CSE 101: Computer Science Principles

Module 22: Principles of Data Visualization

Acknowledgements

- These slides are revised versions from Prof. Kevin McDonnell at Stony Brook University

Exploratory Data Analysis

- *Looking* carefully at your data is important:
  - to identify mistakes in data collection/processing
  - to find violations of statistical assumptions (e.g., that the data follows a **normal distribution**)
  - to observe patterns in the data to make hypotheses
- **Data visualization** is simply the representation of information through graphical means (e.g., charts)

Tabular Data

- Tables can have advantages over plots:
  - Representation of numerical precision
  - Understandable multivariate visualization: each column is a different dimension (e.g., pandas dataframes when displayed "raw")
  - Representation of **heterogeneous** data
  - Compactness for small numbers of points

Edward Tufte

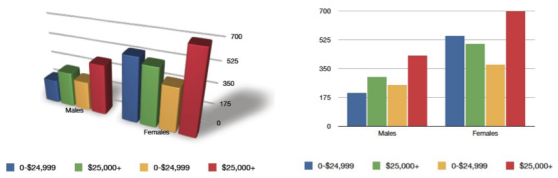
- American statistician and pioneer in data visualization
- Developed simple, but effective, principles for *quantifying* what is visually good or bad with charts and graphs
- Some of the graphics in these notes are from Tufte's books

Tufte's Visualization Aesthetic

- Distinguishing good/bad visualizations requires a **design aesthetic** (a philosophy of what makes a design beautiful or artful), and a vocabulary to talk about visualizations
  - Maximize the **data ink-ratio**
  - Strive for a **lie factor** of 1.0
  - Eliminate **chartjunk**
  - Use proper scales and clear labeling

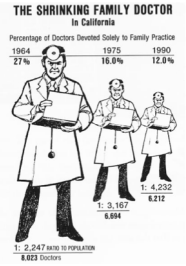
Maximize the Data-Ink Ratio

- Data-ink Ratio = Data Ink / Total Ink Used in Graphic
- Basic idea: make the chart visually simple



Lie Factor

- Lie Factor = $\frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$
- Basic idea: the *change* in graphic size should match the *change* in data
- Want a Lie Factor = 1.0
- The 2D doctor graphics depict a change in a 1D value
- Lie factor = 2.8



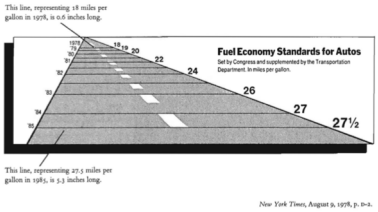
Example: Lie Factor Calculation

- Using Tufte's lie factor, calculate the lie factor for a graph that represents 20 units with a visual length of 1 centimeter, and 30 units with a visual length of 5 centimeters.

- Size of effect in graphic = (5 - 1) / 1 = 4
- Size of effect in data = (30 - 20) / 20 = 10 / 20 = 0.5
- Lie factor = 4 / 0.5 = 8

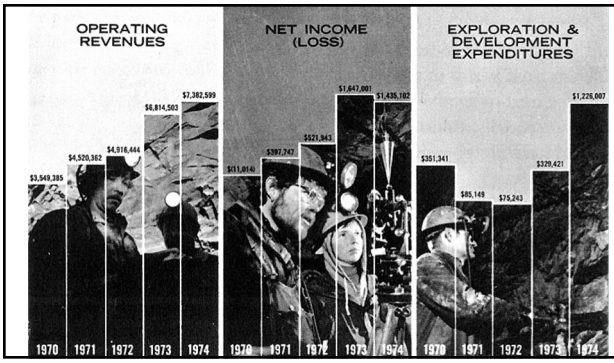
Lie Factor

- Lie factor = 14.8
- Use of a 3D effect to depict a change in a 1D value



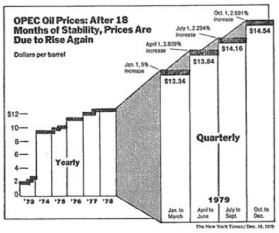
Graphical Integrity: Scale Distortion

- Always start bar graphs at zero
- Always properly label your axes
- The bar charts on the following slide all have different baselines
  - The middle one's baseline is -$4,200,000!



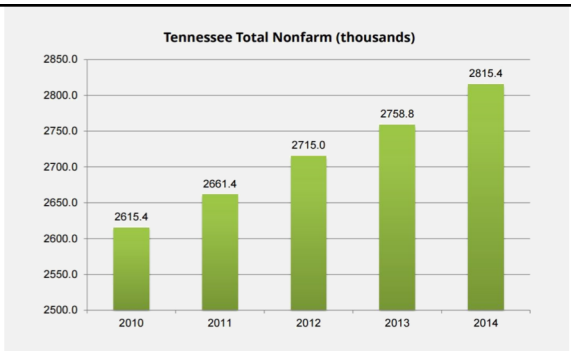
Graphical Integrity: Scale Distortion

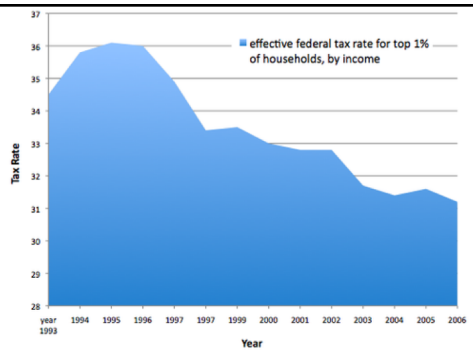
- The 3D effect and a change of time scale (quarterly to yearly) renders this graphic largely useless



The Principle of Proportional Ink

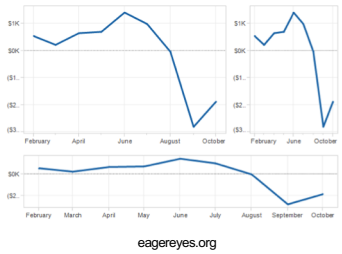
- The principle of proportional ink: the amount of ink used to indicate a value should be proportional to the value itself (Bergstrom and West)
- Suppose in a bar chart, one bar is twice the length of another. We expect that the second bar represents a quantity that is twice as great as the second bar.
- But this is true only when the baseline is zero!





Graphical Integrity: Aspect Ratios

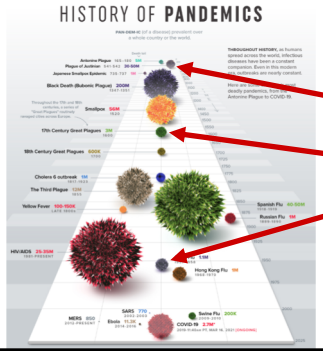
- The steepness of apparent cliffs is a function of **aspect ratio** (width:height)
- Aim for 45° lines or the Golden ratio (~16:10) as most interpretable



eagereyes.org

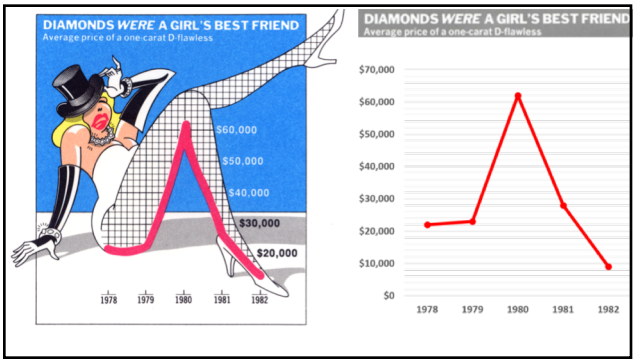
Reduce Chartjunk

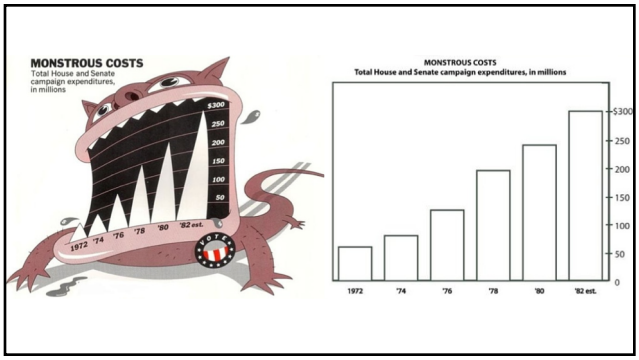
- Extraneous visual elements distract from the message the data is trying to tell
  - Extra dimensionality (e.g., using 3D when 1D or 2D get the job done)
  - Uninformative coloring
  - Excessive grids and figurative decoration
- In an exciting graphic, the data tells the story, not the chartjunk

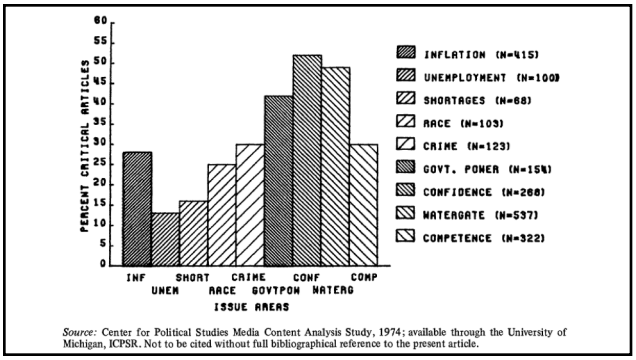


Chartjunk + Lies

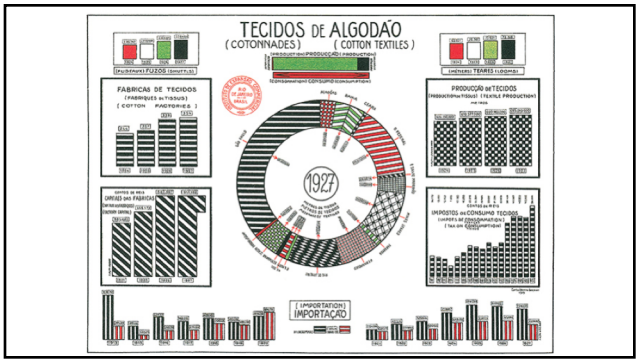
Plague of Justinian (30-50M dead), *17th Century Great Plagues* (3M dead) and *Asian Flu* (1.1M dead) shown as the approximate same size due to 3D effect. (Source)

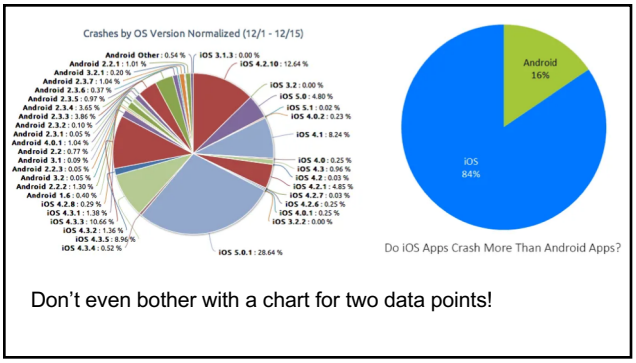






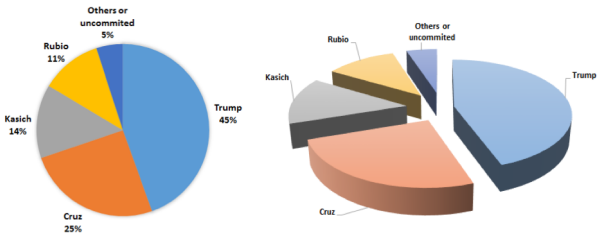
Source: Center for Political Studies Media Content Analysis Study, 1974; available through the University of Michigan, ICPSR. Not to be cited without full bibliographical reference to the present article.

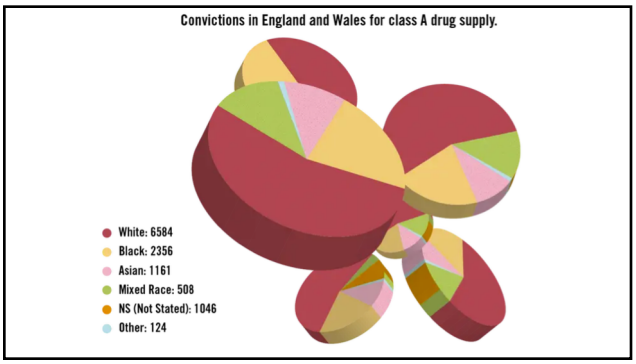


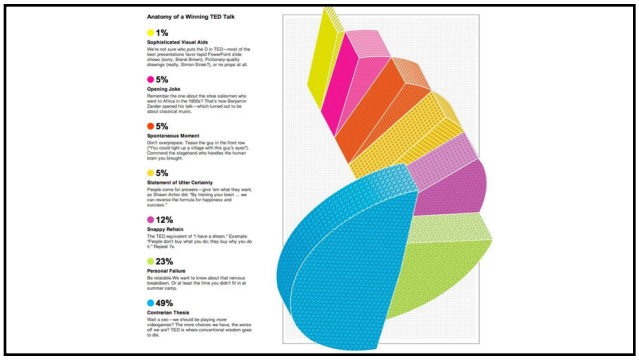


Don't even bother with a chart for two data points!

Which Pie Chart is Better?

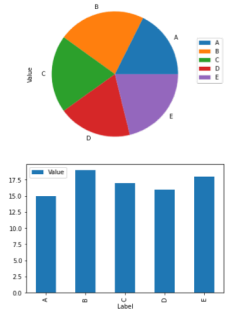


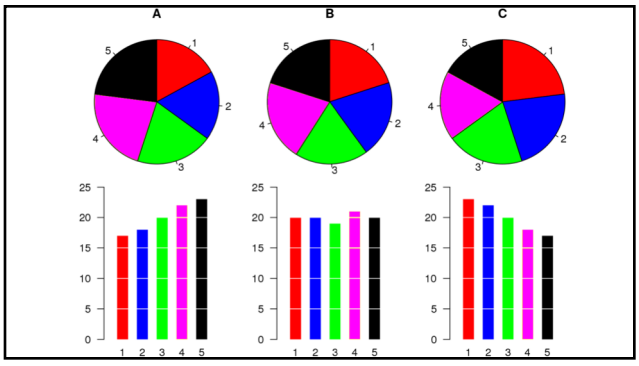




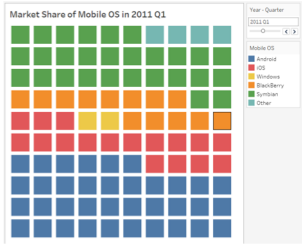
Pie Charts vs. Bar Charts

- Bar charts show the frequency or proportion of items in categories.
- Pie charts use more space and are harder to read and compare.
- The advantages of bar charts are even more apparent for 2D data (next slide).





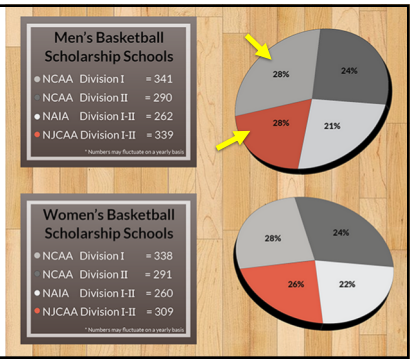
Waffle Charts: Better for Percentages



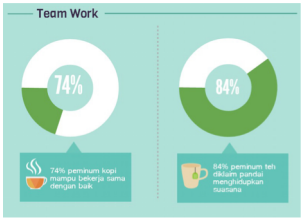
Also called a “square pie chart”.

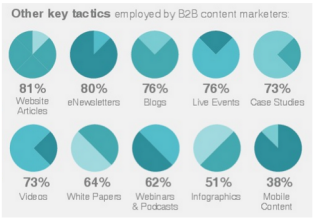
Oval Pie Charts?

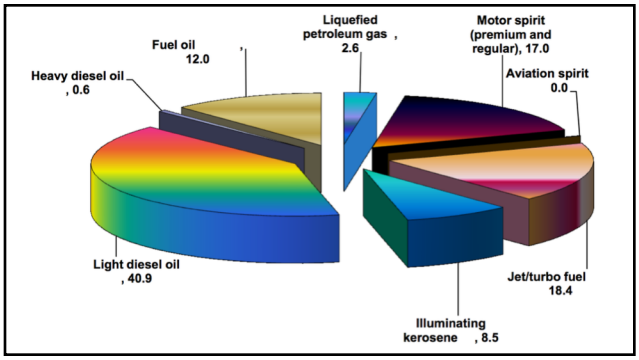
Note how the distortion interferes with interpretation of the data (gray 28% > orange 28%?)

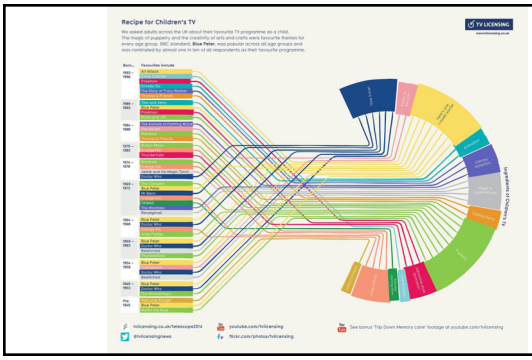


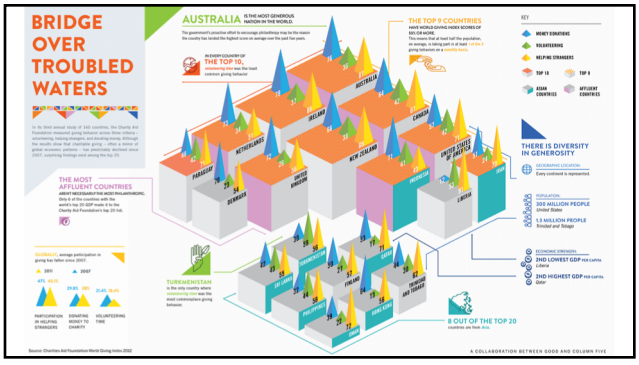
Inconsistent Coloring



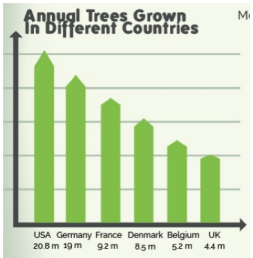


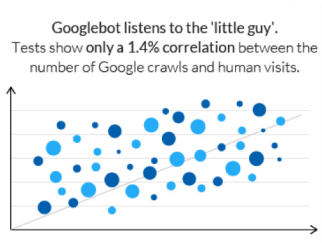




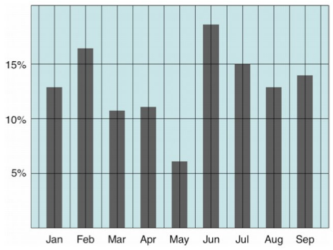


Meaningless Charts

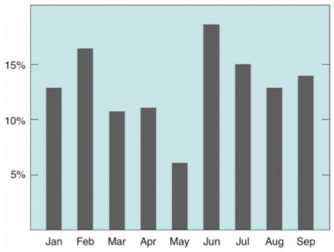




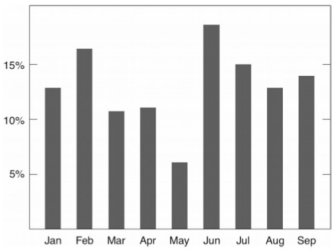
Can You Simplify this Plot?



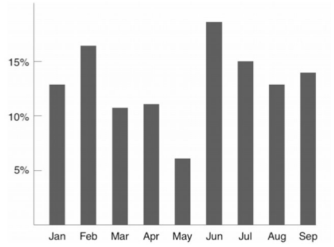
Can You Simplify this Plot Further?



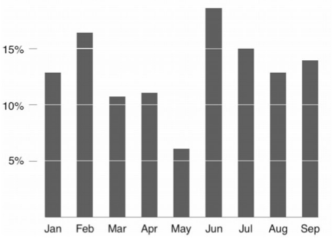
Can You Simplify this Plot Further?



Can You Simplify this Plot Further?



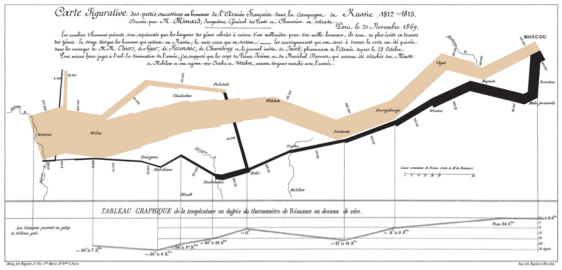
Less is More!



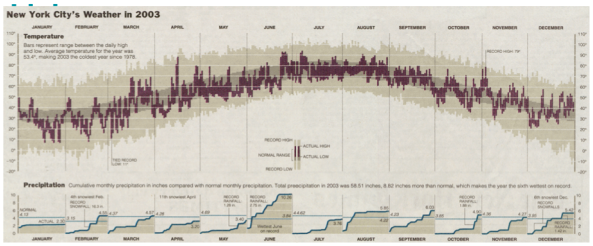
Great Data Visualizations

- Display data accurately and clearly
- Tell a story that the data reveals
- Are rich enough to make you want to look carefully and study the data
- On the following few slides are some of Tufte's favorite visualizations

Napoleon's Advance and Retreat Link

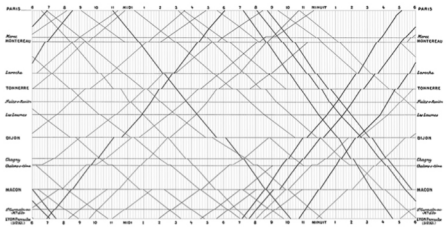


New York's Weather Year in Review



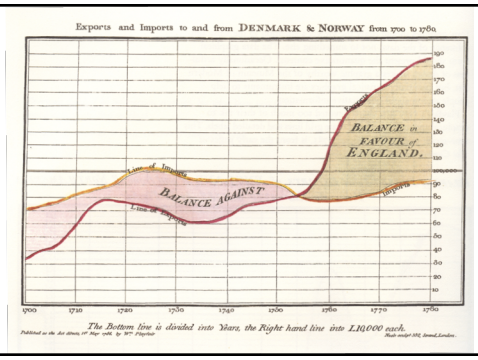
Check out weatherspark.com for an interactive take on this kind of graph.

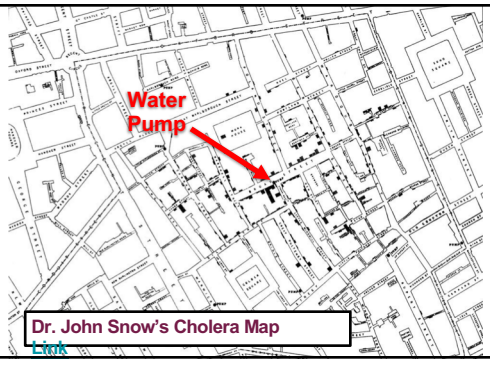
Marey's Train Schedule Link



What can you see here that you cannot with normal train schedules?

The original image had a darker datagrid, making it harder to read.





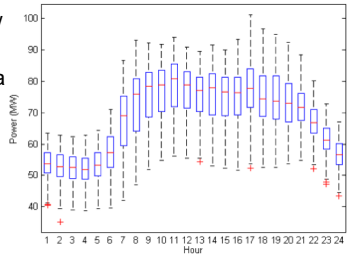
Link

Gapminder: Interactive Visualizations Link

- Gapminder is a website with all sorts of fascinating data about the state of the world
  - Population, health, economy, education, others
- Number of people by income Income per person
- Population by country
- Life expectancy vs. income A **bubble chart** uses color, shape, size, and shading of "dots" enables dot plots to represent additional dimensions.

Creating Effective Visualizations

Box plots concisely show the range and quartiles, median and variance of a distribution

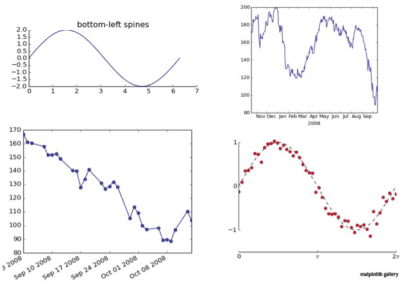


Stacked Areas vs. Line Plots

Note how trends inside the middle of the stack are easier to see in the line plots (e.g., the light blue line)

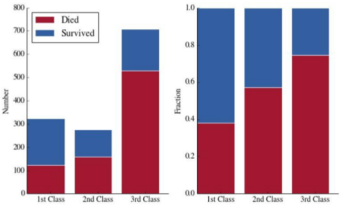
Line Charts

- Show data points, not just fits
- Line segments show connections, so do not use in categorical data
- Connecting points by lines is often chartjunk



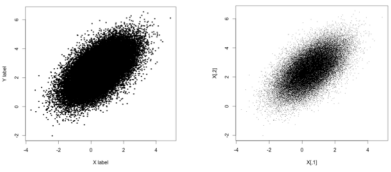
Creating Effective Visualizations

- Sinking of *The Titanic*
- Looking at proportions instead of absolute numbers reveals the underlying story more clearly



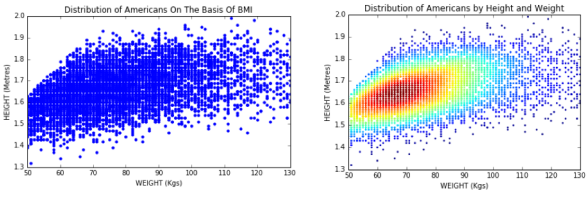
Scatter Plots / Multivariate Data

- Scatter plots show the values of each point, and are a great way to present 2D data sets
- Reduce overplotting by using smaller points



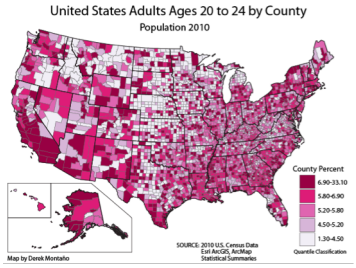
Heatmaps Reveal Finer Structure

Color points on the basis of frequency



Creating Effective Visualizations

A **choropleth map** colors a region in a way proportional to some underlying value or statistic



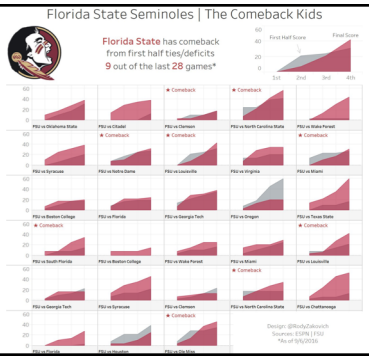
Creating Effective Visualizations

Link

Small multiples can be effective for showing multivariate data

How many different variables are visualized here?

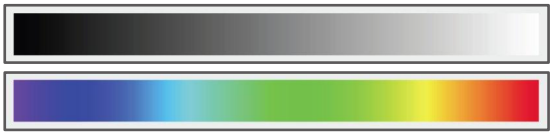
Link



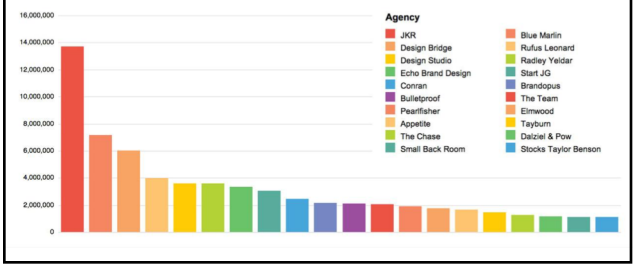
Understanding Color Scales

Rainbow Color Maps

- Rainbows are perceptually non-linear
- Use grayscale gradients or color gradients with 2 or 3 colors maximum



Dramatic Misuse/Reuse of Color



Keep a Critical Eye

- Remember Tufte's principles whenever designing or interpreting data visualizations:
  - Maximize the data-ink ratio
  - Seek a lie factor of 1.0
  - Minimize chartjunk
- Use proper scales and clear labeling
- Beautiful data deserves beautiful visualization