BIG DATA ANALYTICS
Data Visualization

Olga Klopp
klopp@essec.edu

Outline

Examples

Principles for scientific visualization

Basic visualization techniques

Distributions for a single variable

Relationships between two variables

Two categorical variables

Big Data Visualization

Problems

Approaches

Methods

Visualization

- If a picture is worth a thousand words, a data visualization is worth at least a million.
- Data visualization is one of the most impactful ways that data analysts and scientists can communicate their findings.
- Data visualizations manipulate complex pools of data to visually display the datas patterns, trends and correlations.

Outline

Simpsons Paradox:

- The Visualizing Urban Data Idealab (VUDlab) from the University of California-Berkeley
- Visual look at data that disproves the claim in a 1973 suit that charged the school with sex discrimination.

Simpsons Paradox:

- In 1973, the University of California-Berkeley was sued for sex discrimination
- The numbers looked pretty incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants
- When researchers looked at the evidence, though, they uncovered something surprising:

  If the data are properly pooled ... there is a small but statistically significant bias in favor of women

- It was a textbook case of **Simpson's paradox**

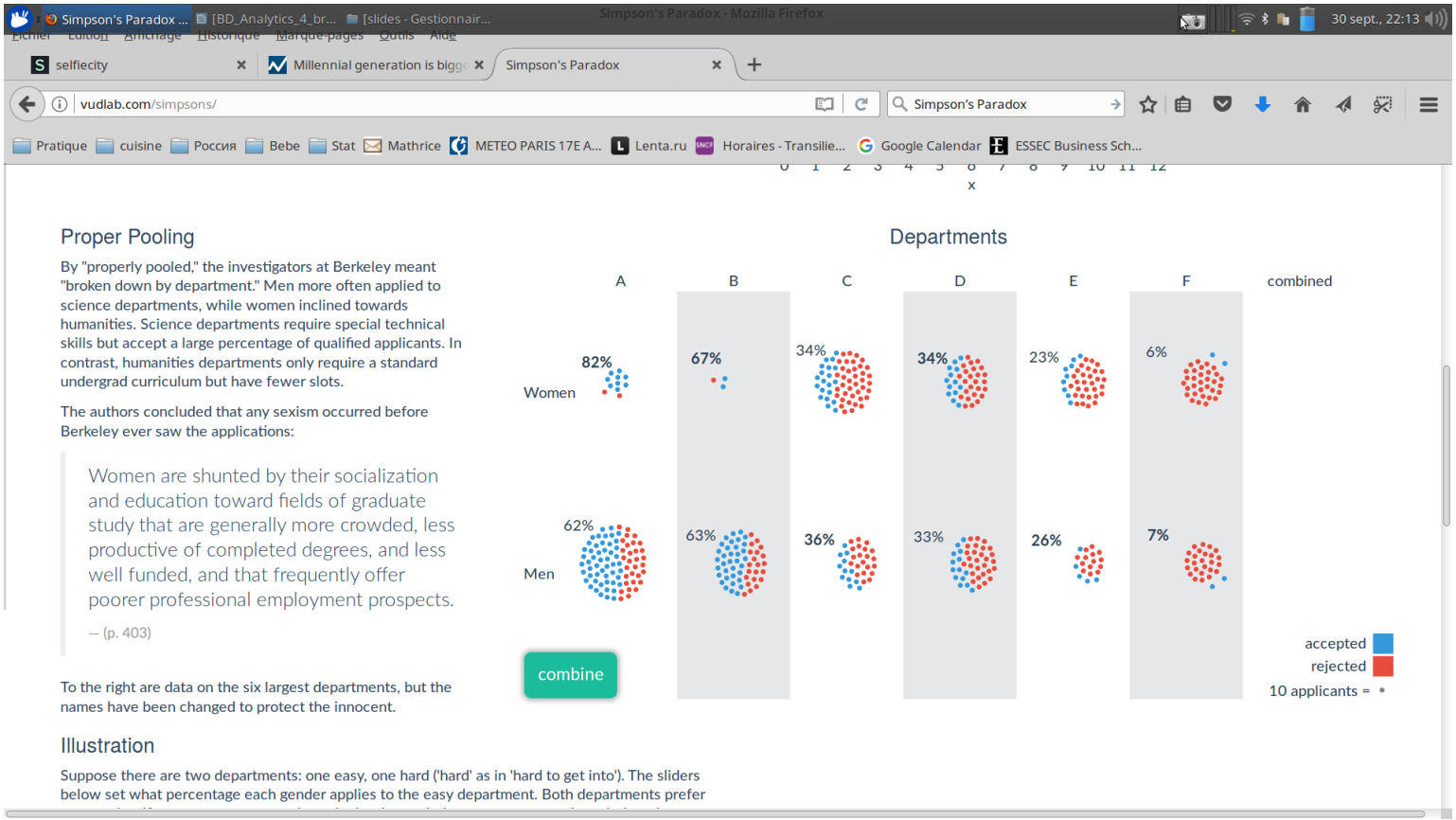
What is Simpson's paradox?

- Simpson's paradox involves at least three variables:
  1. the explained
  2. the observed explanatory
  3. the lurking explanatory
- If the effect of the observed explanatory variable on the explained variable changes directions when you account for the lurking explanatory variable, you've got a Simpson's Paradox.

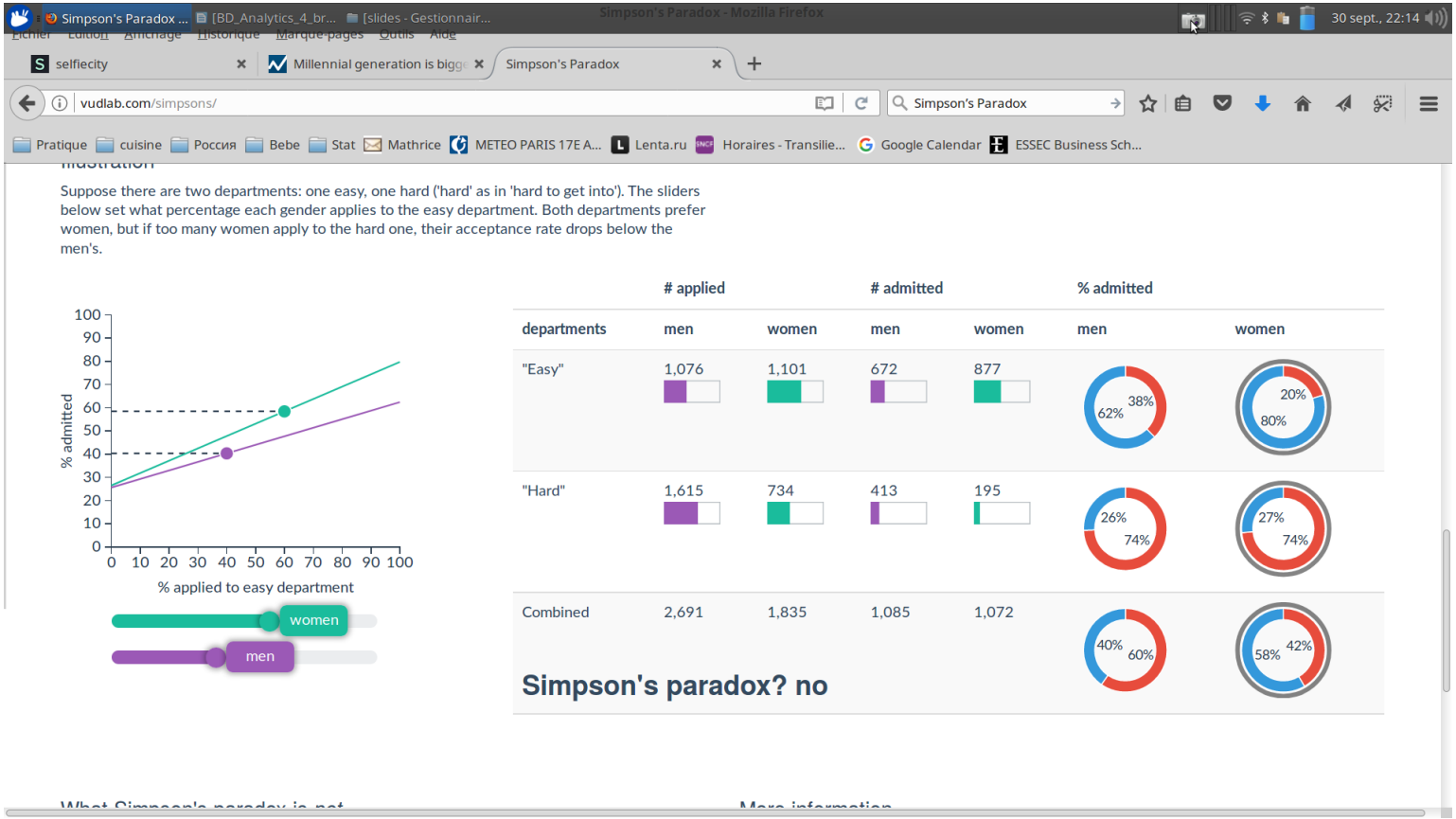
Proper Pooling

- By "properly pooled", the investigators at Berkeley meant "broken down by department":
  - Men more often applied to science departments, while women inclined towards humanities
  - Science departments require special technical skills but accept a large percentage of qualified applicants
  - Humanities departments only require a standard undergrad curriculum but have fewer slots
- The authors concluded that any sexism occurred before Berkeley ever saw the applications

Data on the six largest departments:



Suppose there are two departments: one easy, one hard ('hard' as in 'hard to get into'):

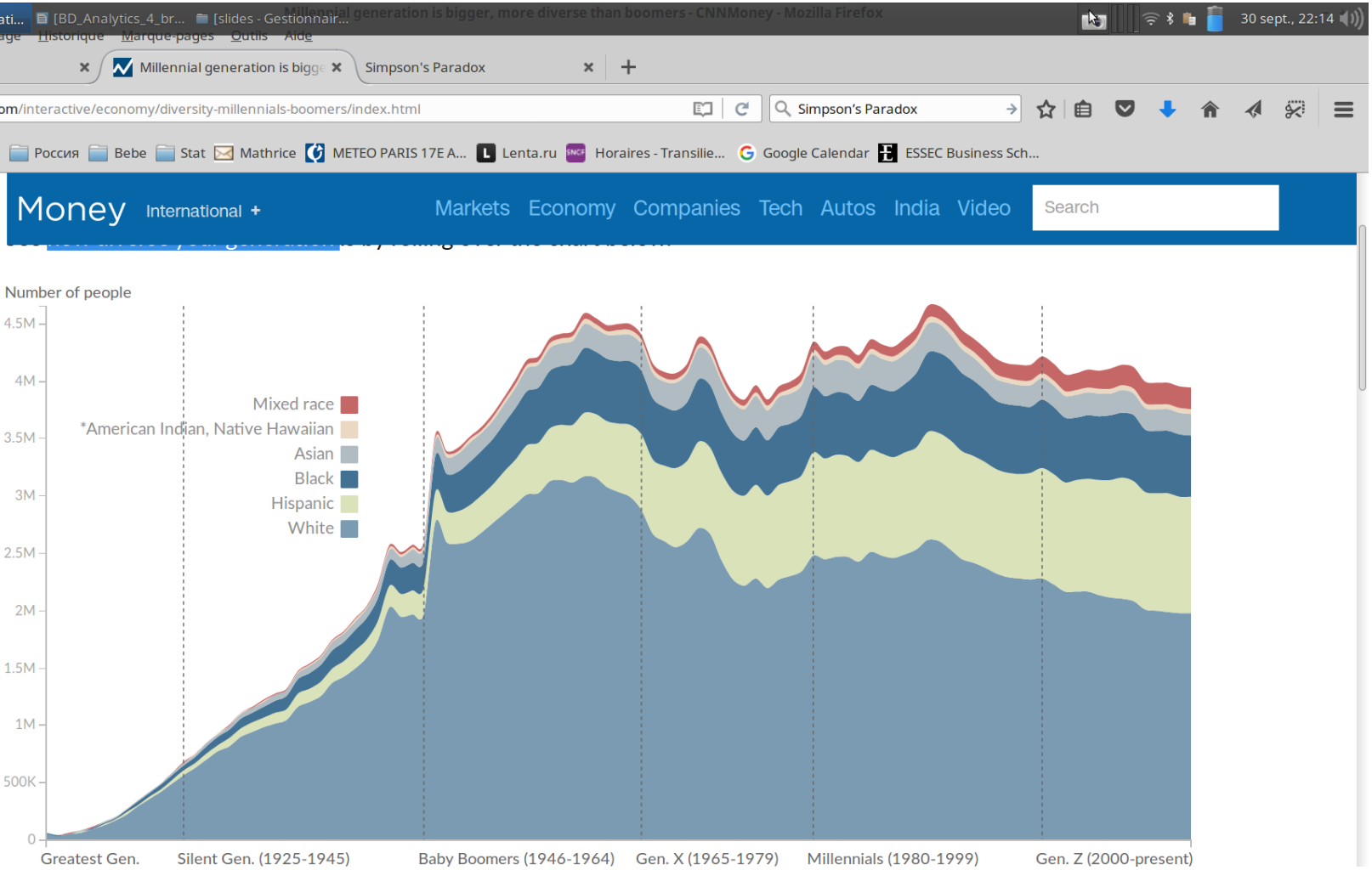


Both departments prefer women, but if too many women apply to the hard one, their acceptance rate drops below the men's.

Millennial Generation Diversity

- CNN Moneys interactive chart shows the size and diversity of generations in U.S.
- It was built using U.S. Census Data
- Illustrates the racial makeup of different age groups from 1913 to present

Millennial Generation Diversity

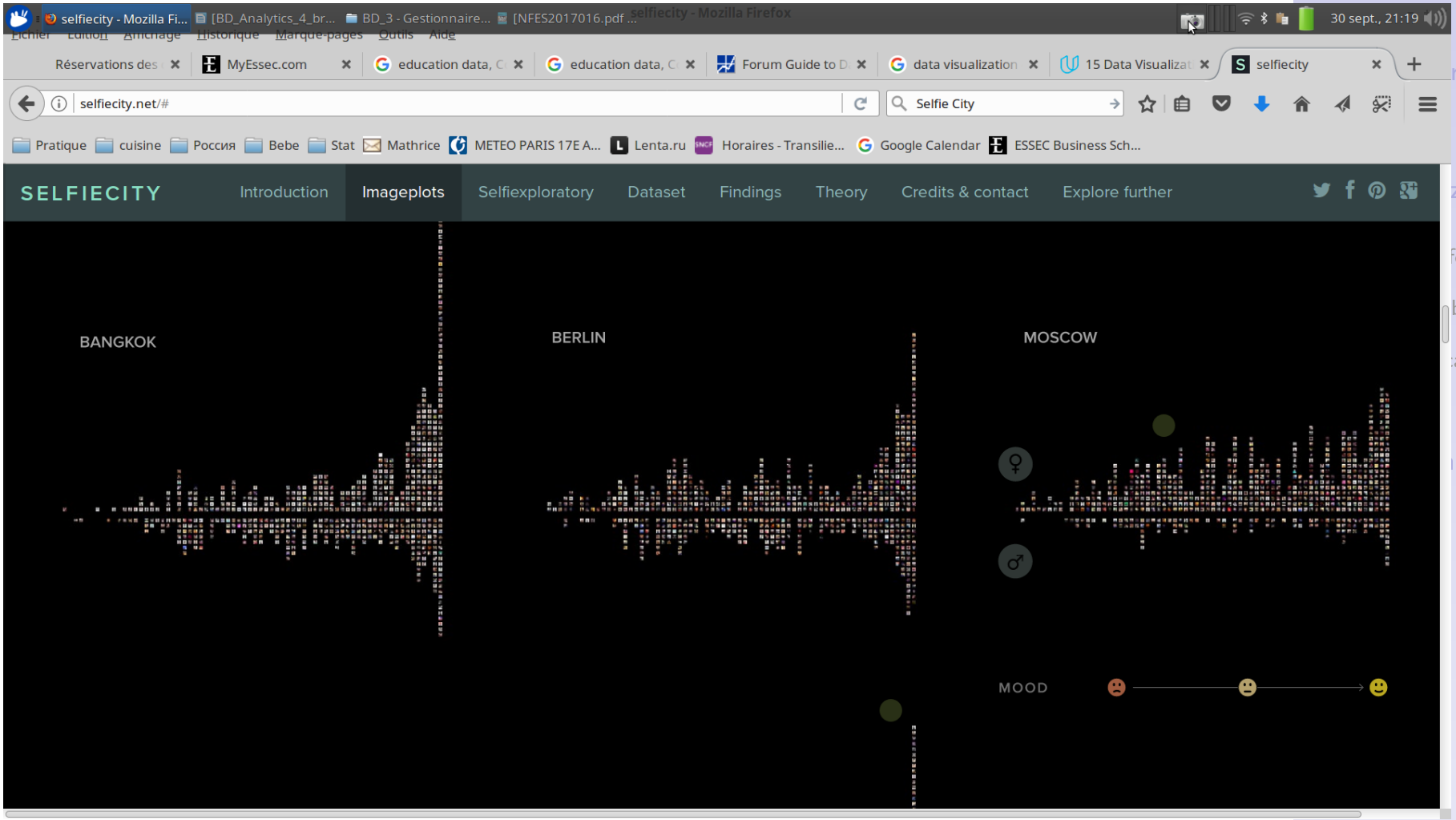


You can move your mouse over the graph to explore the stats of each year

Selfie City

- Multi-component visual exploration of selfies from five major cities around the world
- A close look at the demographics and trends of selfies
- The team behind the project collected and filtered the data using Instagram and Mechanical Turk.

Selfie City



Differences between selfies snapped in, say, New York and Berlin, as well as those between men and women across the world.

Outline

Principles for scientific visualization

- Visualization is the first formal step of the analysis
- Often a tool of choice for data cleaning
- The aim is to find efficient graphical representations that summarize the data and emphasize its main characteristics
- Can also serve as an inferential tool in different stages of the analysis

Understanding the patterns at the data helps creating good models

Principles for scientific visualization

Principles for scientific visualization formulated by William Cleveland:

- A graphic should display as much information as it can, with the lowest possible cognitive effort to the viewer
- Strive for clarity. Make the data stand out:
  - Avoid too many superimposed elements, such as too many curves in the same graphing space
  - Find the right aspect ratio and scaling to properly bring out the details of the data
  - Avoid having the data all skewed to one side or the other of your graph
- Visualization is an iterative process. Its purpose is to answer questions about the data

Principles for scientific visualization

- E.g. a million points of data can be plotted in a graph and offer a view of the density of data
- But, plotting a million points on the graph may produce a blurred image which may hide rather than highlight the distinctions
- Binning or selecting the top few frequent categories may deliver greater insights.

Principles for scientific visualization

- During the visualization stage:
  - graph the data
  - learn what you can
  - then regraph the data to answer the questions that arise from your previous graphic
- Different graphics are best suited for answering different questions

Outline

Checking distributions for a single variable:

Visualization can help you answer questions like these:

- What is the peak value of the distribution?
- How many peaks are there in the distribution (unimodality versus bimodality)?
- How normal (or lognormal) is the data?
- How much does the data vary? Is it concentrated in a certain interval or in a certain category?

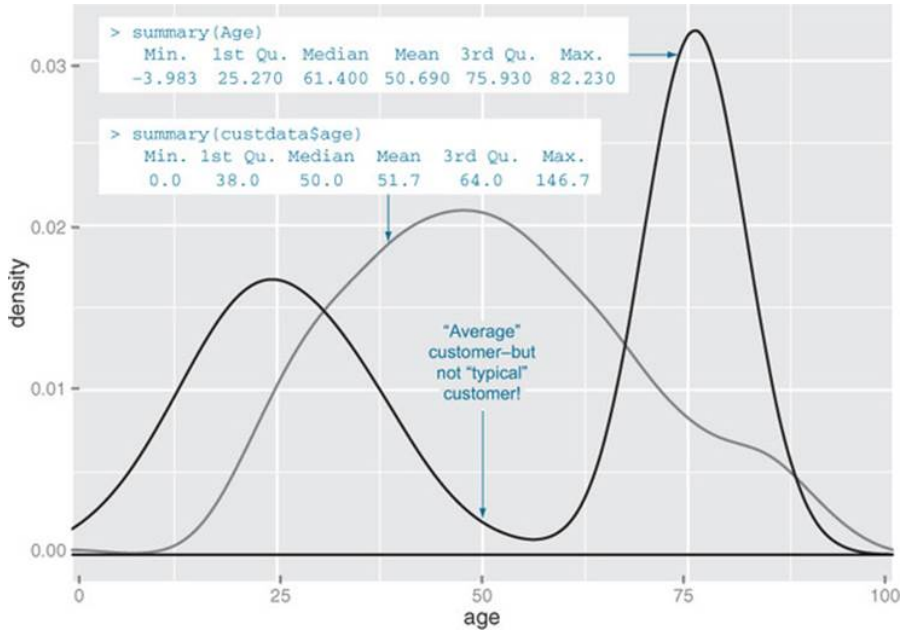
Shape of the data distribution

- One of the things that's easier to grasp visually is the shape of the data distribution:
  - many summary statistics assume that the data is approximately normal in distribution (at least for continuous variables)
  - verify whether this is the case.

Unimodal vs Multimodal

Check if your data is unimodal or multimodal:

- Unimodal distribution: one population of subjects
  - "typical" customer is middle-aged
- Multimodal: several populations
  - Two populations probably with very different behavior patterns



HISTOGRAMS

- Histogram bins a variable into fixed-width buckets and returns the number of data points that falls into each bucket
- For example, you could group your customers by age range, in intervals of five years: 20 - 25, 25 - 30, 30 - 35, and so on
- Disadvantage: you must decide ahead of time how wide the buckets are:
  - If too wide, you can lose information about the shape of the distribution
  - If too narrow, the histogram can look too noisy to read easily.

DENSITY PLOTS

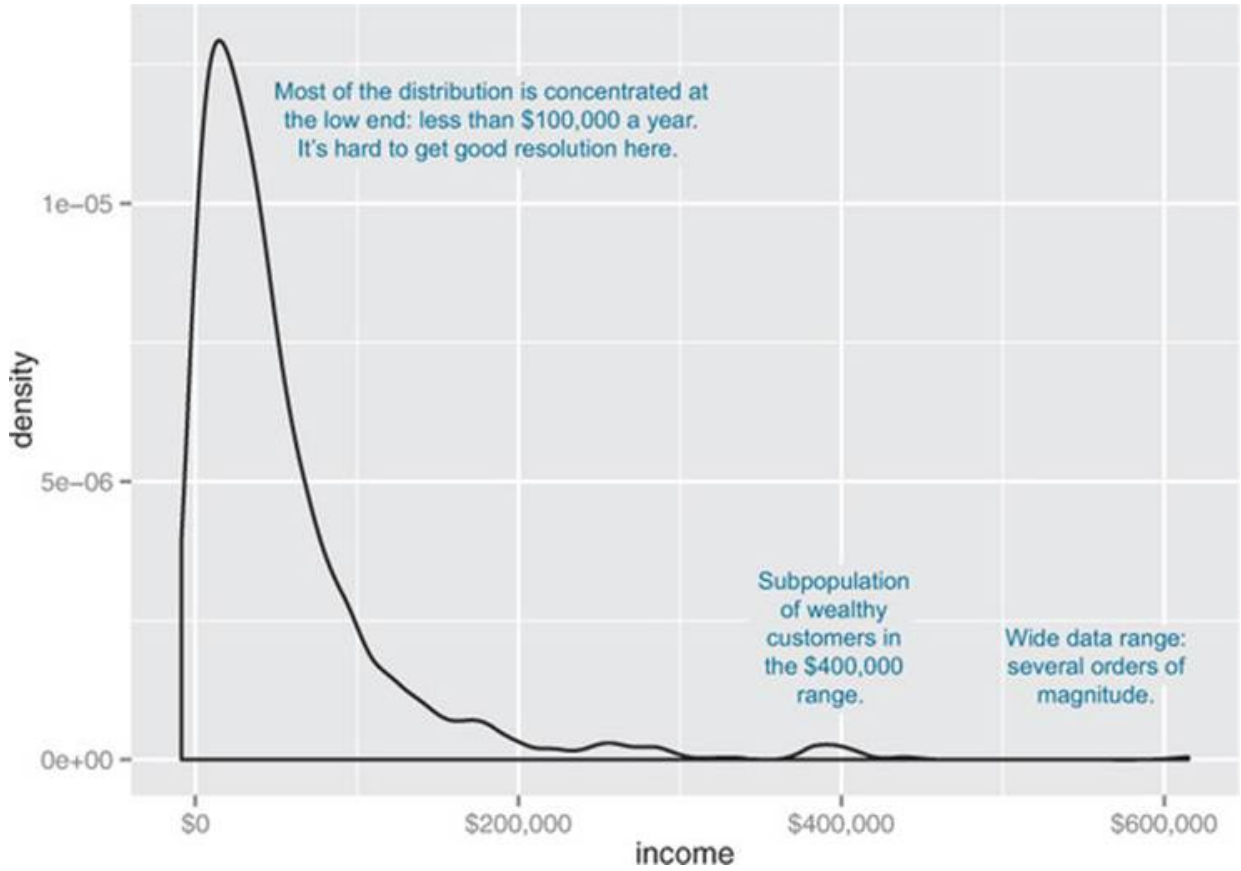
- Density Plot visualizes the distribution of data over a continuous interval
- It computes and draws kernel density estimate:

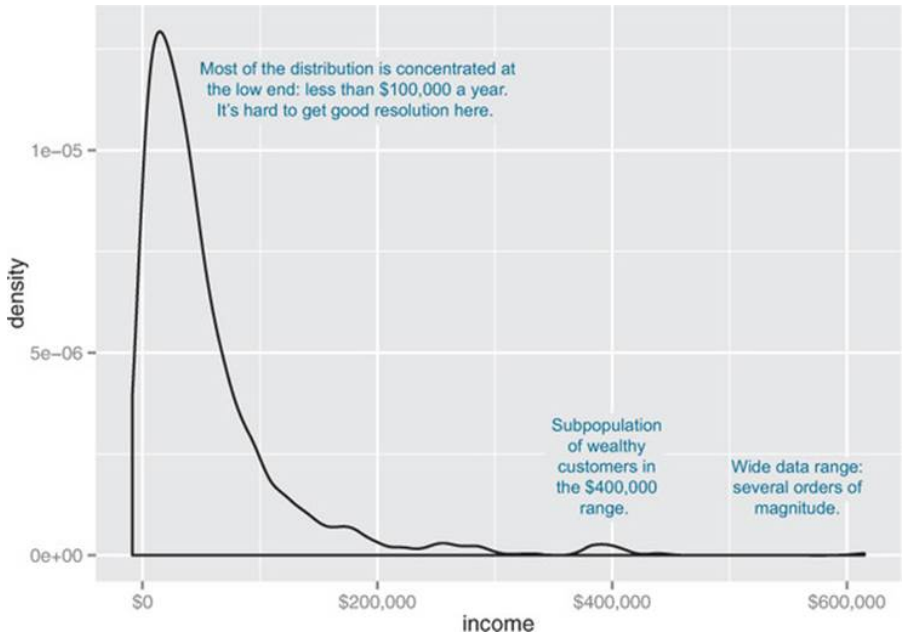
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

  - K is the kernel - a non-negative function that integrates to one
  - and $h > 0$ is a smoothing parameter called the bandwidth
- Smoothed version of the histogram allowing for smoother distributions.

DENSITY PLOTS

- Density plot: the overall shape of the curve
- E.g., the peaks of a Density Plot help display where values are concentrated over the interval

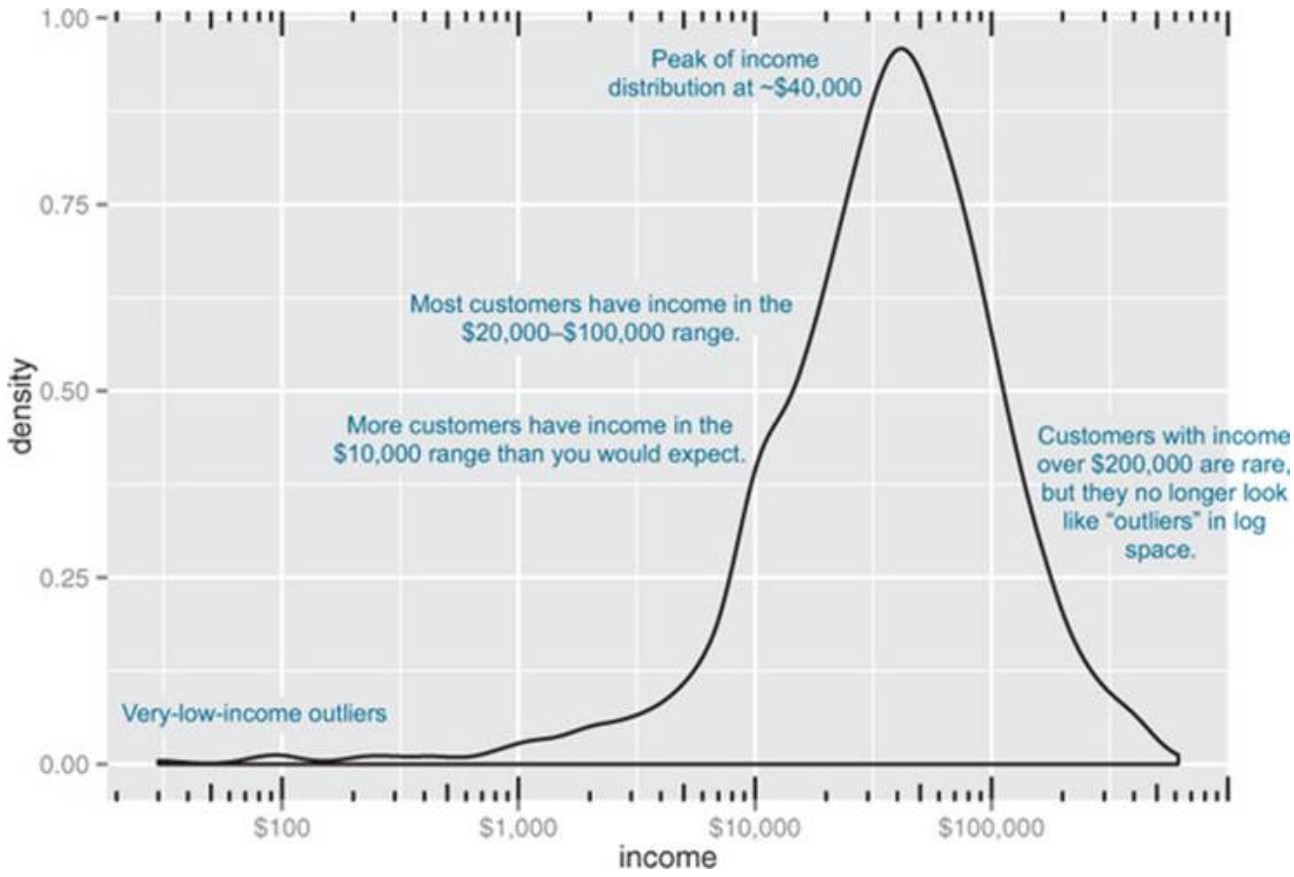




- The data range is very wide and the mass of the distribution is heavily concentrated to one side
- It's difficult to see the details of its shape: hard to tell the exact value where the income distribution has its peak

Log-scale

If the data is non-negative, then one way to bring out more detail is to plot the distribution on a logarithmic scale:

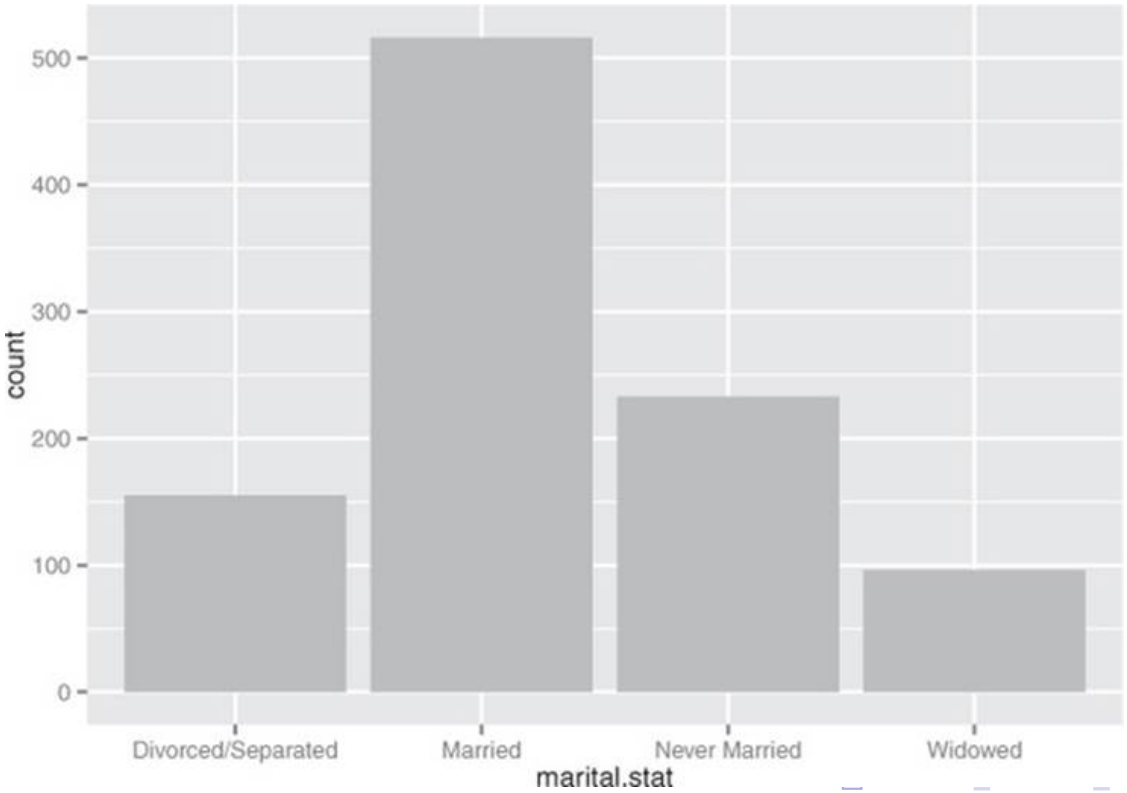


When should you use a logarithmic scale?

- To better visualize data that is heavily skewed
- When percent change is more important than changes in absolute units
  - e.g. in income data, a difference in income of five thousand dollars means something very different in a population where the incomes tend to fall in the tens of thousands of dollars than it does in populations where income falls in the hundreds of thousands or millions of dollars
  - what constitutes a significant difference depends on the order of magnitude of the incomes youre looking at

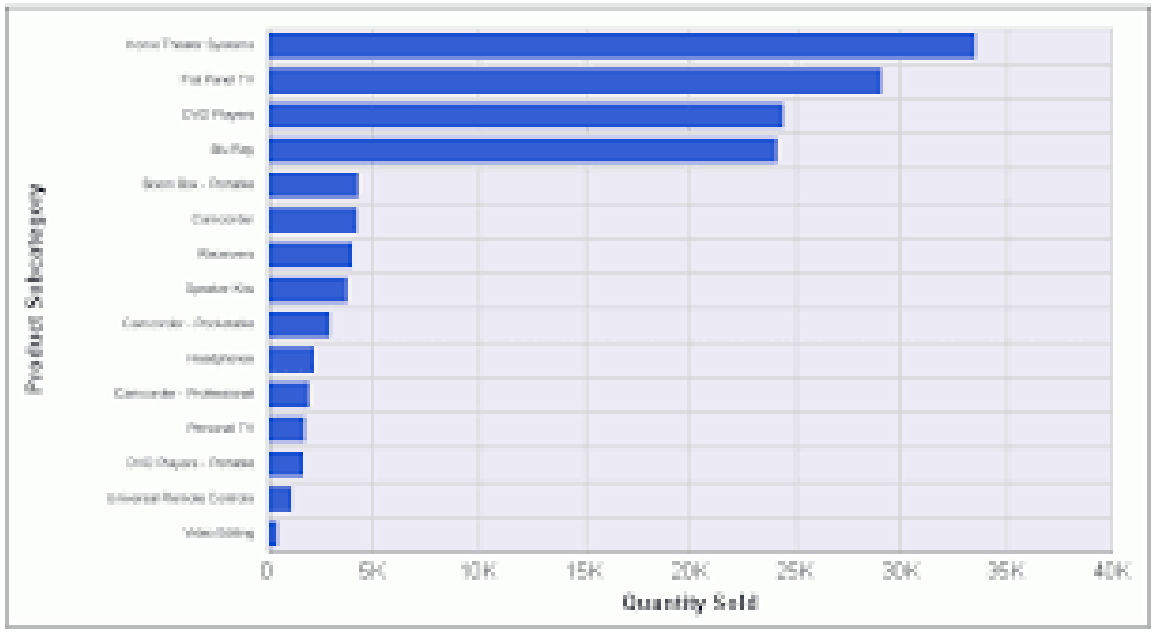
BAR CHARTS

- A bar chart is a histogram for discrete data: it records the frequency of every value of a categorical variable
  - e.g. if you believe that marital status helps predict the probability of health insurance coverage $\rightarrow$ check that you have enough customers with different marital statuses



BAR CHARTS

- Doesn't really show any more information than summary would show
- Most useful when the number of possible values is fairly large, like state of residence:
  - Horizontal graph is more legible than a vertical graph
  - The data in a bar chart can also be sorted to more efficiently extract insight from the data.



Checking relationships between two variables

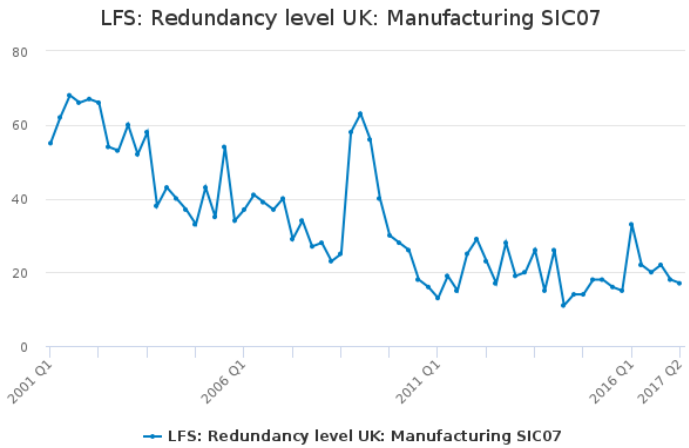
You'll often want to look at the relationship between two variables:

- Is there a relationship between the two inputs in my data?
- What kind of relationship and how strong?
- Is there a relationship between the input and the output variables? How strong?
- Precise quantification: during the modeling phase
- Exploring them gives a feel for the data and helps determine which variables are the best candidates to include in a model

LINE CHART

Shows the relationship between two continuous variables:

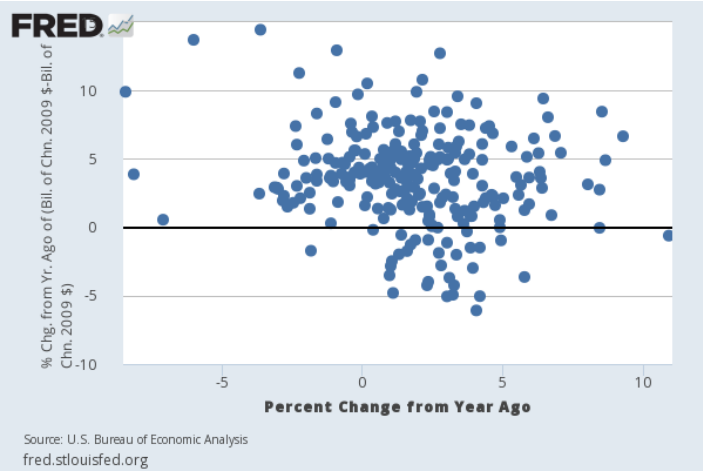
- Best when that relationship is functional, or nearly so: each x value has a unique (or nearly unique) y value
- Displays information as a series of data points called 'markers' connected by straight line segments
- Often used to visualize a trend in data over intervals of time - a time series
- When the data is not so cleanly related, line plots aren't as useful: use the scatter plot instead.



Source:

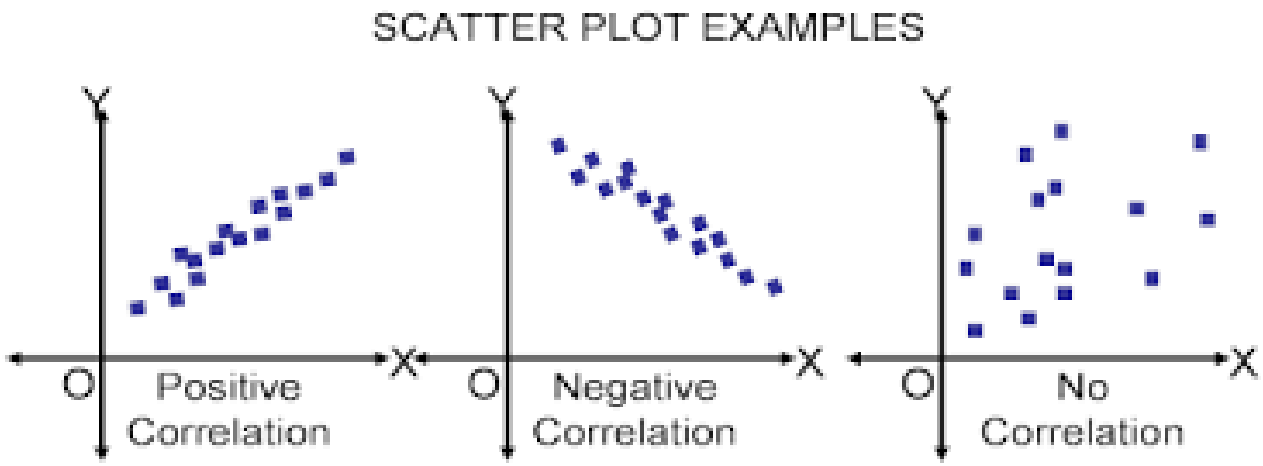
SCATTER PLOTS

- A scatter plot is a type of plot to display values for typically two variables for a set of data
- If the points are color-coded, one additional variable can be displayed
- The data is displayed as a collection of points
- Scatter plots are a straightforward way to visualize the data distribution in a XY plane, especially when we are looking for trends or clusters.



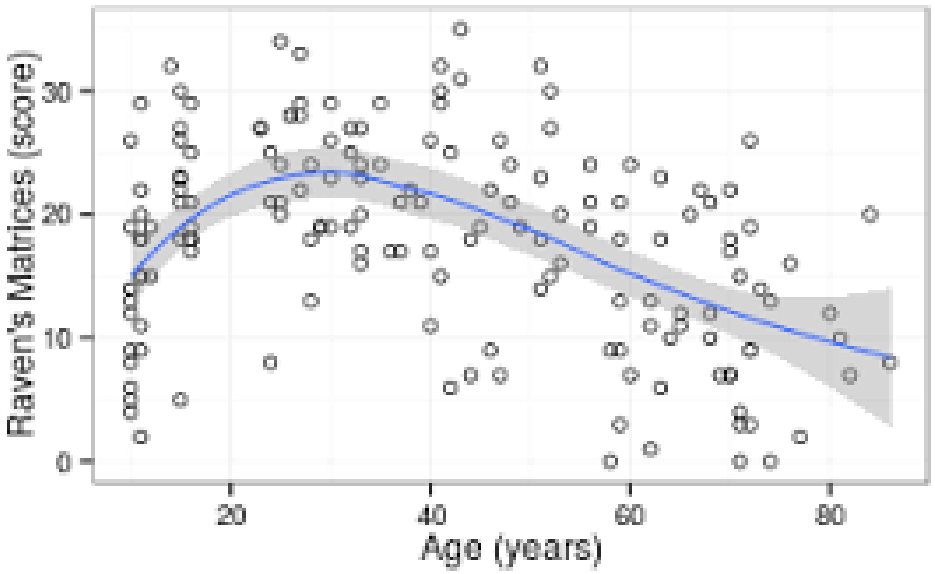
SCATTER PLOTS

- Can suggest various kinds of correlations between variables:
  - If the pattern of dots slopes from lower left to upper right, it indicates a positive correlation between the variables
  - If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation



SMOOTHING CURVES

- The relationship between two variables may not be easy to see: try to plot a linear fit through the data
- If the linear fit doesn't really capture the shape of the data: plot a smoothing curve through the data
- A scatter plot with a smoothing curve also makes a good visualization of the relationship between a continuous variable and a Boolean



SCATTER PLOTS

- If the dataset is small enough, the scatter plots will still be legible
- A scatter plot can only represent the point density up to a certain threshold
- For a dataset with a large number of points, many of these data points can overlap:
  - This overlapping effect can make it difficult to see any trends or clusters
  - Another type of visualization: binning methods which plot the point density rather than the points themselves

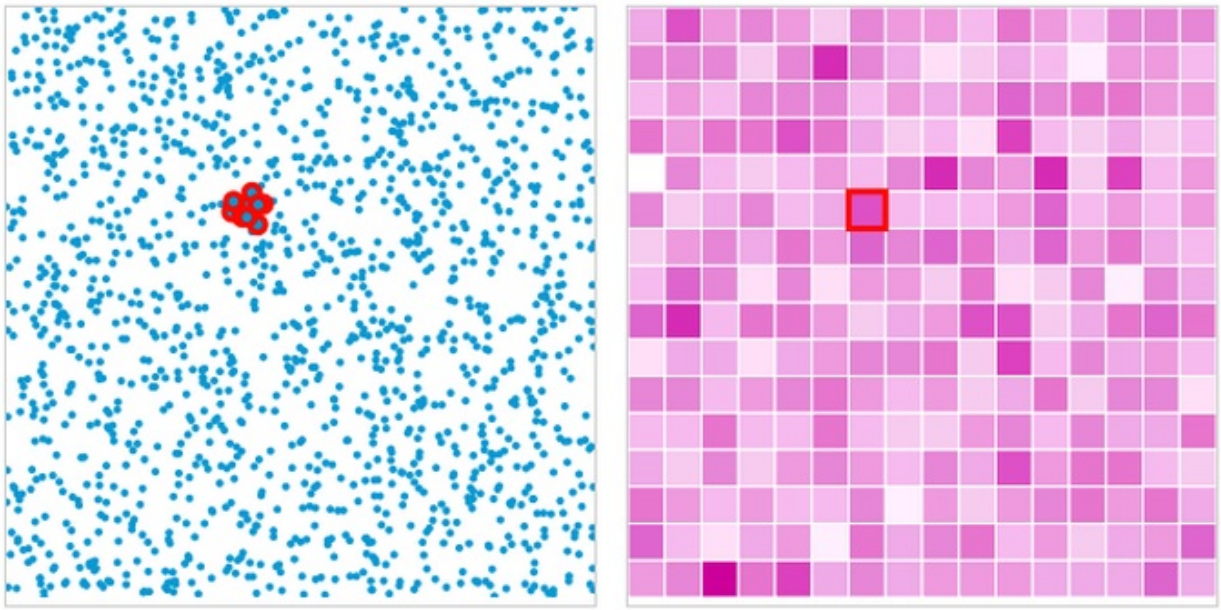
Binning

A binned representation is a technique of data aggregation which may reveal patterns not readily apparent in a scatter plot

- the XY plane is uniformly tiled with polygons (squares, rectangles or hexagons)
- the number of points falling in each bin (tile) are counted and stored in a data structure
- the bins with count > 0 are plotted using a color range or varying their size in proportion to the count
- If we consider the case of one-dimensional datasets, the binning technique generates histograms

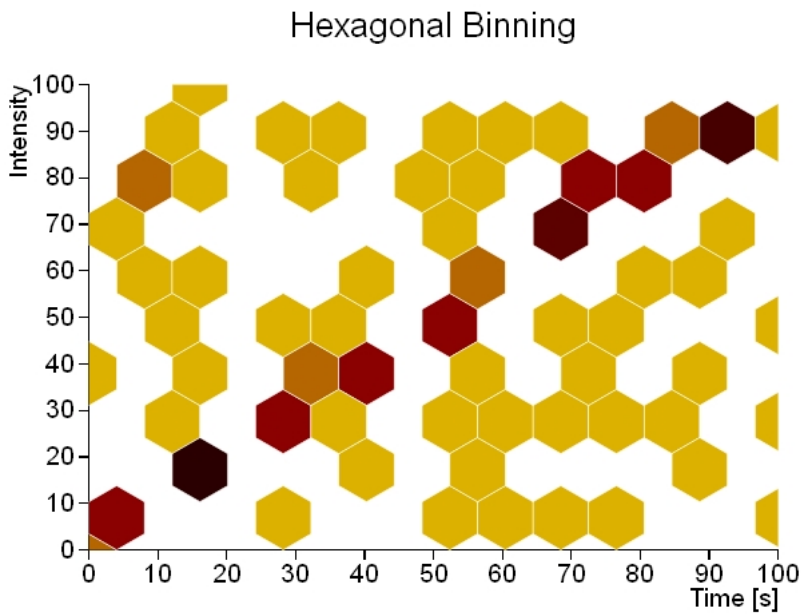
Rectangular Binning

- Simplest binning method
- Uses square tiles
- Advantage: its computational simplicity



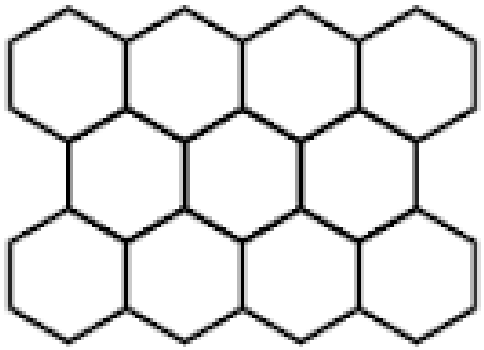
Hexagonal binning

- First described in 1987: D.B.Carr et al. Scatterplot Matrix Techniques for large N
- The hexagonal binning is the most efficient and compact division of 2D data space
- This is a consequence of the **hexagonal tessellation**

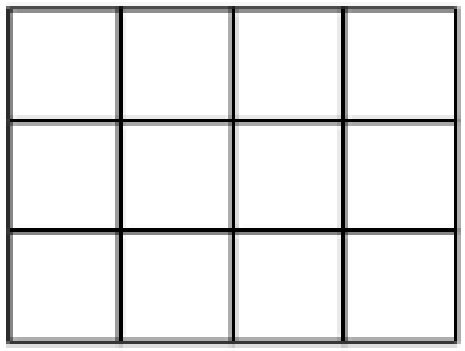


Regular tessellation:

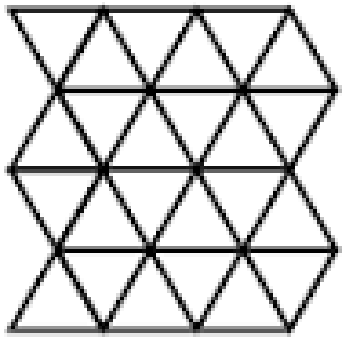
- The hexagon is the polygon with the maximum number of sides for a regular tessellation of a 2D plane:
  - Regular tessellation is not possible if you are using the same polygon with more than 6 sides
  - Only triangles, squares and hexagon can create a regular tessellation:



$\{6, 3\}$



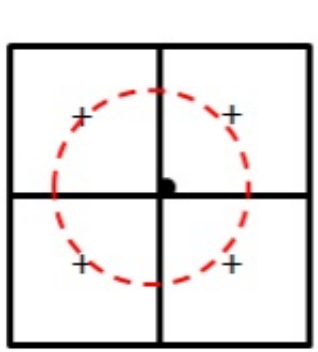
$\{4, 4\}$

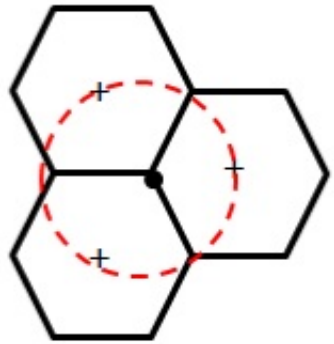


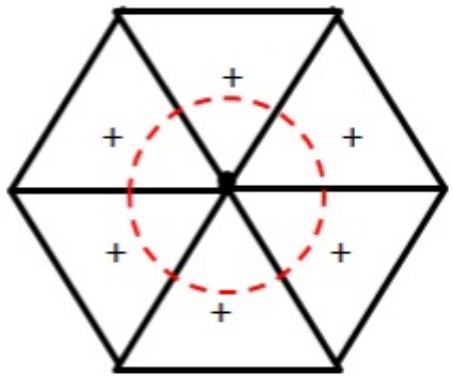
$\{3, 6\}$

Hexagonal tessellation

- In an hexagonal binning, adjacent hexagons share edges borders and not only vertices
- In square and triangular binning, triangles and square share only a vertex with some of its adjacent
- Any point inside a hexagon is closer to the center of any given point in an equal area square or triangle
- This translates into a more efficient data aggregation around the bin center

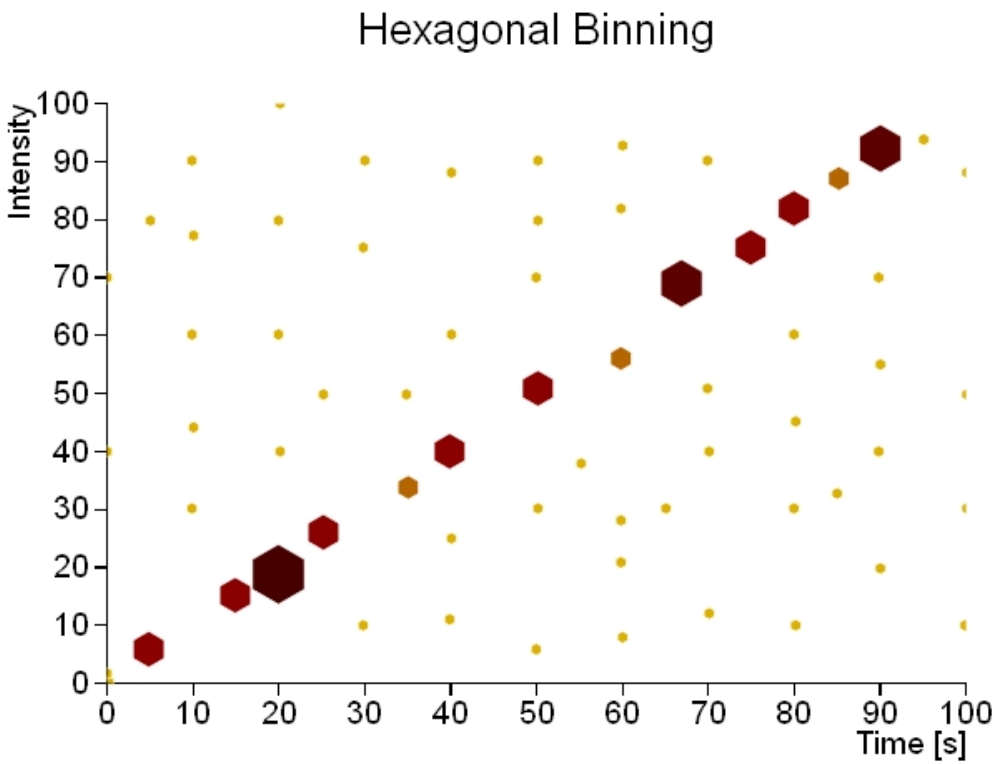






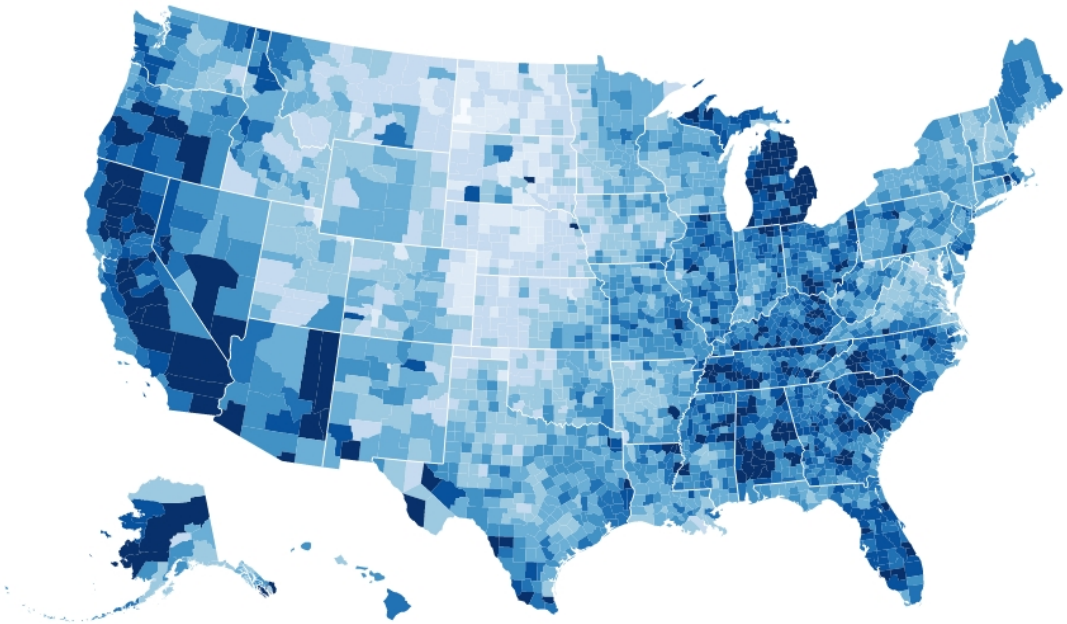
Multivariate Hexagonal Binning

- Possibility to draw hexagons with different sizes
- Two options here:
  - the variable distribution represented by color value/saturation is the same as that represented by size
  - they could be different, that is, for example, the size could represent the standard deviation

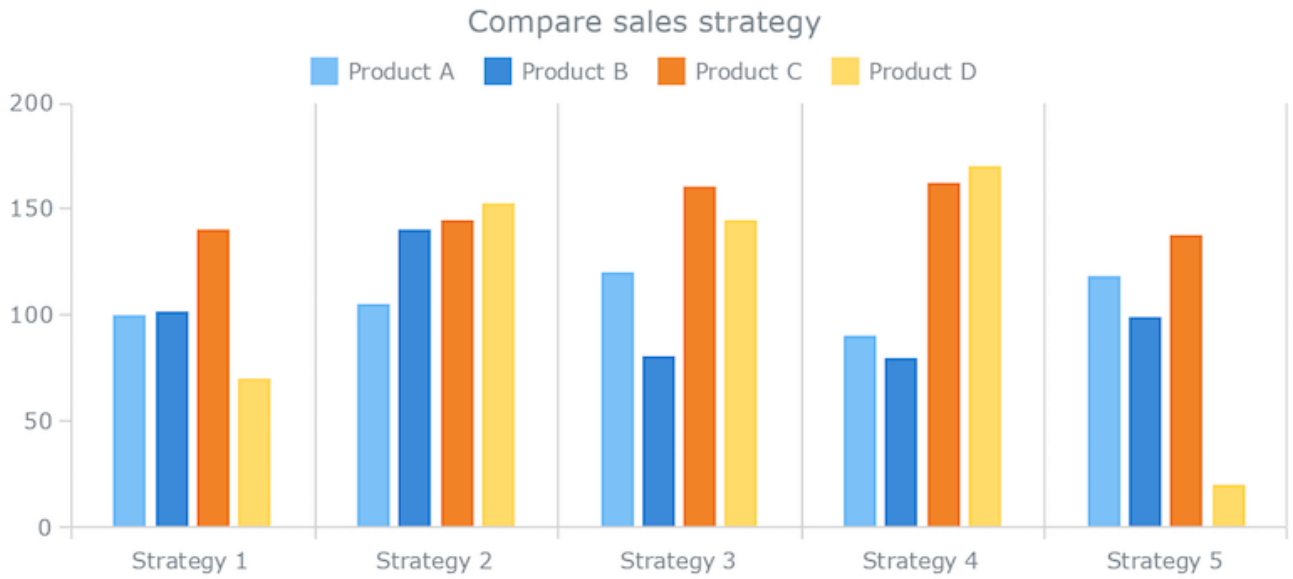


Choropleth maps

- The hexagonal binning is having a rapid spread in a specific field: cartography
- The choropleth maps are thematic maps in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map
- The idea is using hexagonal bins to represent data on maps



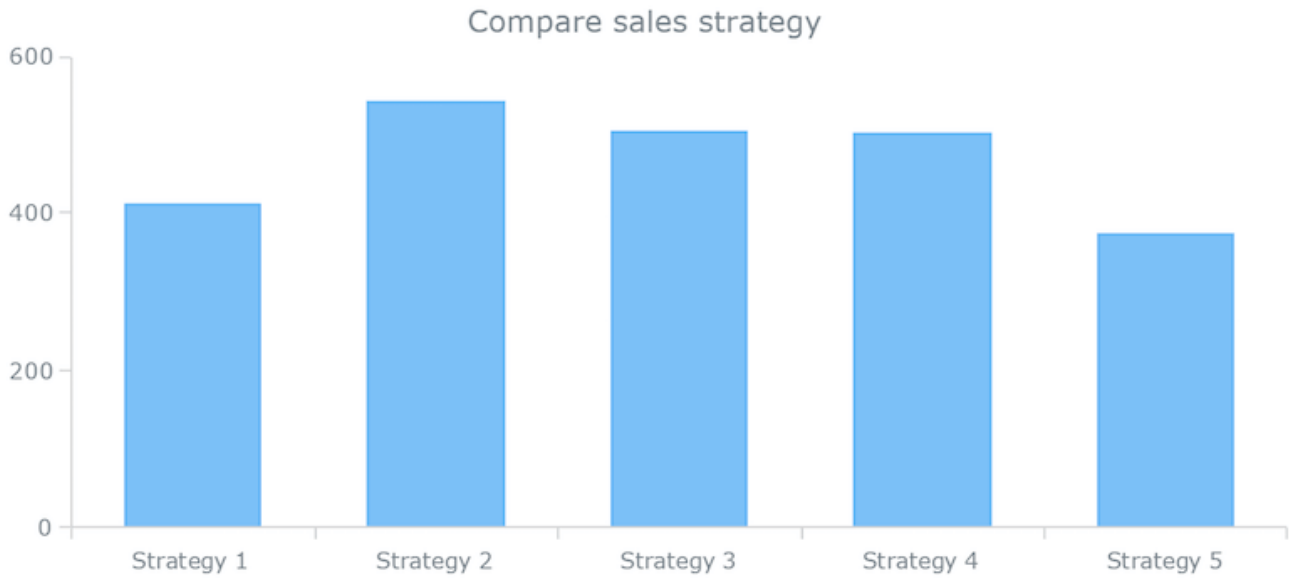
Multiseries bar chart



- Displays sales of each product within each sales strategy:
  - Which strategy generated the most sales of every single product?
  - How did the products perform individually within a given strategy?

Single-Series Bar Charts

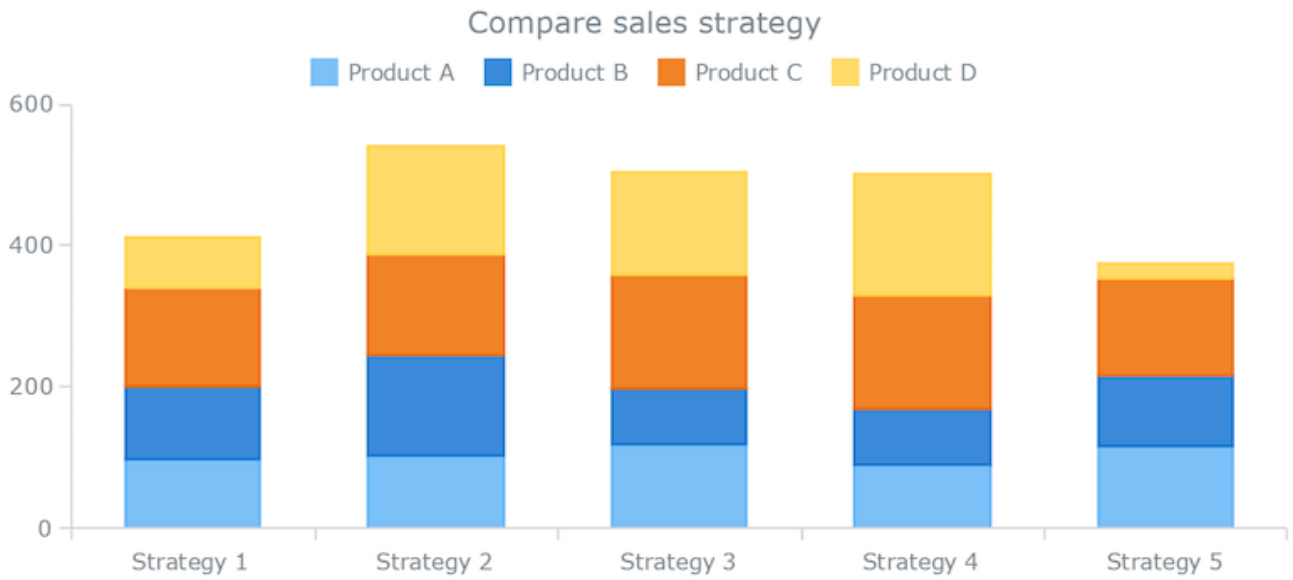
- If we only want to see the overall sales for each strategy and then compare them with each other: Single-Series Bar Charts
- We simply add up the product values and represent them as columns in a single-series bar chart
- Is clear that the Strategy 2 was overall most effective and Strategy 5 was the least.



Stacked bar chart

- The relationships between what we noticed in the first (multi-series) and second (single-series) graphs: plot both category totals and product-specific data on one stage
- Stacked bar charts help simultaneously compare totals and notice sharp changes at the item level that are likely to have the most influence on movements in category totals.

Stacked bar chart

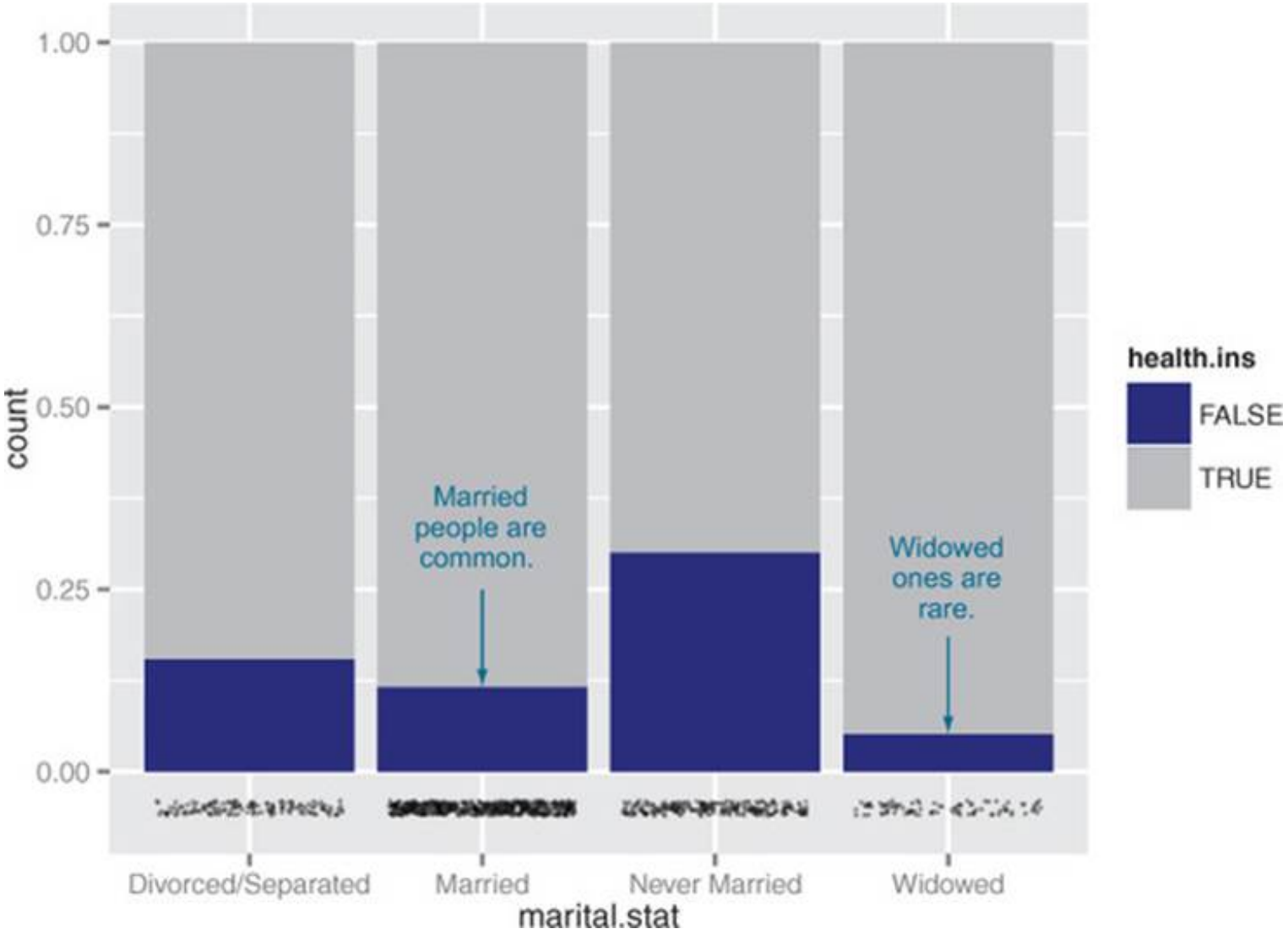


- Strategy 5 was the least effective overall
- Mainly because sales from Product D

Filled bar

- Compare the relative frequencies of each value of var2 within each value of var1 (works best when var2 is binary)
- To get a simultaneous sense of both the population in each category and the relative frequencies of each value of var2 within each value of var1, you can add a rug to the filled bar chart:
  - *A rug is a series of ticks or points on the x-axis. The rug is dense where you have a lot of data, and sparse where you have little data*

Filled bar



Visualization

- Which bar chart you use depends on what information is most important for you
- You should try different kinds of graphs to get different insights from the data
- Visualization is an interactive process: one graph will raise questions that you can try to answer by replotting the data again, with a different visualization

Goal: to explore your data enough to get a sense of it, to feel your data and to spot most major problems and issues.

Outline

Visual Noise

Typically, for Big Data, the analyst cannot observe the whole dataset, find anomalies in it, or find any relations from the first glance:

- **Visual Noise**: the presentation of whole array of data, can become a total mess on a screen
- Sometimes, the analyst cannot get even a bit of useful information from whole data visualization without any preprocessing tasks

Visual Noise

- Possible solution: data distribution above a larger screen
  - Occasionally, it ends up in another problem: large image perception
  - There is a certain level of human perception for different data visualization
  - After achieving this level of perception, we just loose the ability to acquire any useful information from the data overloaded view.

Information loss

- Another solution: reduction of visible data sets
  - Operates with data aggregation and filtration, based on some relations of objects in concrete dataset by one or more criteria
  - Solves the first problems
  - Leads to another problem: **information loss**
    - The analyst may miss some interesting hidden objects
  - Sometimes, complex aggregation process can consume a large amount of time and performance resources
  - This approach may be also difficult to customize: unknown nature of the incoming data
    - Filtering task in this approach can consist only of simple steps, such as excluding each second row, or removing some factors from data

Big Data Visualization Problems

Effectiveness and even applicability of methods can become a real problem with data volumes growth and data production speed:

(1) the need of artificial preparation of data slices for partial data visualization

(2) visual limitation to the number of perceived data factors

Big Data visualization results in loss of analysis quality.

More Than One View per Representation Display

- A simple approach: placing different classical data views which include only a limited set of factors
  - easily find some relations between these views or in one concrete view
  - any method of data visualization can be used here
  - often similar or near to similar graphical objects, e.g. diagrams
  - key point: ability to select desirable data areas onto all related representations.

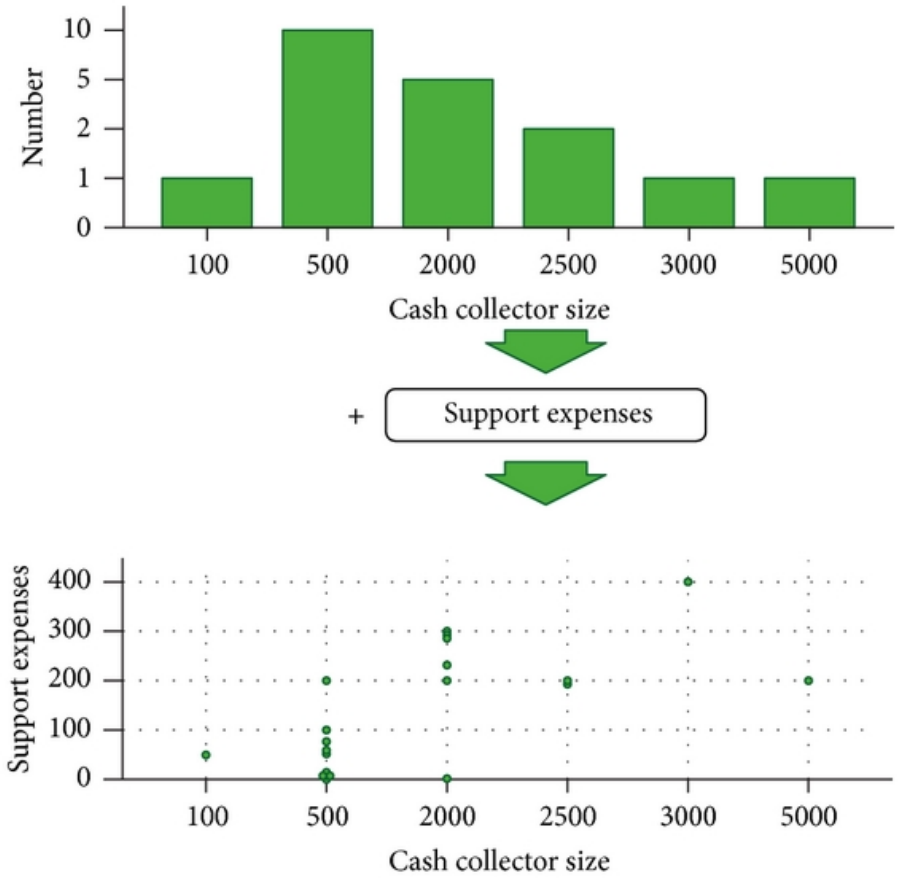
More Than One View per Representation Display

FIGURE 2: Data area selection onto related representations.

Dynamical change in the number of factors

- Assume that the analyst has chosen one factor: he is willing to see a classical histogram which shows the distribution of records number depending on record type
- After the analyst has chosen another factor, e.g., support expenses, the diagram type also changed into point diagram
- Continuing on, we can vary number of factors consequentially, lowering or increasing the number of visible factors and we will see changes in the diagram.

Dynamical change in the number of factors

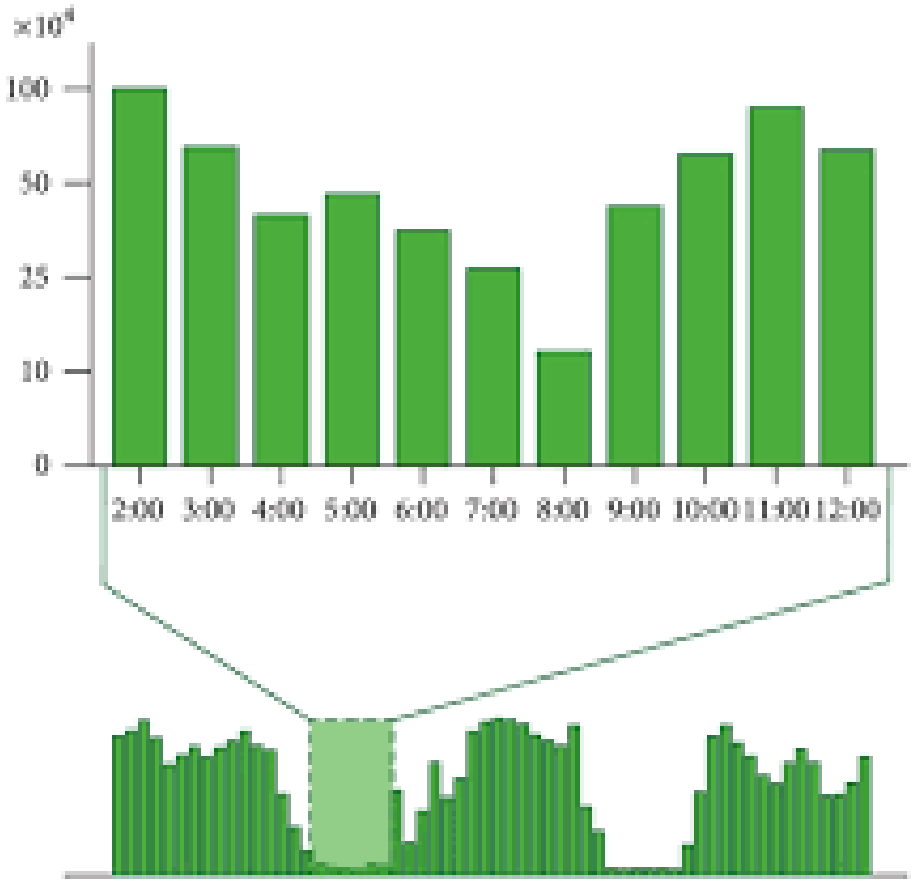


Filtering

- Human being cannot properly percept a large number of visible objects at once
- The analyst usually wants to see both whole data representation and a partial and more detailed data representation lying in his area of interest
- The area of interest is not static and can dynamically change during research process

Filtering

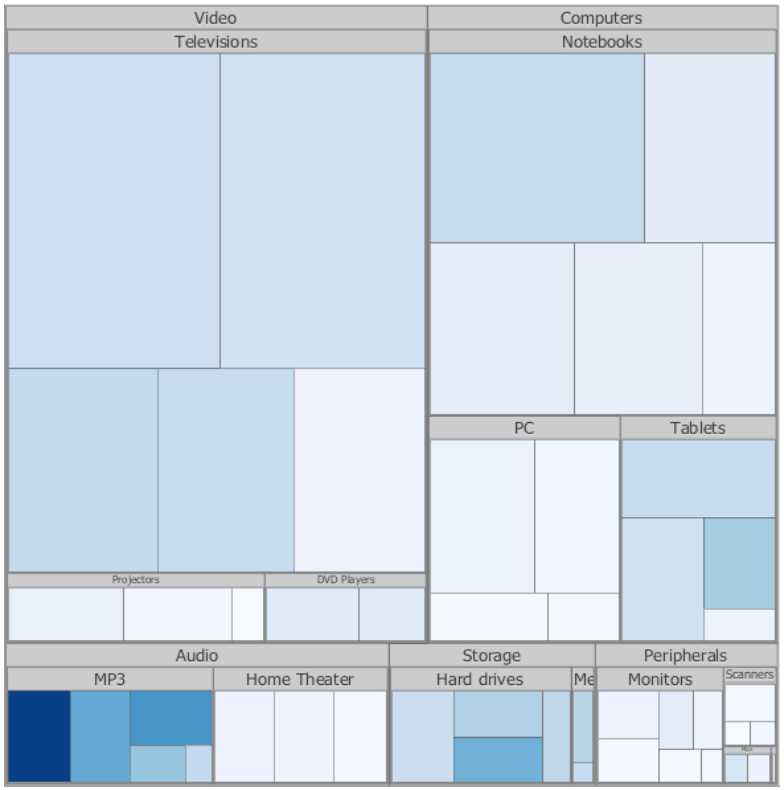
The analyst can change the range on an overview map and see the detailed visualization of data in that range:



Tree mapping

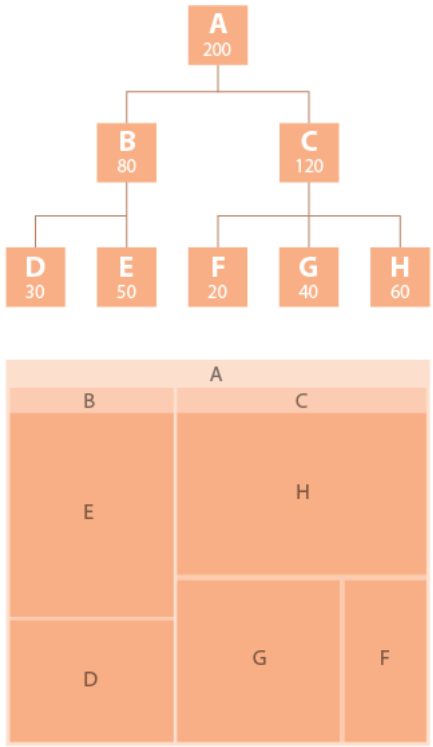
Tree mapping: is a method for displaying hierarchical data using nested figures, usually rectangles.

Example: Treemap showing sales (color) and profits (size) for all the product categories and type:



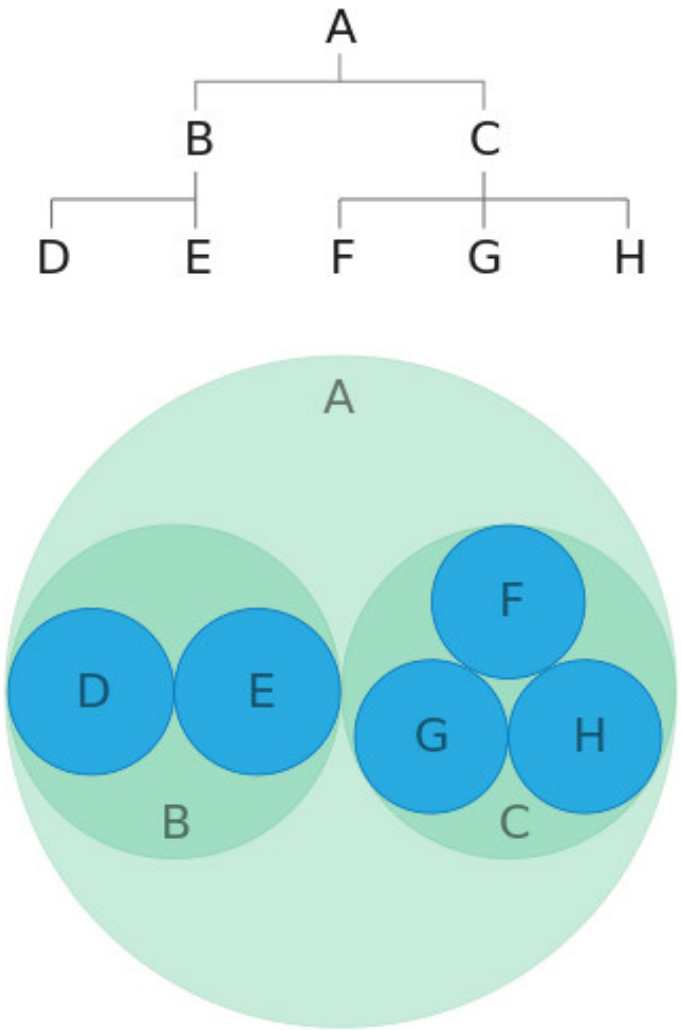
Tree mapping

- Each branch of the tree is given by a rectangle, which is then tiled with smaller rectangles representing sub-branches
- A leaf node's rectangle has an area proportional to a specified dimension of the data



Circle Packing

Circle Packing is a variation of a Treemap that uses circles instead of rectangles:

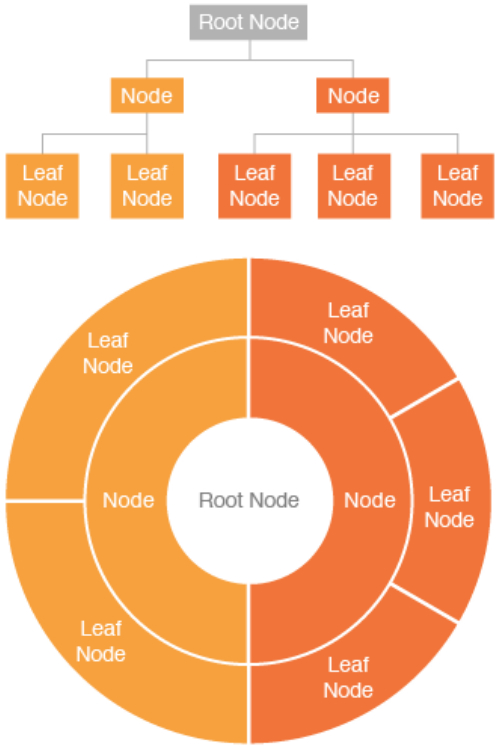


Sunburst

- An alternative to Treemap
- Uses Treemap visualization, converted to polar coordinate system
- The main difference between these methods is that the variable parameters are not width and height, but a radius and arc length
- This difference allows us not to repaint the whole diagram upon data change, but only one sector containing new data by changing its radius
- This method can be adapted to show data dynamics

Sunburst

- Shows hierarchy through a series of rings, that are sliced for each category node
- Each ring corresponds to a level in the hierarchy, with the central circle representing the root node and the hierarchy moving outwards from it:



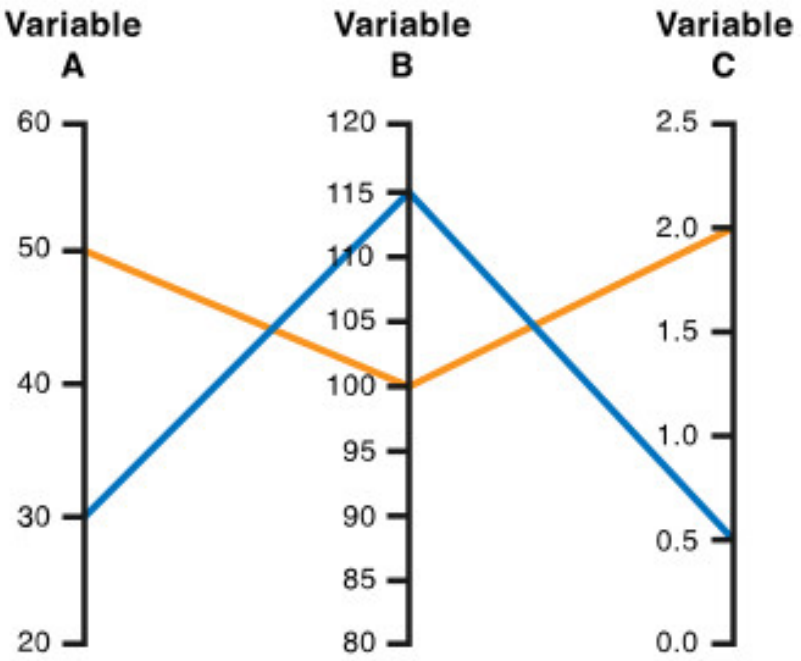
Parallel Coordinates Plots

- Parallel Coordinates Plot is used for plotting multivariate, numerical data
- Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them:
  - E.g. to compare an array of products with the same attributes (comparing computer or cars specs across different models)

Parallel Coordinates Plots

- Each variable is given its own axis and all the axes are placed in parallel to each other
- Each axis can have a different scale, as each variable works off a different unit of measurement, or all the axes can be normalized to keep all the scales uniform
- Values are plotted as series of lines connected across each axis
- Each line is collection of points placed on each axis, that have all been connected together

Parallel Coordinates



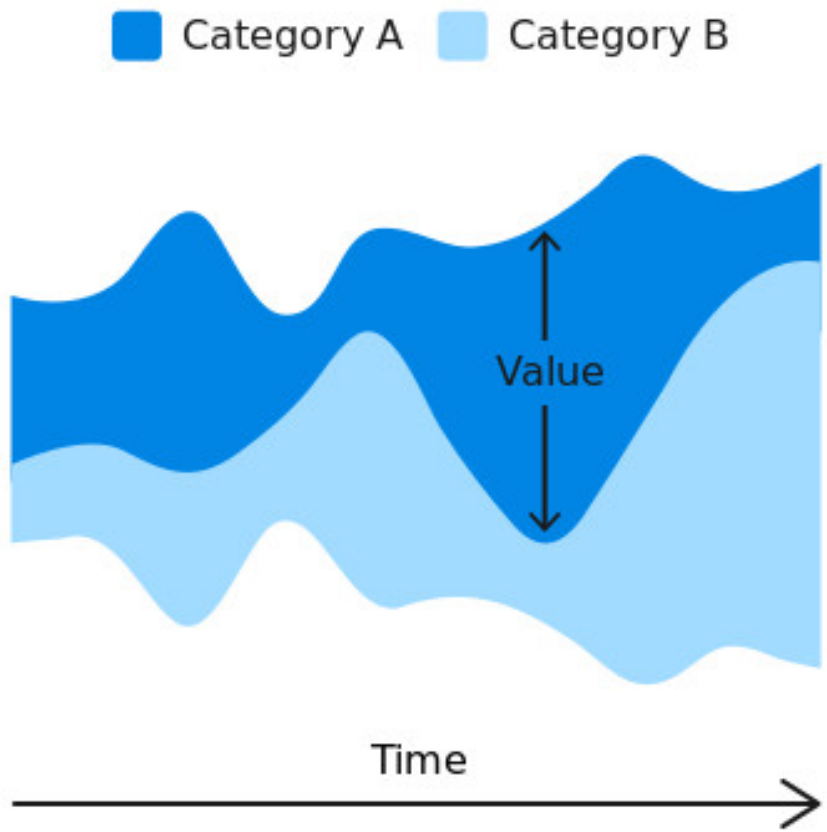
Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

Parallel Coordinates

- The downside: can become illegible when they're very data-dense:
  - The way to remedy this problem: "Brushing"
  - Brushing highlights a selected line or collection of lines, while fading out all the others.

Streamgraph

Stream Graphs display the changes in data over time of different categories through the use of flowing, organic shapes that somewhat resemble a river-like stream.



Streamgraph

- Aesthetically pleasing and more engaging to look at
- The size of each individual stream shape is proportional to the values in each category
- The axis that a Stream Graph flows parallel to, is used for the time scale
- Color is used to distinguish each category
- Can be used for displaying high-volume datasets, to discover trends and patterns over time across a wide range of categories

References

- https://datavizcatalogue.com
- *Practical Data Science with R*, Nina Zumel, John Mount, Manning (2014)