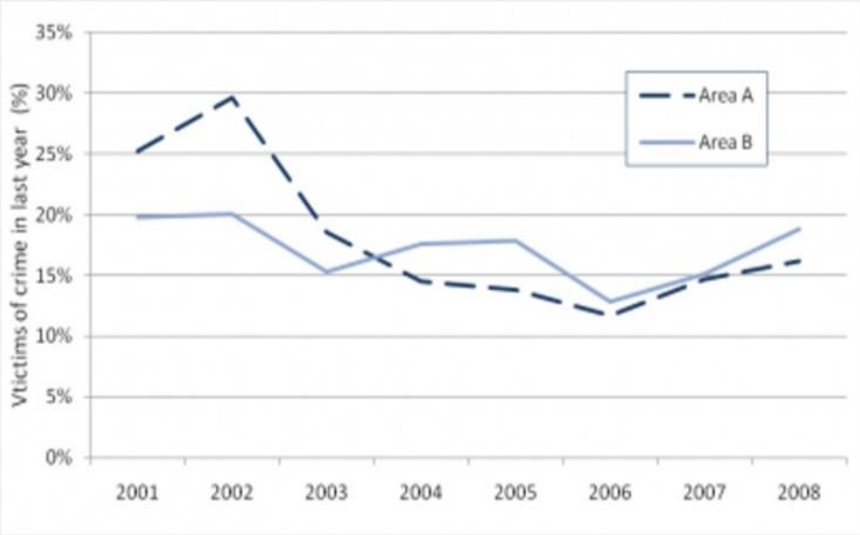
Keep the zero on the axis scale

Before	*After*
	

Set the base of the bars to zero

Before	*After*
	

Avoid line charts if only few time points available

Before	*After*
	

Keep it clear

- Avoid using visualisation effects such as 3D that can hide the data*
- When choosing colours to use, limit the number of colours used and ensure that different colours can be distinguished.
- Where colour is needed, use solid blocks of colour and avoid fill patterns
- Avoid using strong or bold colours for the background in a visualisation*
- When creating choropleth maps, choose colours to help users identify patterns and relationships between areas

Using Colors

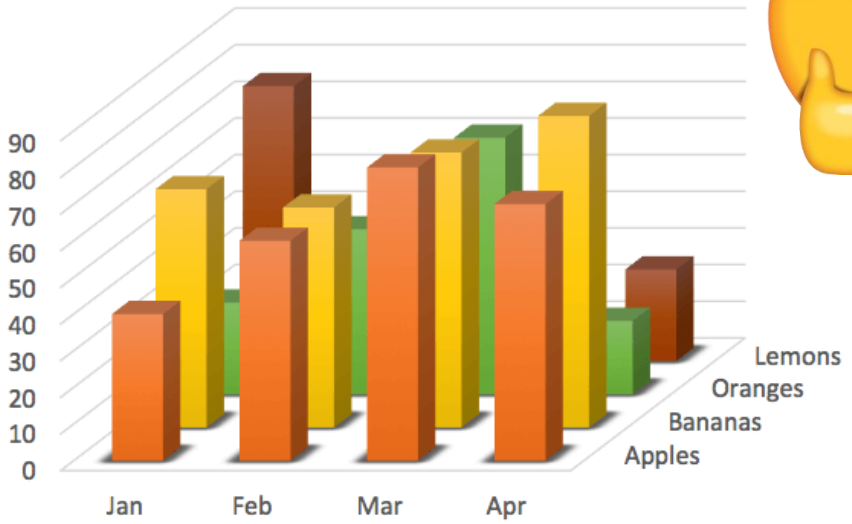
(Grey)Value is perceived as ordered (O)

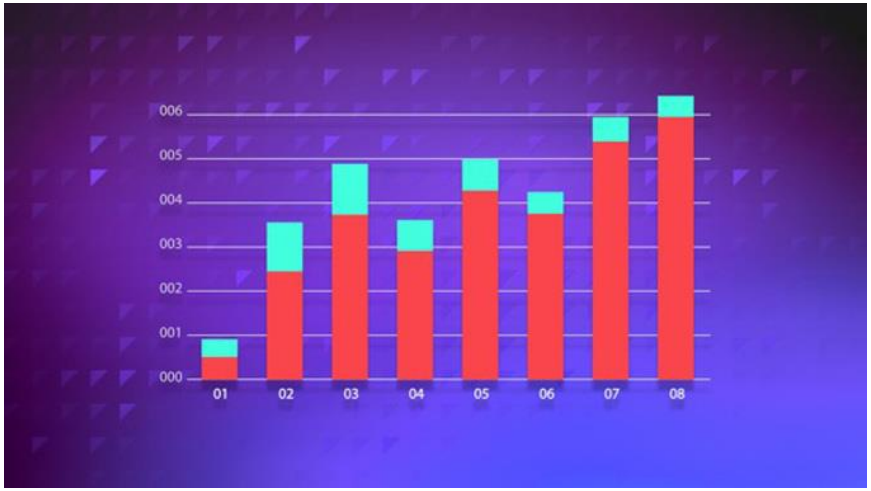
Can encode quantitative values (Q) [not as well]

Hue is normally perceived as unordered (N)

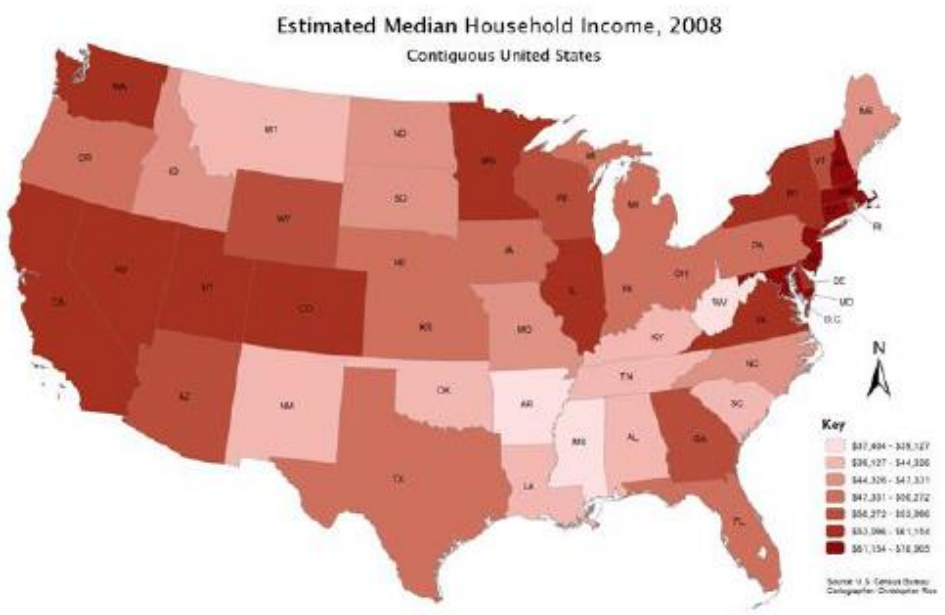
- **Ordinal data**: Use colors of different brightness (intensity) or saturation (shade).
- **Nominal data**: Use different hues since ordering is not important. Different hues will set apart each data.

Avoid Visualization Effects





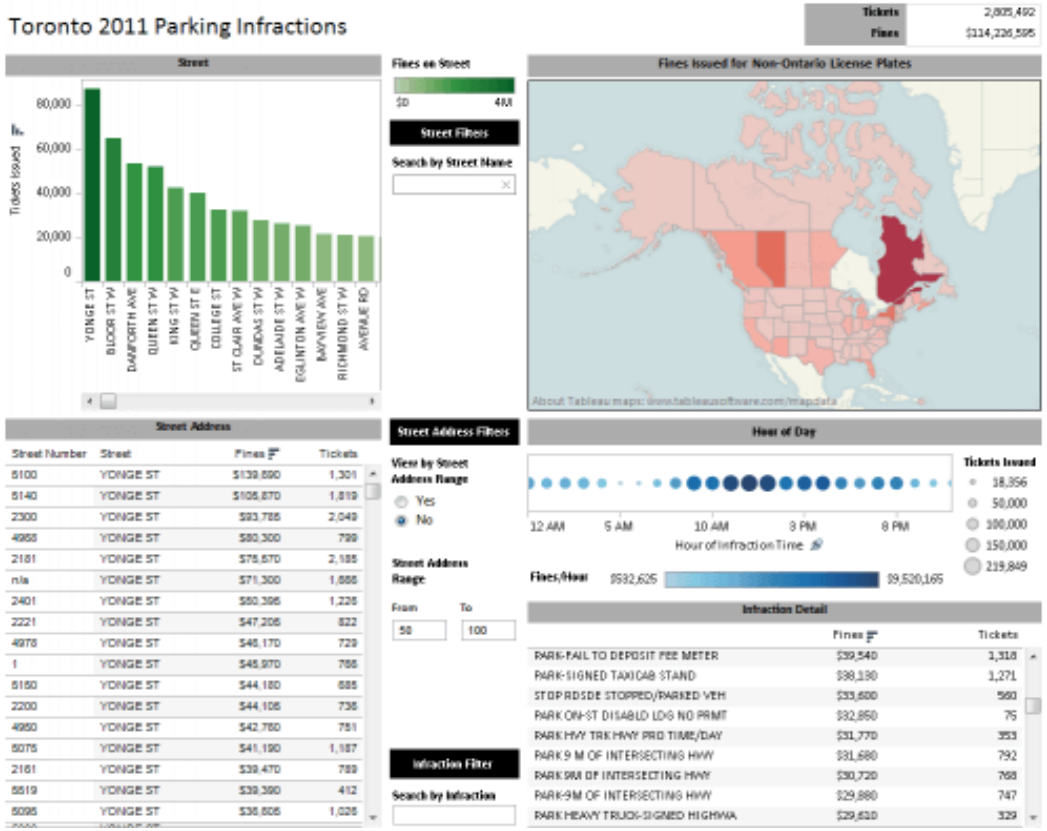
Choropleth Map



- A **choropleth map** is a thematic map where areas are coloured/shaded/patterned in proportion to the measurement of a variable displayed

- Crucial when creating: Choose a good range of colours to help users identify patterns and relationship between areas
- Values of the variable of interest should also be normalized

Dashboards



- "A dashboard is a visual display of the most important information needed to achieve one or more objective; consolidated and arranged on a single screen so the information can be monitored at a glance." – Stephen Few, Information Dashboard Design (2006)

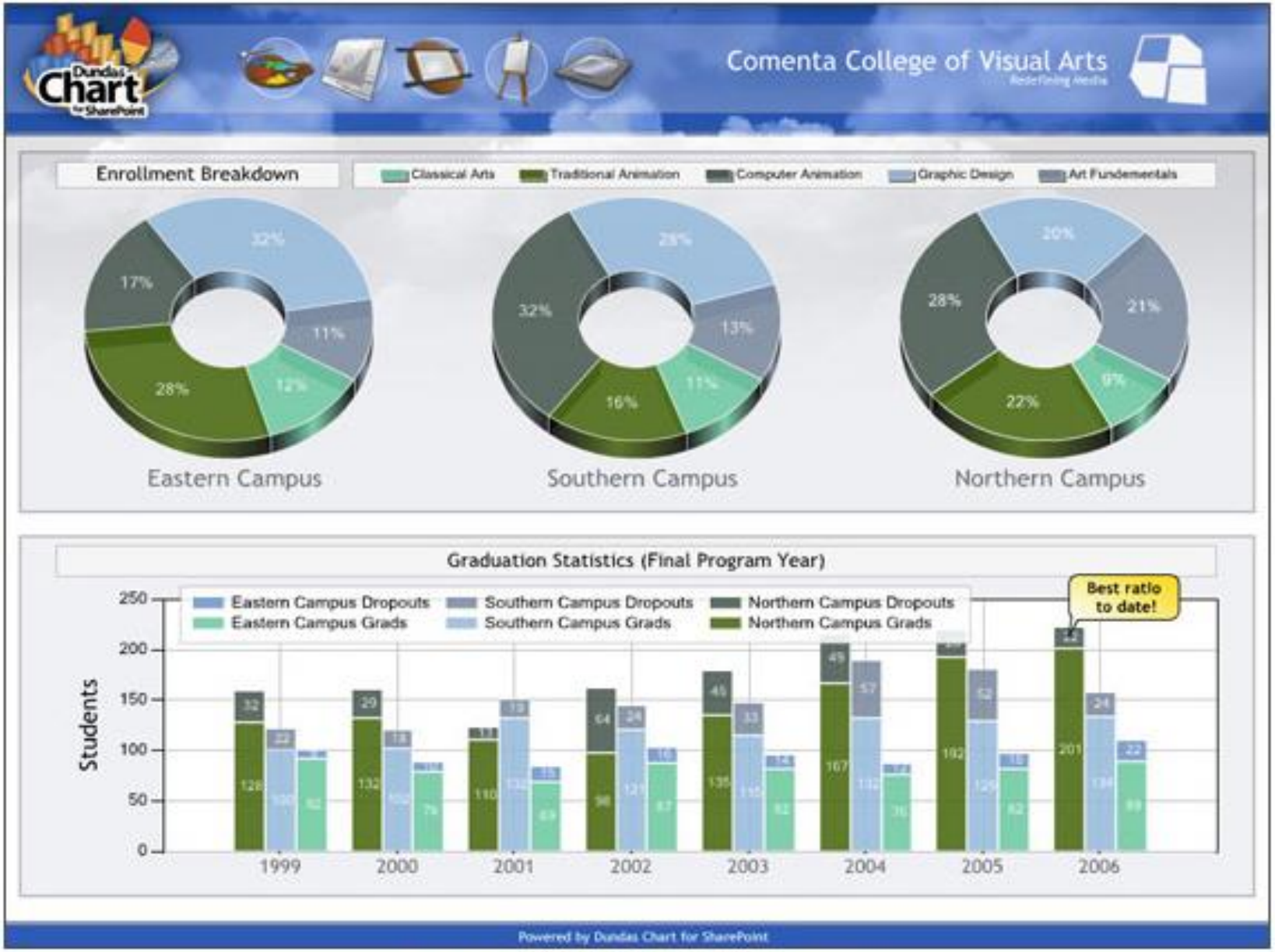
Interactive Visualization: http://www.opendataexplorer.ca/toronto-parking-tickets-study.html

Source: https://unilytics.com/how-to-design-and-build-effective-dashboards-in-tableau

Dashboards Design Issues

- They don't say enough
- What they do say is not said well.
- Lack context
- When monitoring what's going on, people usually need answers to these questions:
  - Are we doing well or poorly?
  - How well or how poorly?
  - What has led to what's happening today?

Source: https://www.matillion.com/insights/dashboard-examples-the-good-the-bad-and-the-ugly/



Digital Dashboard, Education Metrics

Source: https://www.matillion.com/insights/dashboard-examples-the-good-the-bad-and-the-ugly/

Law of Simplicity

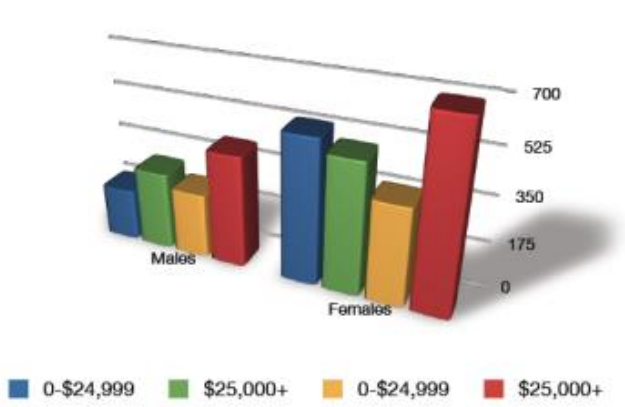
- John Maeda, in The Laws of Simplicity, offers a maxim about design simplicity, which Stephen Few (2008) has massaged into the following statement:

  "Simplicity is about eliminating the obvious (and everything else that doesn't support your purpose), and enhancing the meaningful."

Maximize Data-Ink Ratio

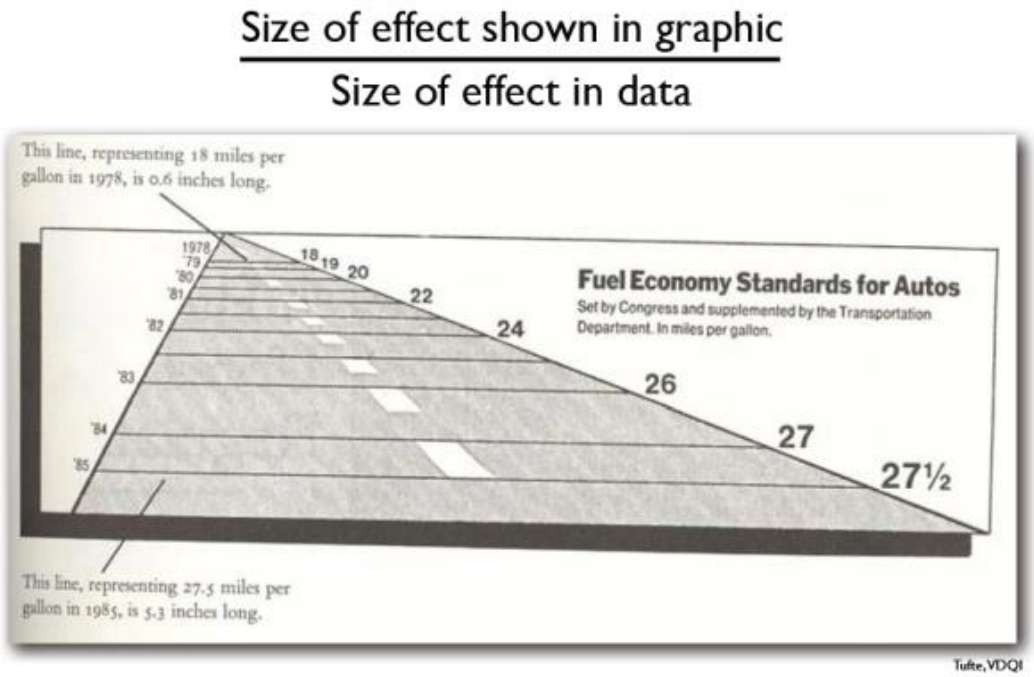
- One of Edward Tufte's Principles

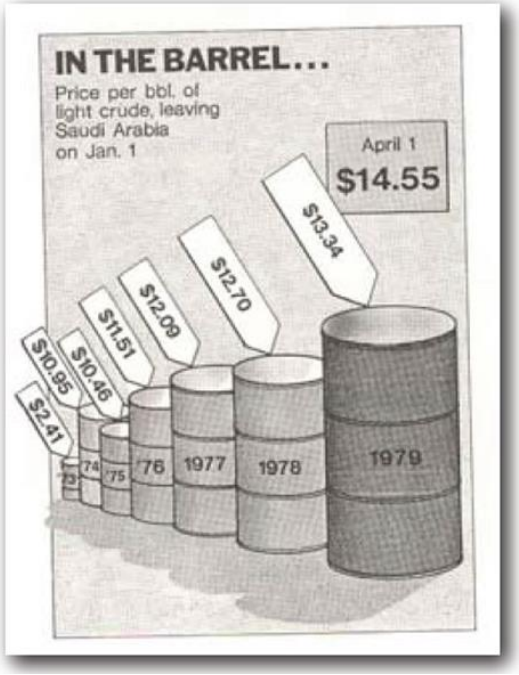
$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



Lie Factor

- Distorting the data with graphics that aren't accurate or representative of scale





Reading Material

- Edward Tufte, "The Visual Display of Quantitative Information", 1983 (original), 2001 reprint.
- Stephen Few, "Show me the numbers: Designing tables and graphs to enlighten", 2005, (slides!)
- Bill Shander, "How to tell stories and weave cohesive narrative with data"
- Unilytics, "Top 7 Tips for Dashboard Visualization", 2015.
- Tableau, "Data is beautiful: 10 of the best data visualization examples from history to today"