

VISUALIZATION

Visualizing Tables

Visualizing Tables

- Introduction
- Comparisons
- Compositions
- Correlations
- Time series data
- Distributions

Visualizing Tables

- Introduction
- Comparisons
- Compositions
- Correlations
- Time series data
- Distributions

Representing tables

- Two main ways of representing tables
 - Directly (textually): Tables/Spreadsheets
 - Symbolically (pictures): Graphs
- Use tables when
 - The document will be used to look up individual values
 - The document will be used to compare individual values
 - Precise values are required
 - The quantitative info to be communicated involves more than one unit of measure
- Use graphs when
 - The message is contained in the shape of the values
 - The document will be used to reveal relationships among values

Graph

- Visual display that illustrates one or more relationships among entities
- Shorthand way to present information
- Allows a trend, pattern or comparison to be easily comprehended

Graph Components

- Framework
 - Measurement types, scale
- Content
 - Marks, lines, points
- Labels
 - Title, axes, ticks

Tasks

- Search
 - Find a known/unknown item
- Query
 - Find out about characteristics of item by itself or relative to others
- Analyze
 - Compare
 - Trends over time
 - Correlations
 - Compositions
 - Distributions

Basic Plots for Basic Tasks

➔ Search

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

➔ Query

➔ Identify



➔ Compare



➔ Summarize



➔ ALL DATA

➔ Trends



➔ Outliers



➔ Features



➔ ATTRIBUTES

➔ One

➔ Distribution



↓ Extremes

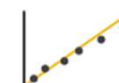


➔ Many

➔ Dependency



➔ Correlation



➔ Similarity



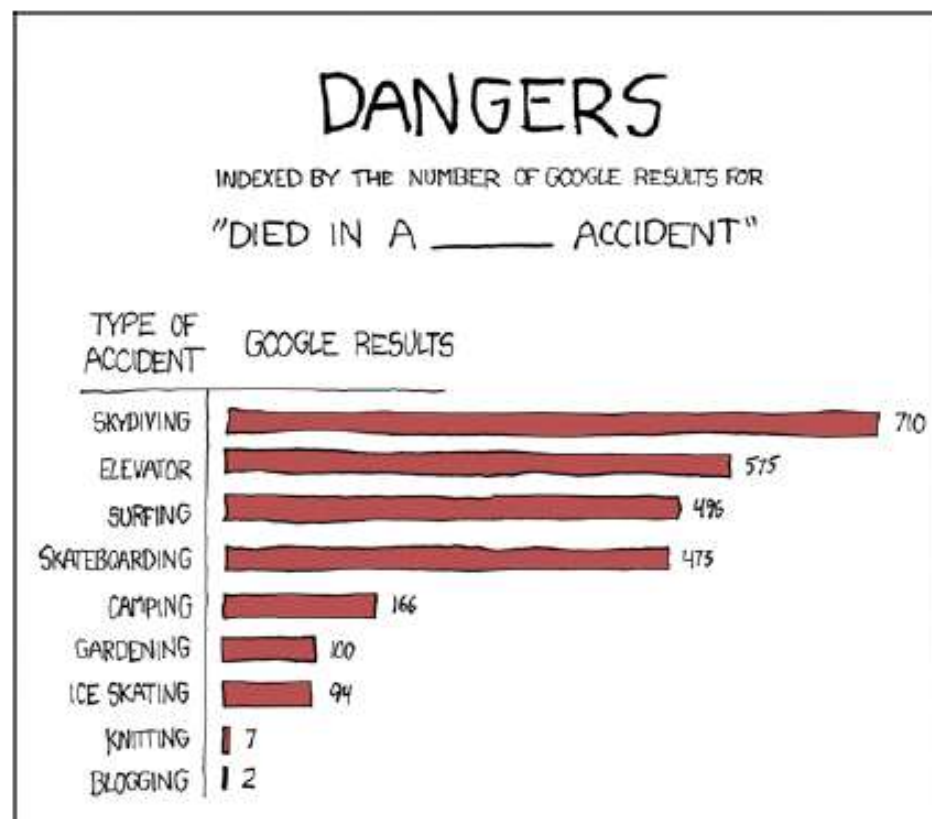
Few's Selection & Design Process

- Determine your message and identify your data
- Determine if a table, or graph, or both is needed to communicate your message
- Determine the best means to encode the values
- Determine where to display each variable
- Determine the best design for the remaining objects
 - Determine the range of the quantitative scale
 - If a legend is required, determine where to place it
 - Determine the best location for the quantitative scale
 - Determine if grid lines are required
 - Determine what descriptive text is needed
- Determine if particular data should be featured and how

Visualizing Tables

- Introduction
- **Comparisons**
- Compositions
- Correlations
- Time series data
- Distributions

Bar Charts



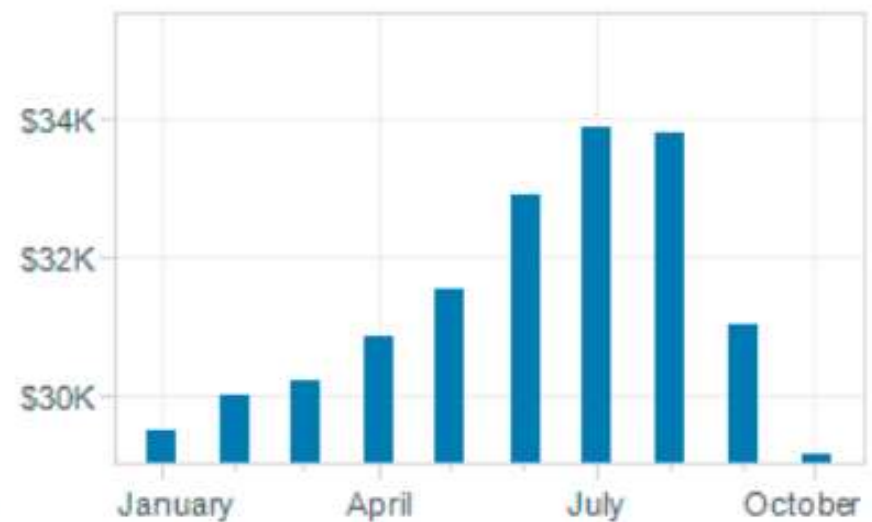
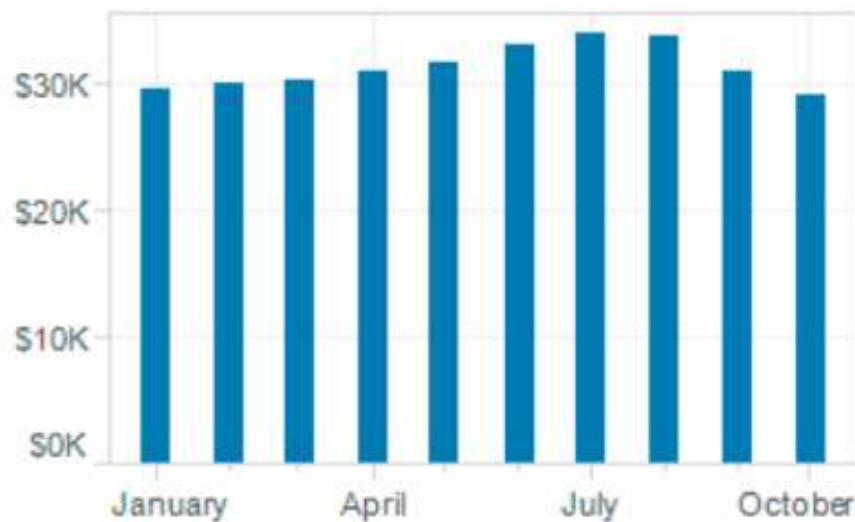
Good for showing comparisons

Bar Charts



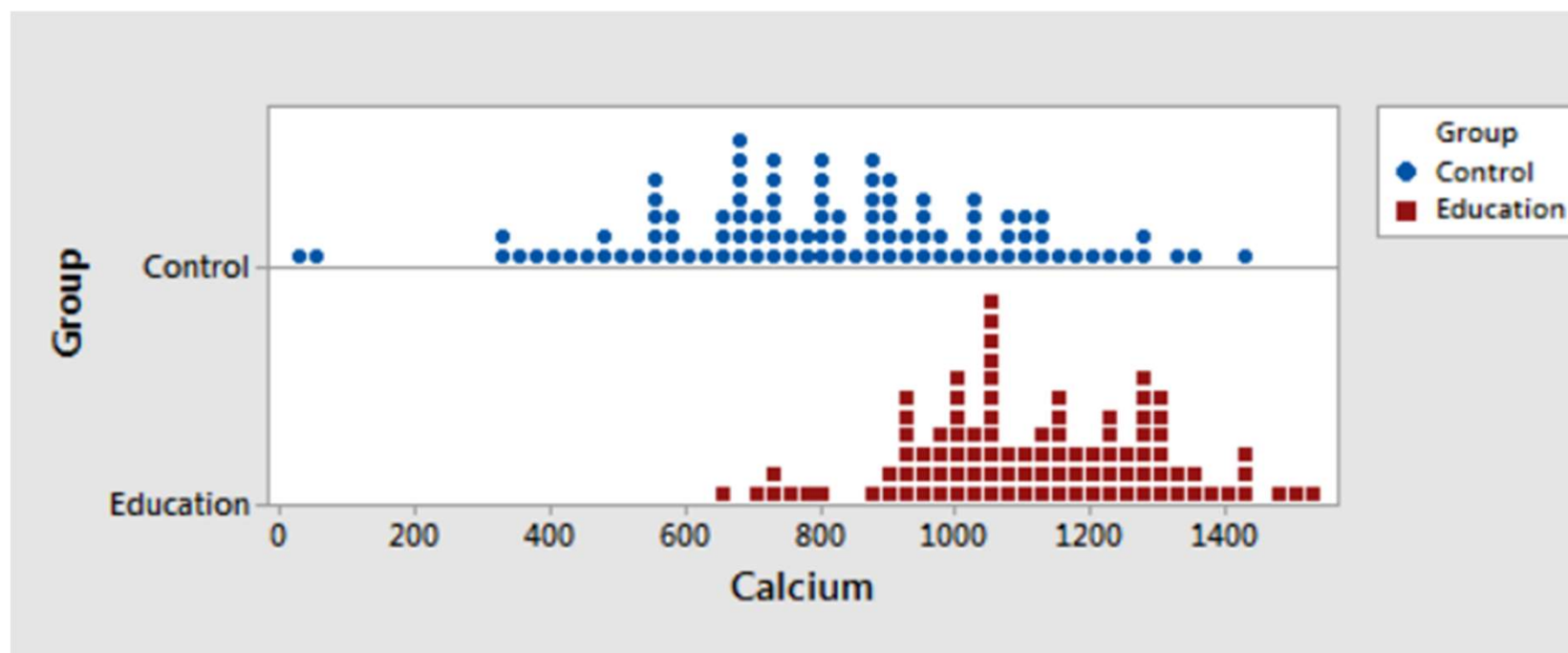
Good for showing comparisons

Baseline Problem



Dot Plots

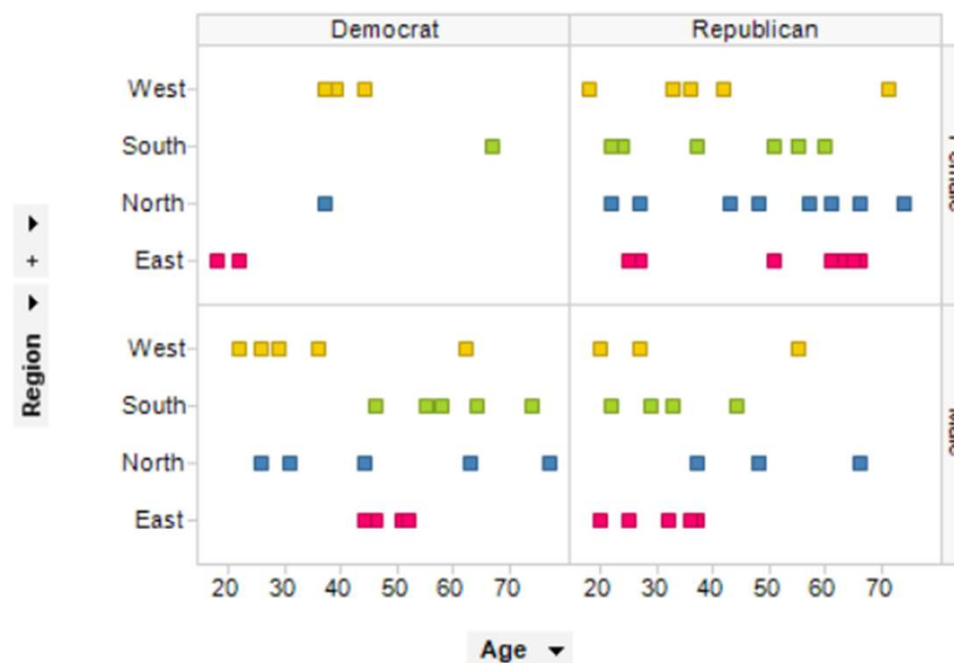
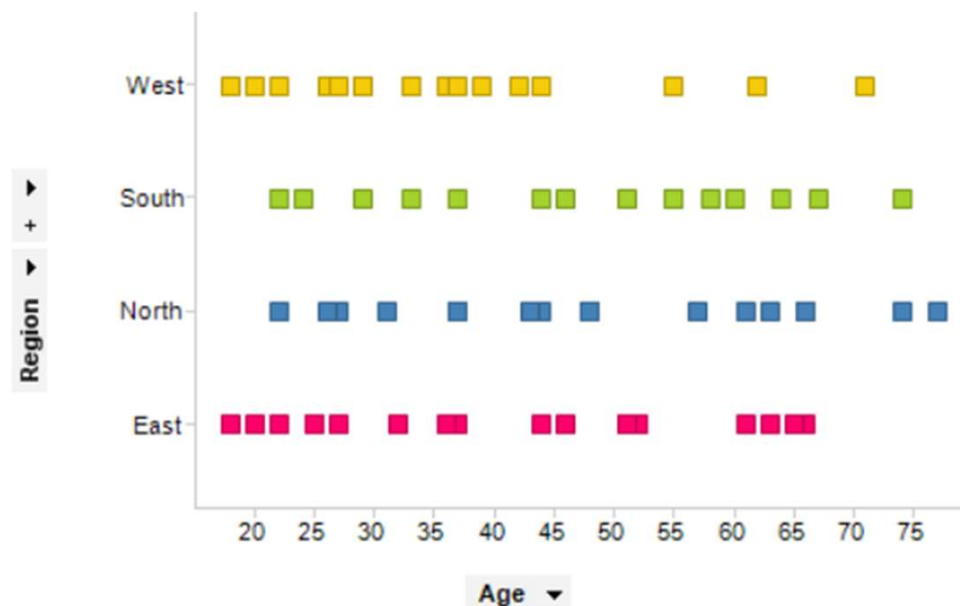
A **dot plot** is a chart consisting of data points plotted on a scale, typically using filled in shapes.



Trellis Visualization

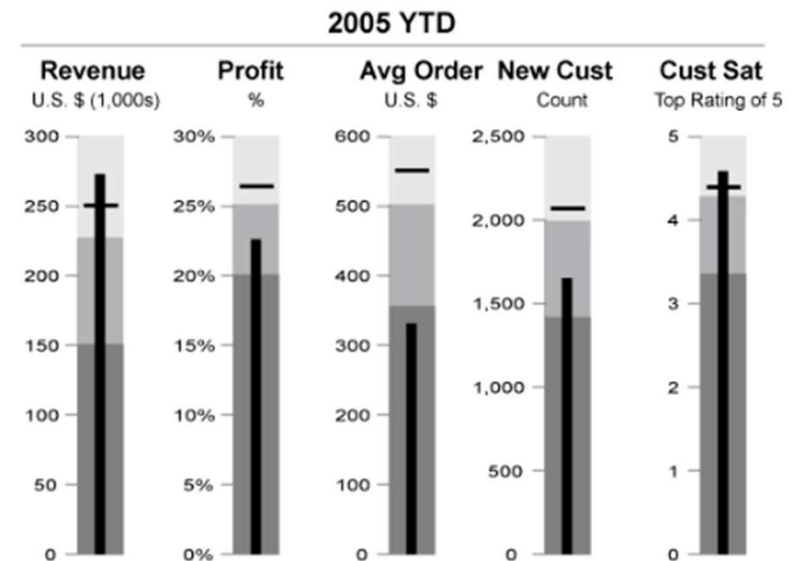
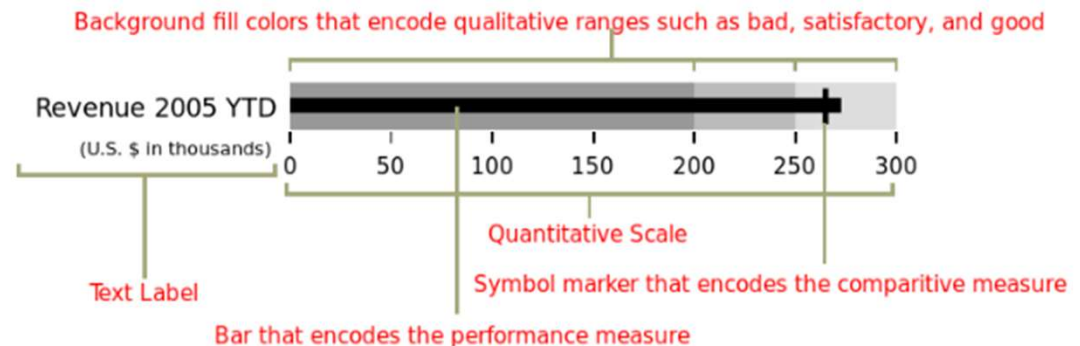
- Trellised visualizations enable you to quickly recognize similarities or differences between different categories in the data. Each individual panel in a trellis visualization displays a subset of the original data table, where the subsets are defined by the categories available in a column or hierarchy.
- For example, if you choose to trellis a visualization based on the two variables "Gender" and "Political affiliation", this will result in four separate panels representing the combinations Female-Republican, Female-Democrat, Male-Republican, and Male-Democrat.
- If the "Gender" variable is used in conjunction with another variable that has five different values, this will yield ten panels.
- From this follows that variables with a continuous distribution and a wide range of values (for example, Real values) should be binned before they are used to form a trellis visualization. Otherwise the number of panels quickly becomes unmanageable.

Trellis Visualization - Example



Bullet Graph

- A **bullet graph** is a variation of a [bar graph](#) developed by Stephen Few.
- The bullet graph features:
 - A single, primary measure,
 - Compares that measure to one or more other measures to enrich its meaning (for example, compared to a target)
 - Displays it in the context of qualitative ranges of performance, such as poor, satisfactory, and good.
 - The qualitative ranges are displayed as varying intensities of a single [hue](#)
 - Bullet graphs may be horizontal or vertical and may be stacked to allow comparisons of several measures at once



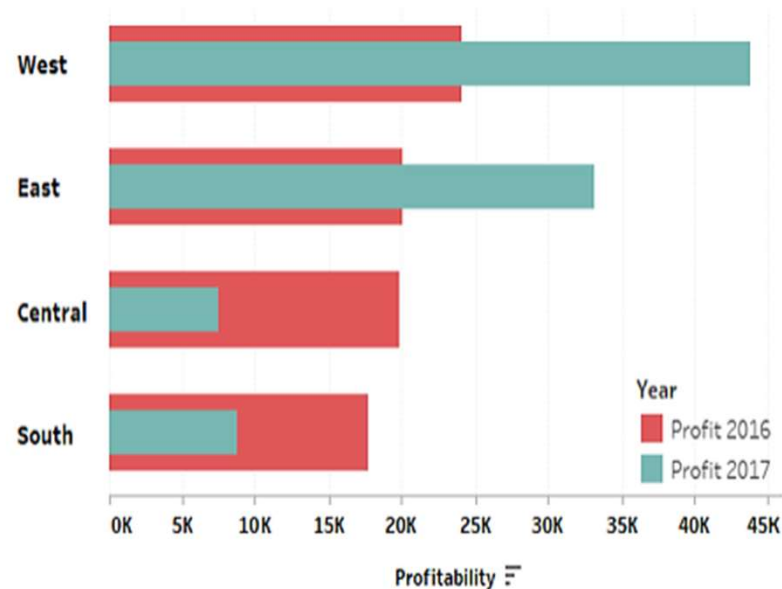
Visualizing Tables

- Introduction
- Comparisons
- **Compositions**
- Correlations
- Time series data
- Distributions

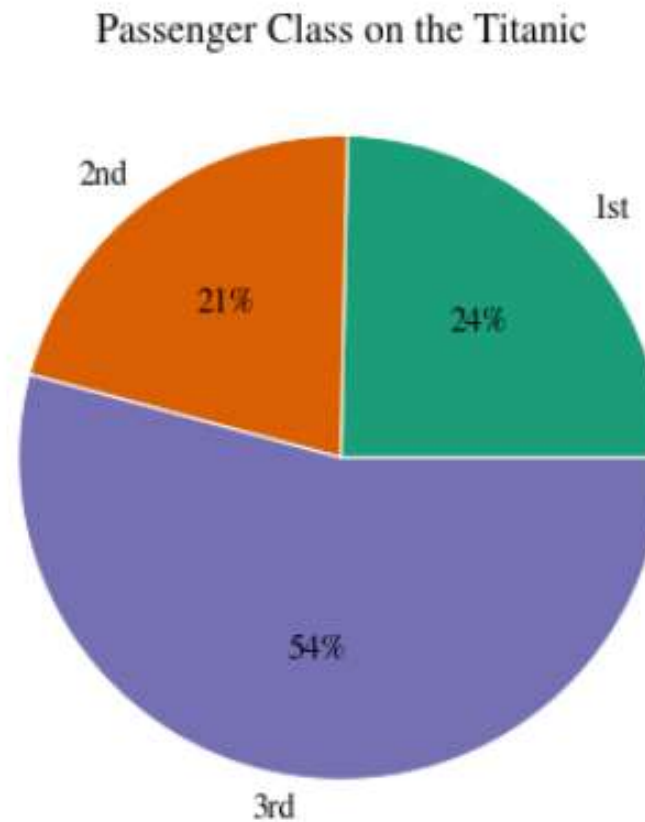
Bar in bar Graph

Bar-in-bar chart is simply two interlocking bar charts. Its useful when you need to illustrate change between two points. E.g. Profit between two timelines for different products or regions.

Example of a bar-in-bar chart



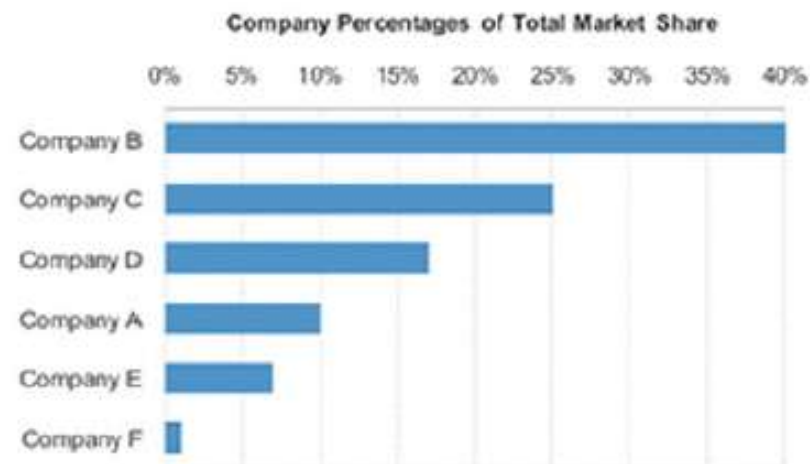
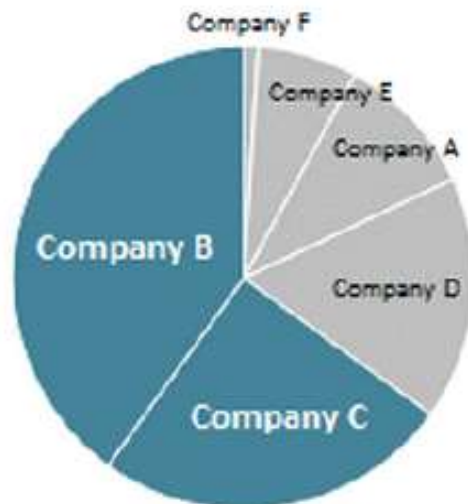
Pie Charts



Pie charts are good for showing compositions

Pie vs Bar Charts

65% of the market is controlled by companies B and C

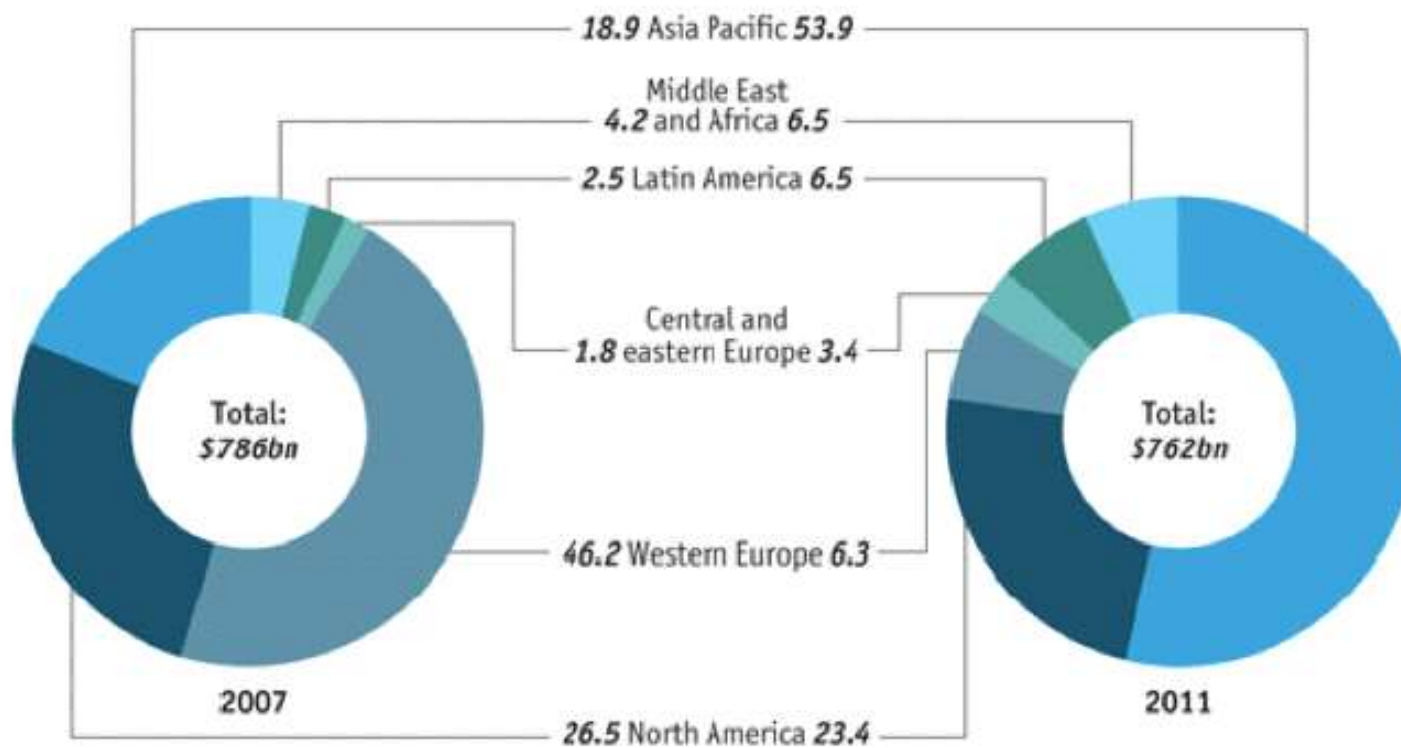


Donut Charts



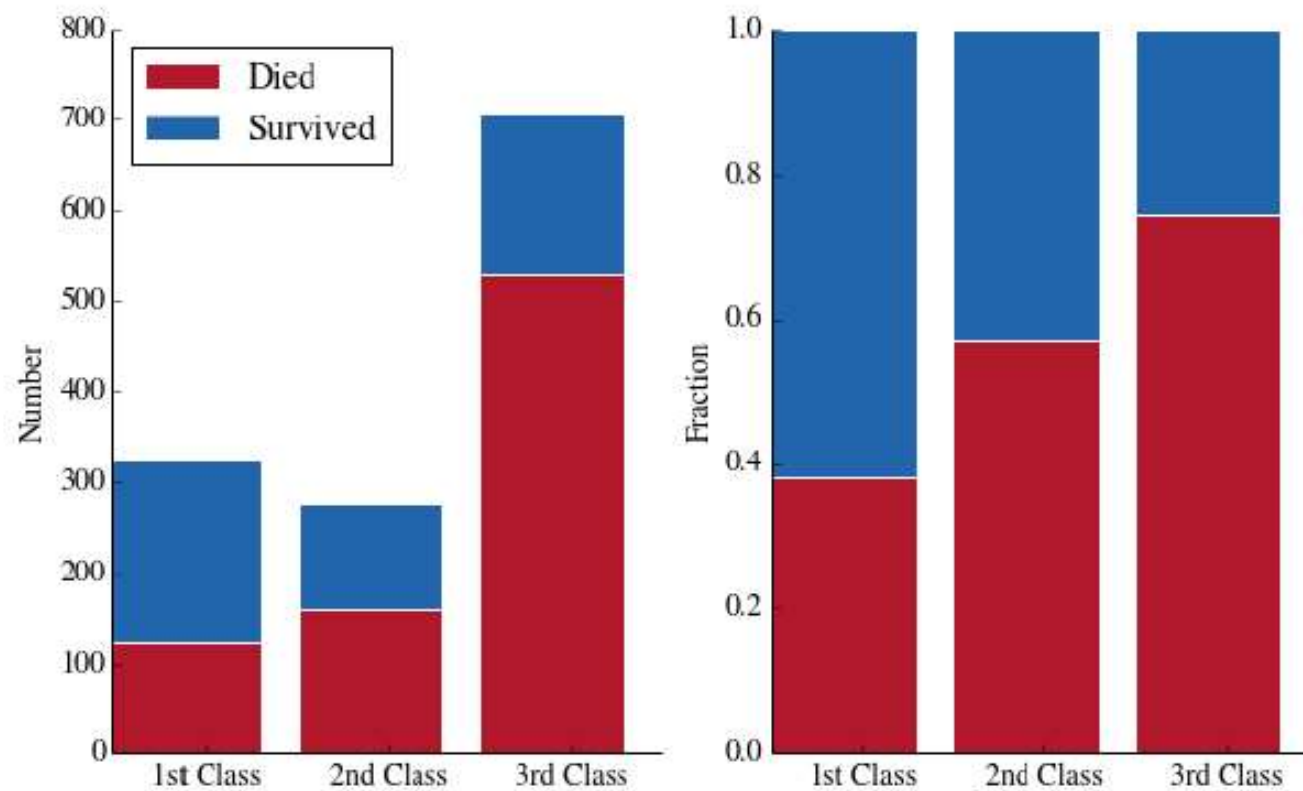
Pre-tax profits of the 1,000 largest banks

By tier-one capital and domicile, % of total

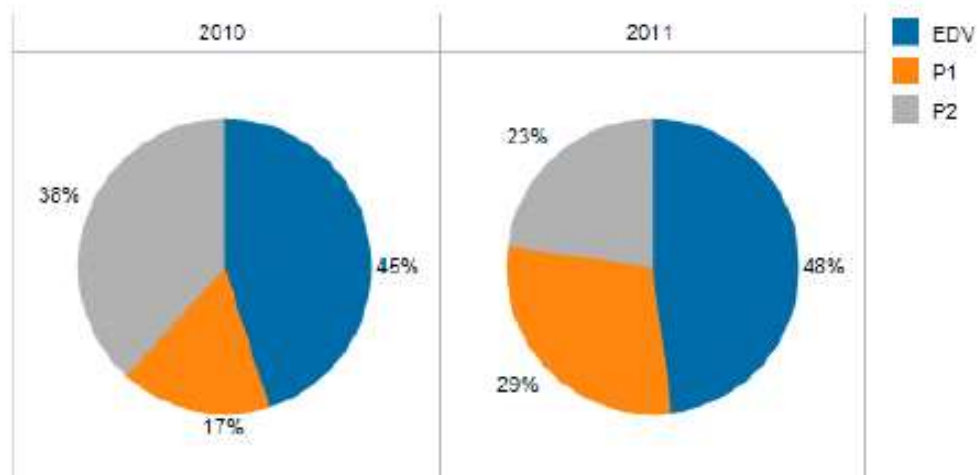


Source: *The Banker Top 1000*

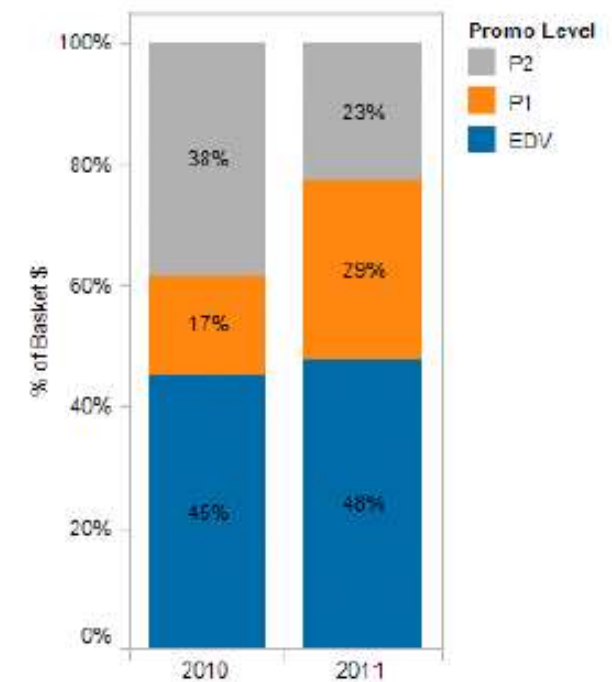
Stacked Bar Charts



Stacked Bar vs Pie Charts

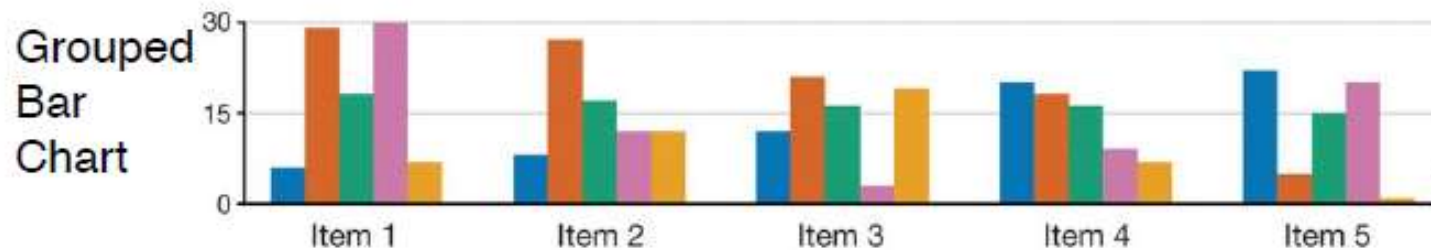
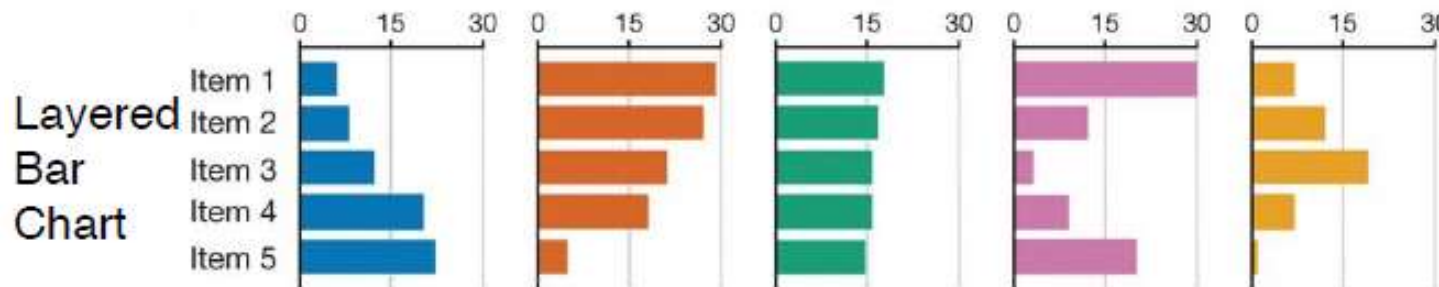
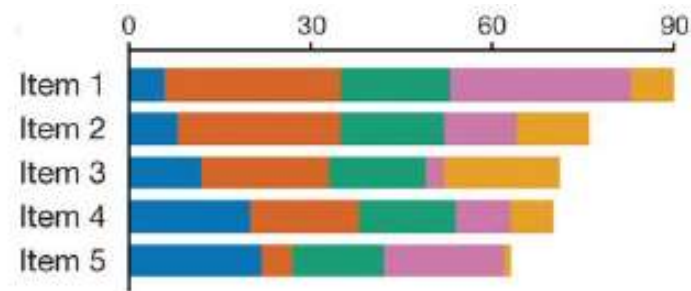
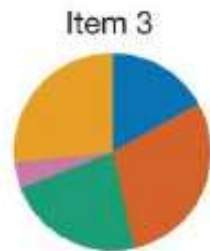


vs.



Comparison of Chart types

Category 1 ●
 Category 2 ●
 Category 3 ●
 Category 4 ●
 Category 5 ●

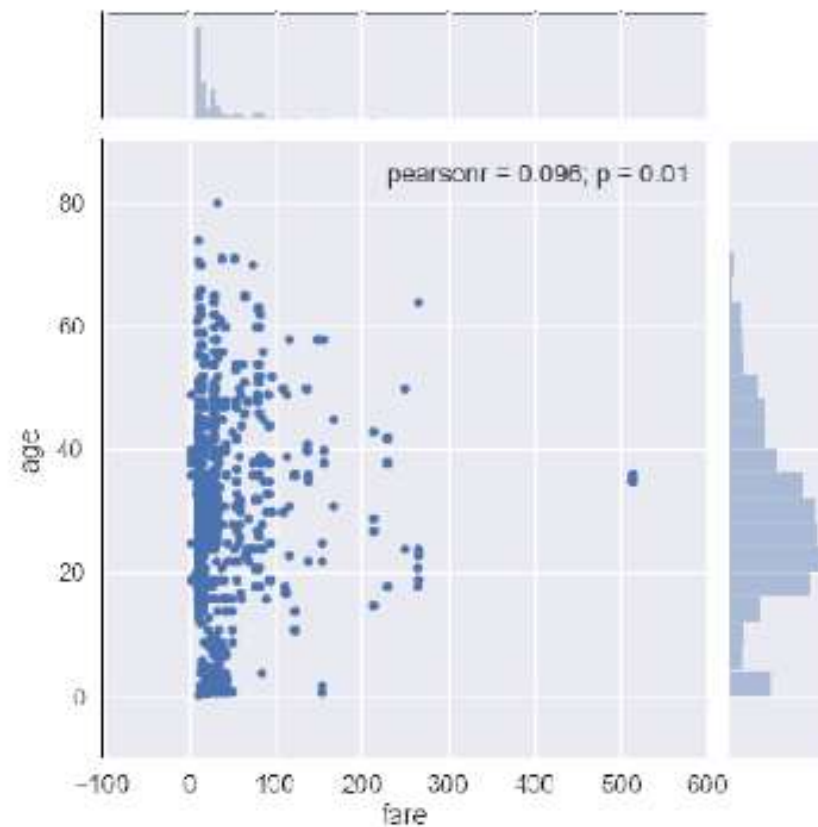


Small Multiples

Visualizing Tables

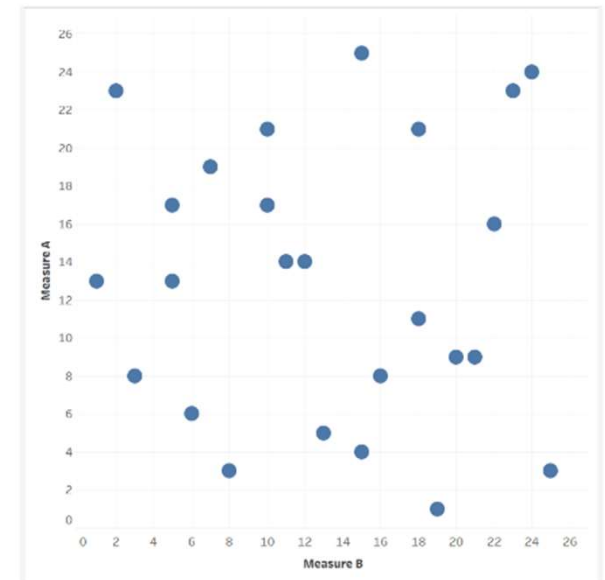
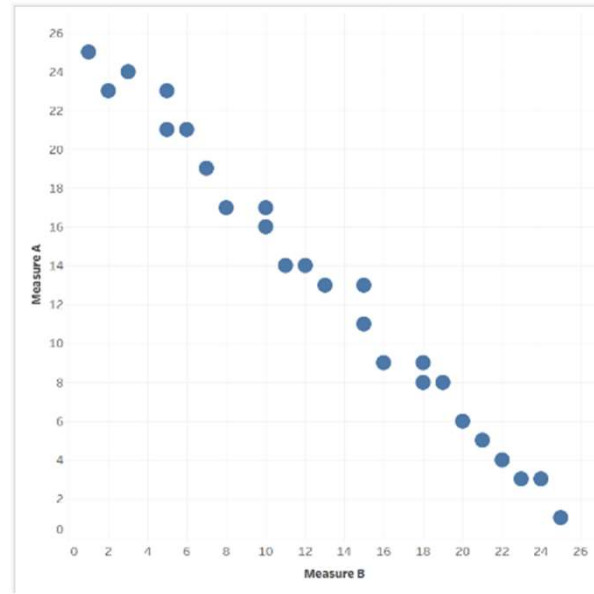
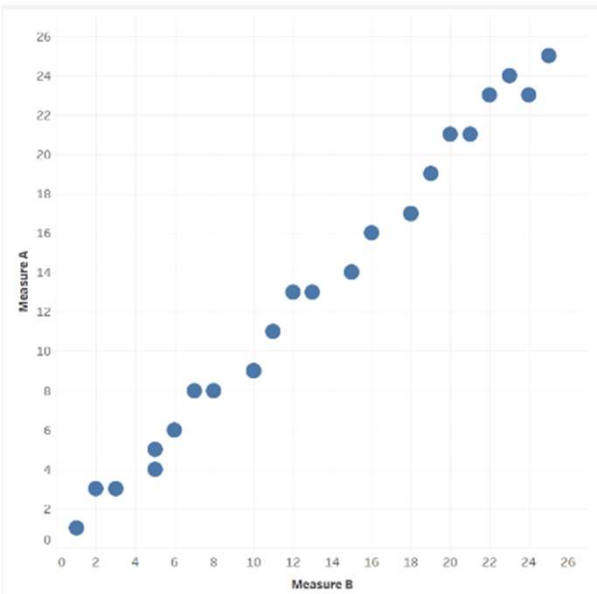
- Introduction
- Comparisons
- Compositions
- **Correlations**
- Time series data
- Distributions

Correlations

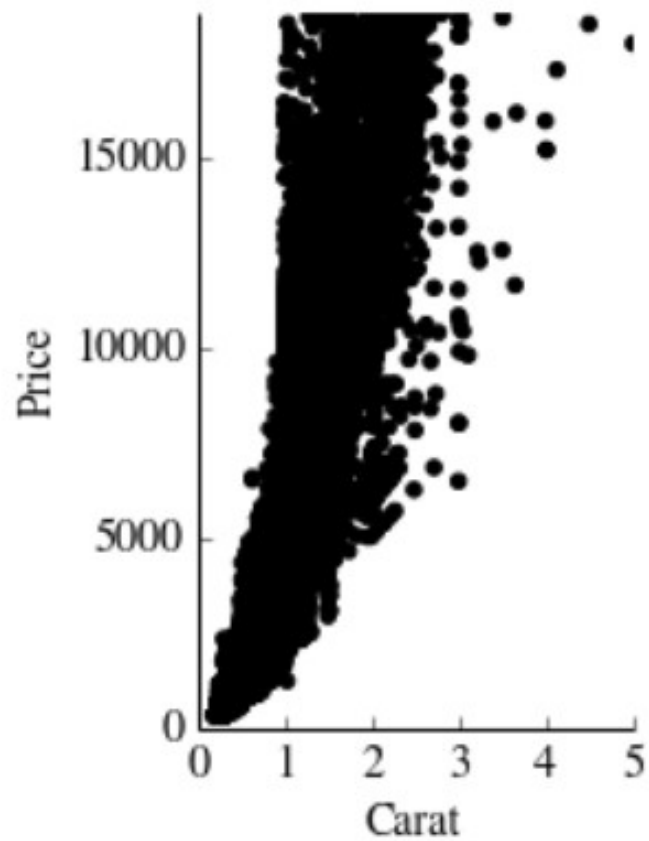


Scatterplots are very useful for showing correlations in 2 dimensional data

Correlations



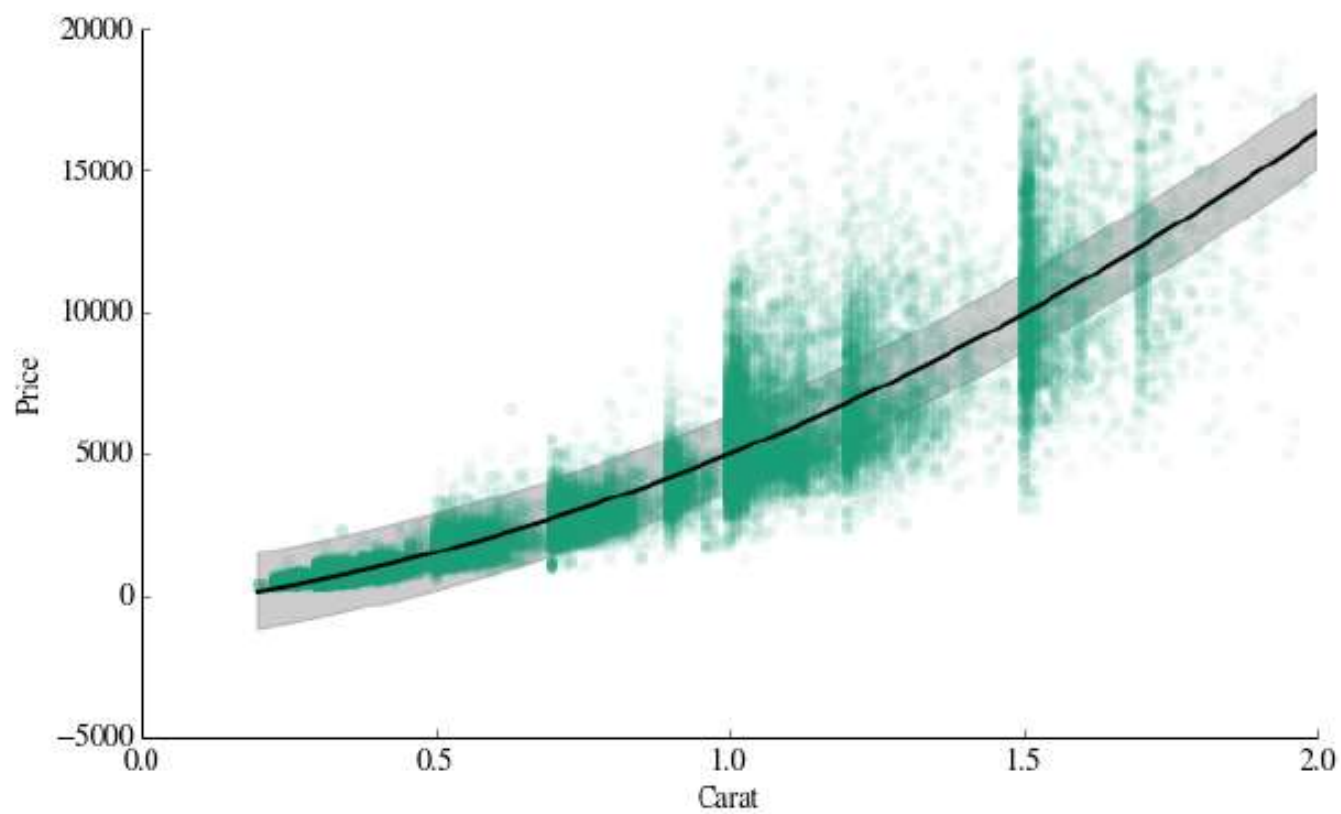
Overplotting



How to deal with overplotting

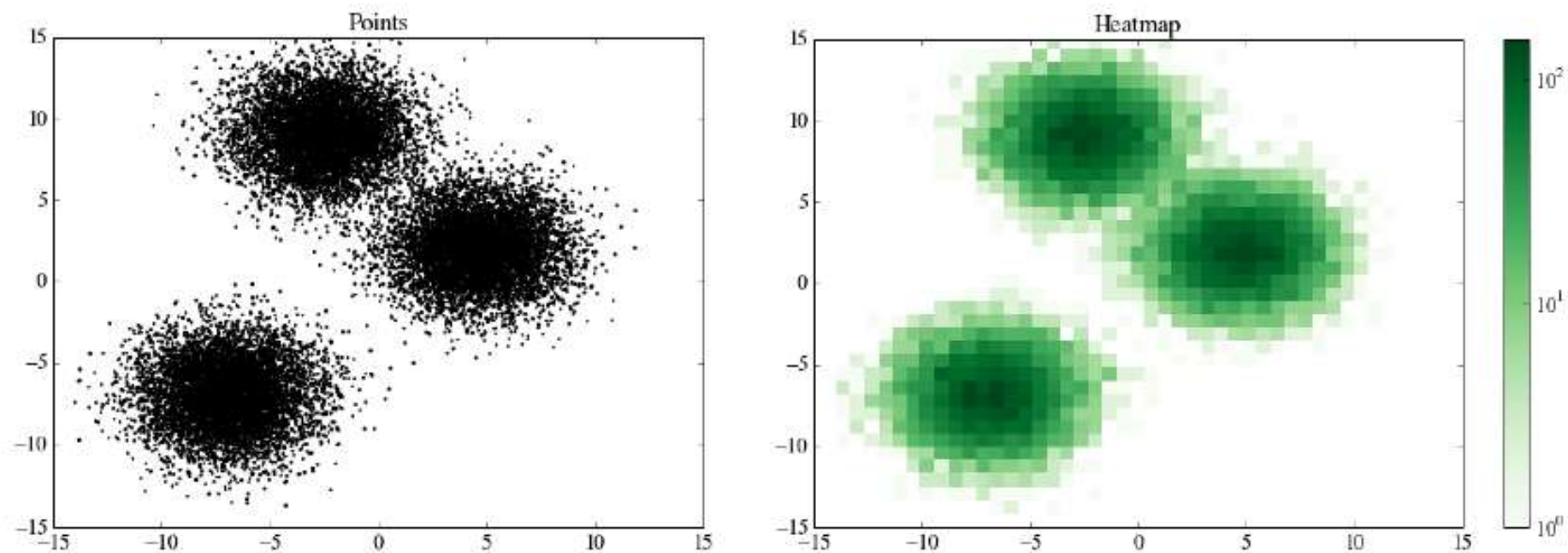
- Reducing size of data objects
- Removing all fill color from data objects
- Changing the shape of data objects
- Jittering data objects
- Making data objects transparent
- Encoding the density of values
- Reducing the number of values
 - Aggregating the data
 - Filtering the data
 - Breaking the data into a series of separate graphs
 - Statistically sampling the data

Trend Lines



Heat Maps

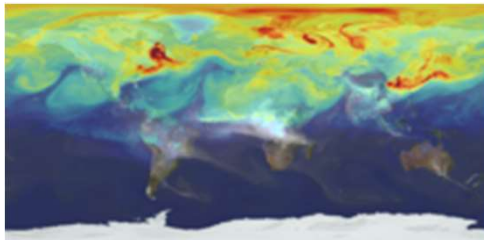
A heat map is a **two-dimensional representation of data in which values are represented by colors.**



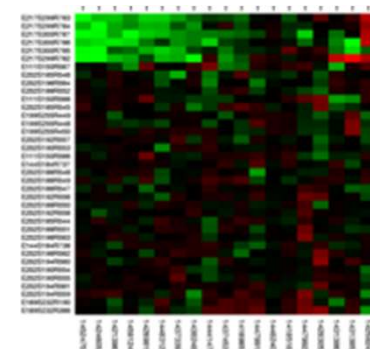
Binning of scatterplots

Heat Maps - Types

- There are two main type of heat maps: spatial, and grid.
 - A **spatial heat map** displays the magnitude of a spatial phenomena as color, usually cast over a map.
 - A **grid heat map** displays magnitude as color in a two-dimensional matrix, with each dimension representing a category of trait and the color representing the magnitude of some measurement on the combined traits.



Spatial Heat Map Example:
Displays temperature across a world image with red being the highest and blue being the lowest degree in temperatures.



Grid Heat map generated from [DNA microarray](#) data reflecting [gene expression](#) values in several conditions

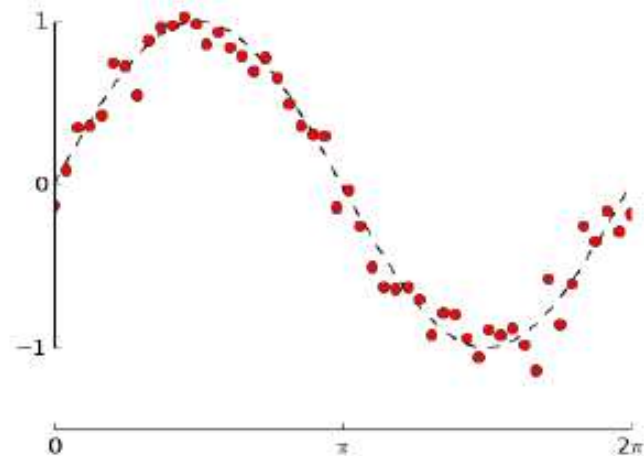
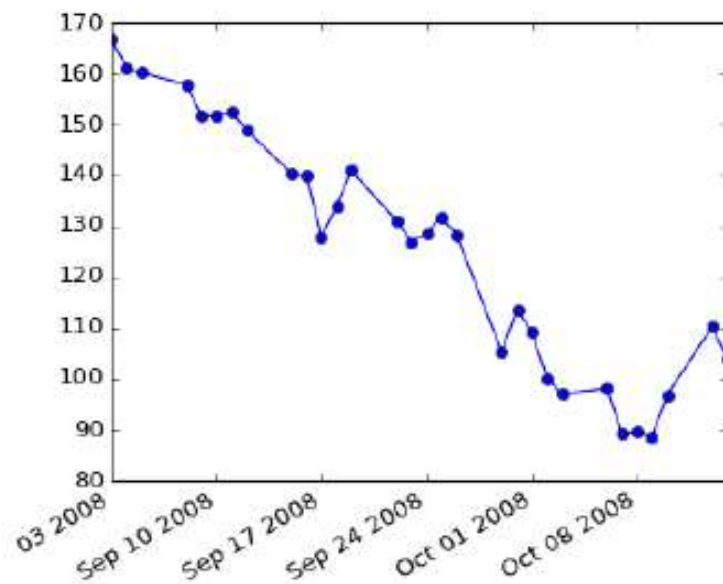
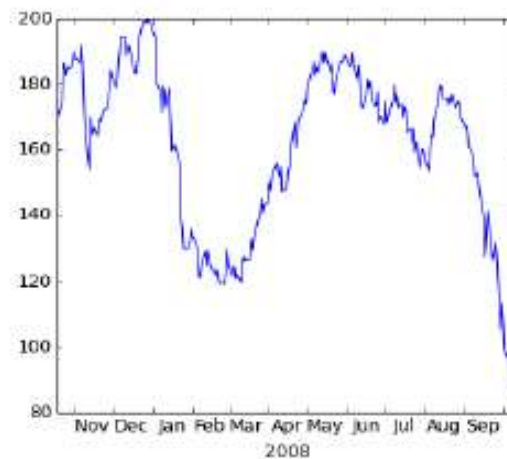
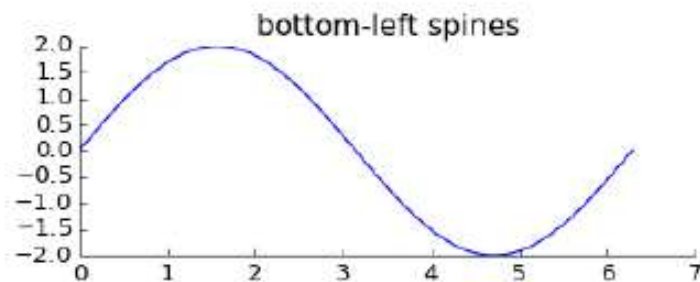
Visualizing Tables

- Introduction
- Comparisons
- Compositions
- Correlations
- **Time series data**
- Distributions

Time Series Data

- A **time series** is a series of data points listed in time order.
- Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.
- Time series data visualization can be used to show patterns like:
 - Trend
 - Variability
 - Rate of change
 - Covariation
 - Cycles
 - Exceptions

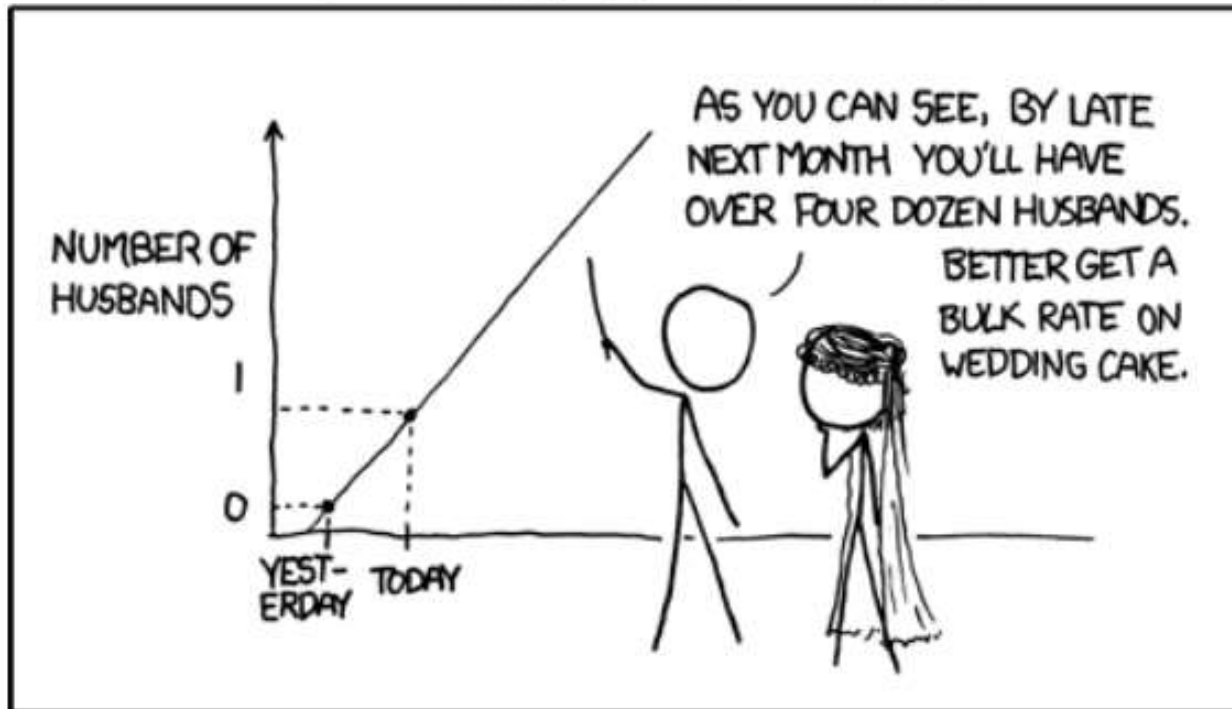
Line Graph



matplotlib gallery

Useful to show trends over time

Wrong Extrapolation!



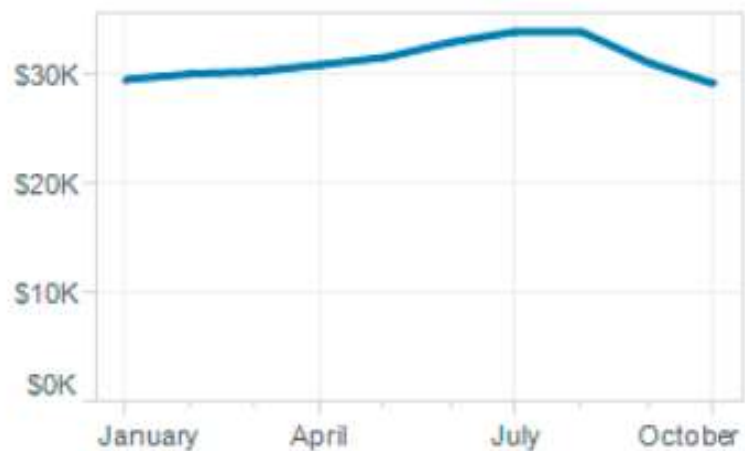
<http://xkcd.com/605/>

Bars vs. Lines

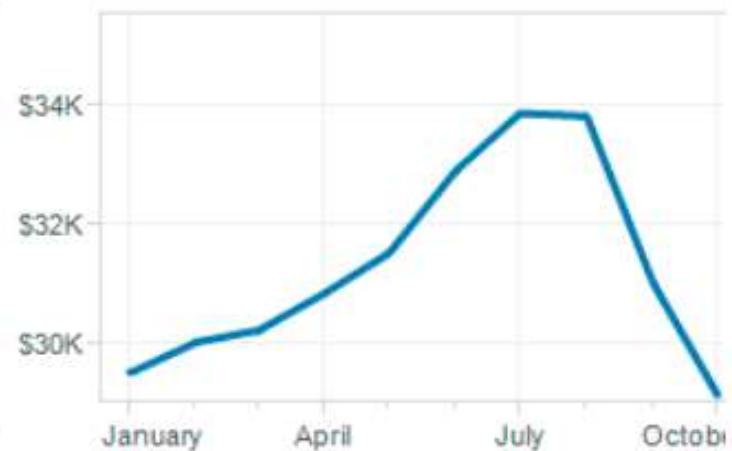
- Lines imply connections & sampling from continuous data.
 - Do not use for categorical data.
- Bars may not be useful for continuous data

Baseline Problem

True Baseline



Clipped Baseline

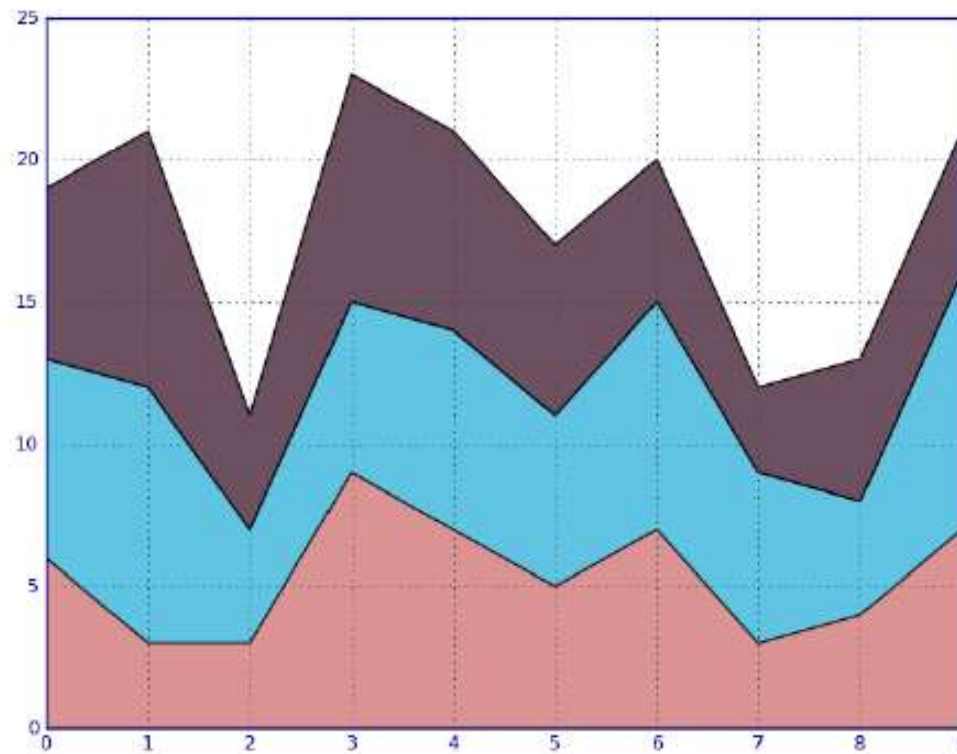


Linear vs Logarithmic Scale



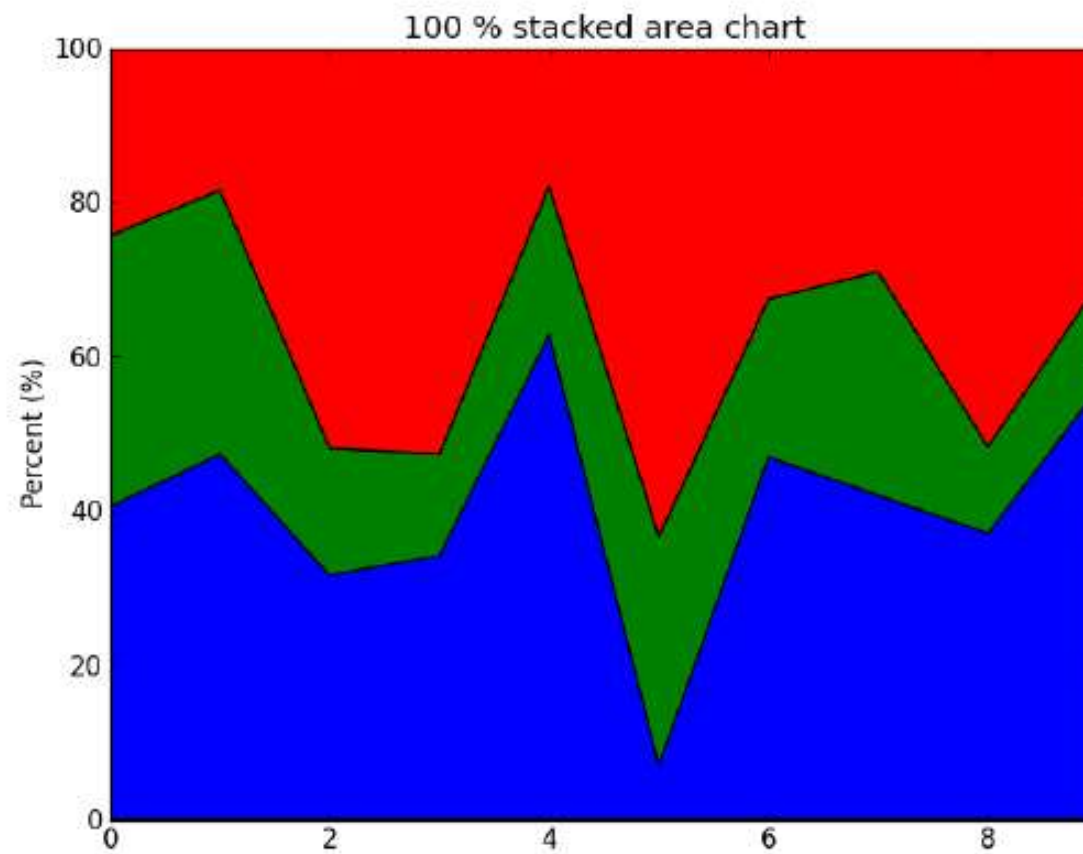
Apple Stock Price

Stacked Area Chart

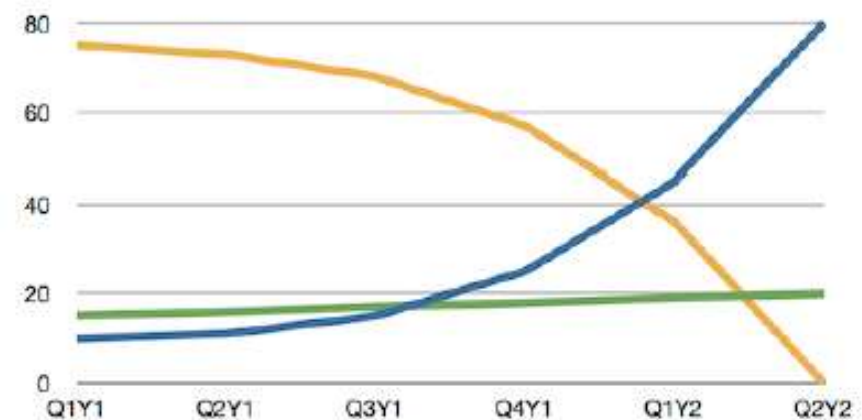
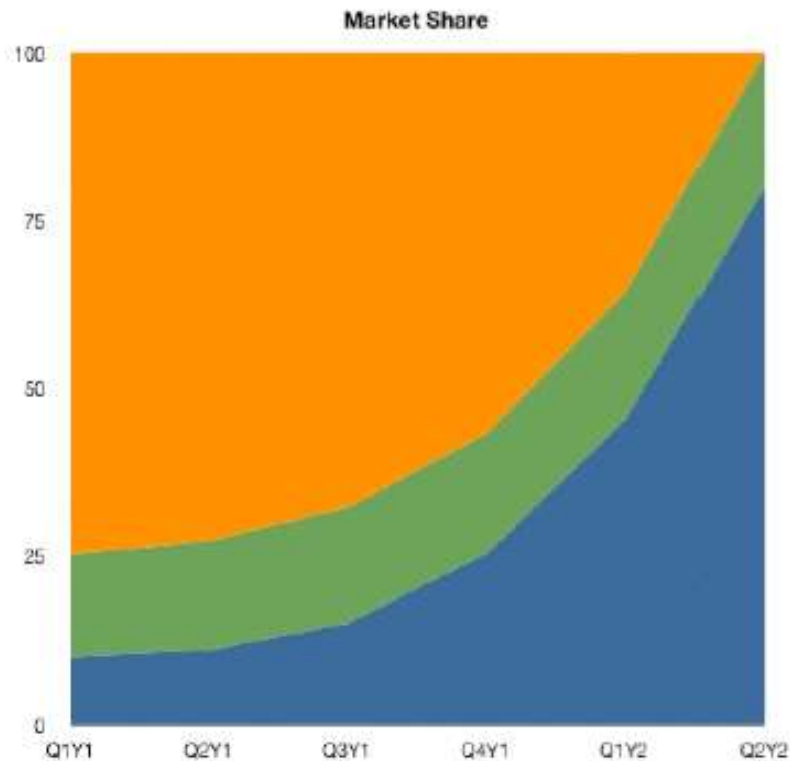


Good for showing trends over time among related attributes

100% Stacked Area Charts



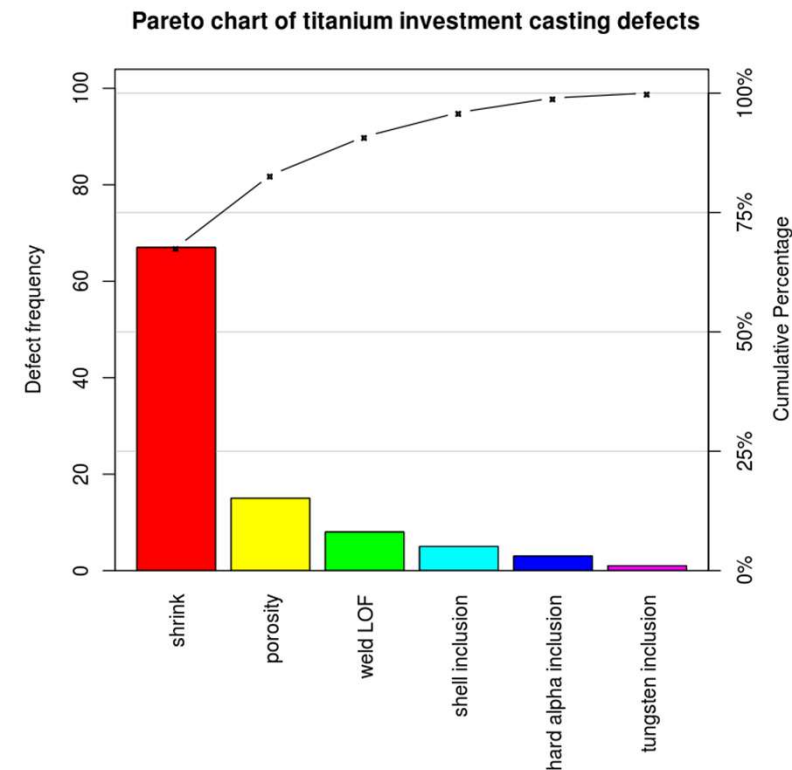
Stacked Area vs Line Charts



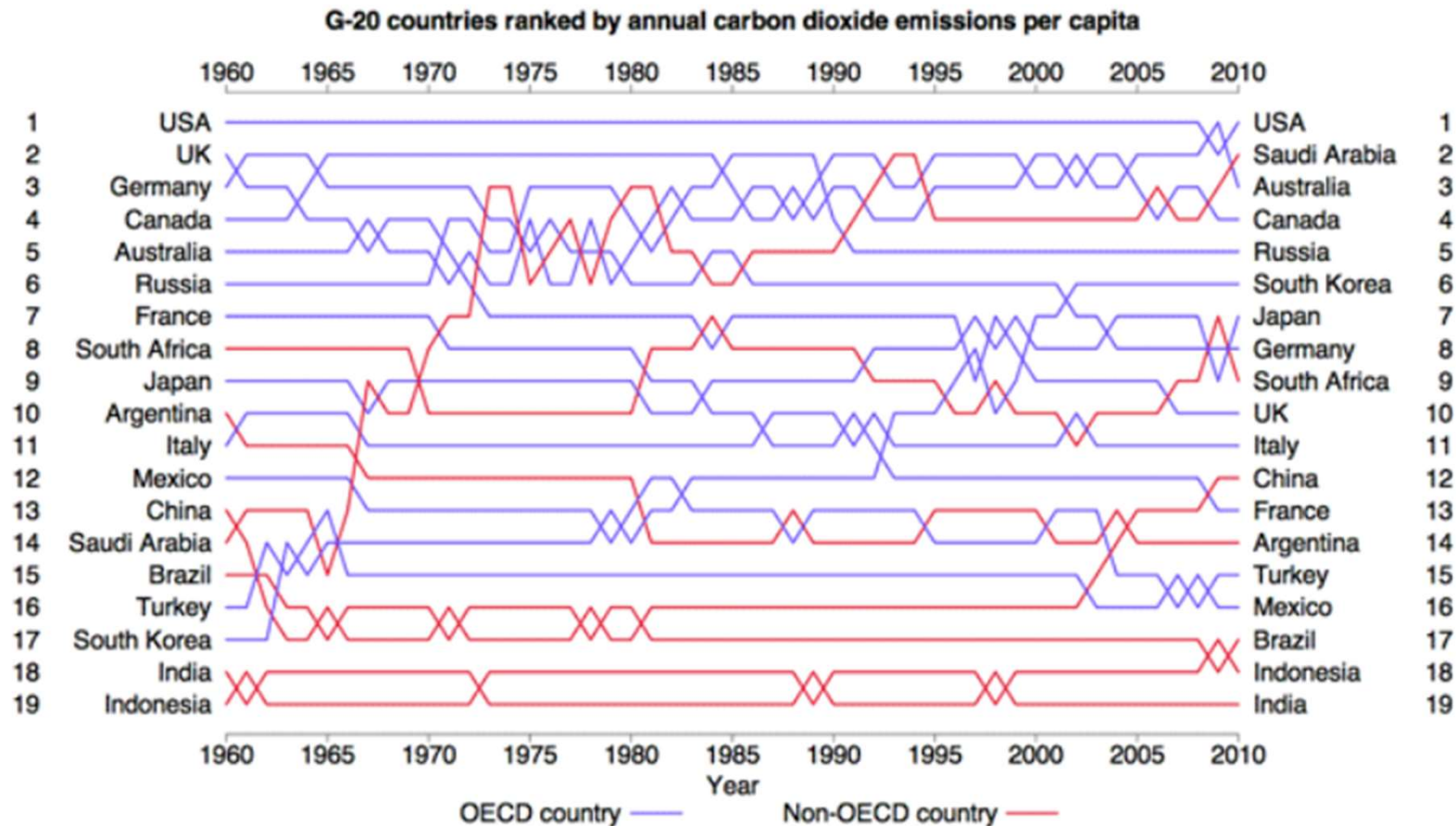
Difficult to figure out the market share of Green from the Stacked Area Chart

Pareto Chart

- A **Pareto chart** is a type of chart that contains both [bars](#) and a [line graph](#), where individual values are represented in descending order by bars, and the cumulative total is represented by the line.
- The left vertical axis is the [frequency of occurrence](#)
- The right vertical axis is the cumulative percentage of the total number of occurrences, or total of the particular unit of measure.
- The purpose of the Pareto chart is to highlight the most important among a (typically large) set of factors. In [quality control](#) Pareto charts are useful to find the defects to prioritize in order to observe the greatest overall improvement.



Bump Chart



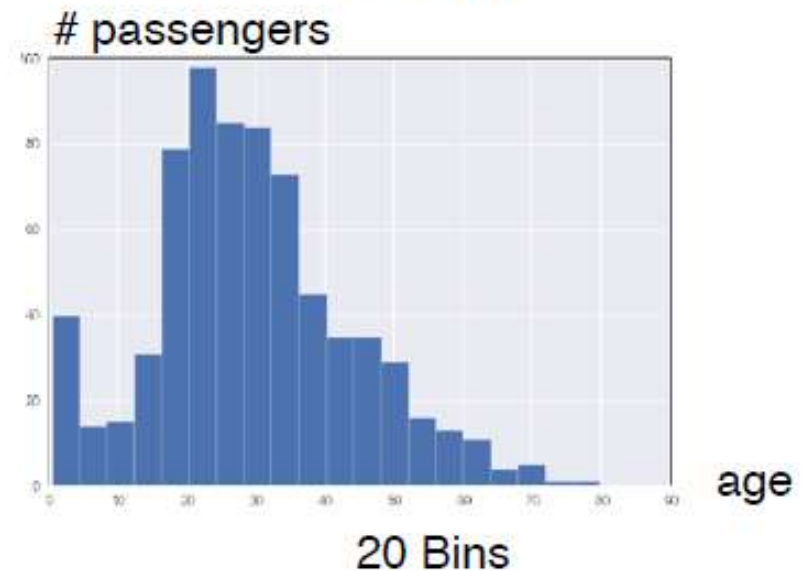
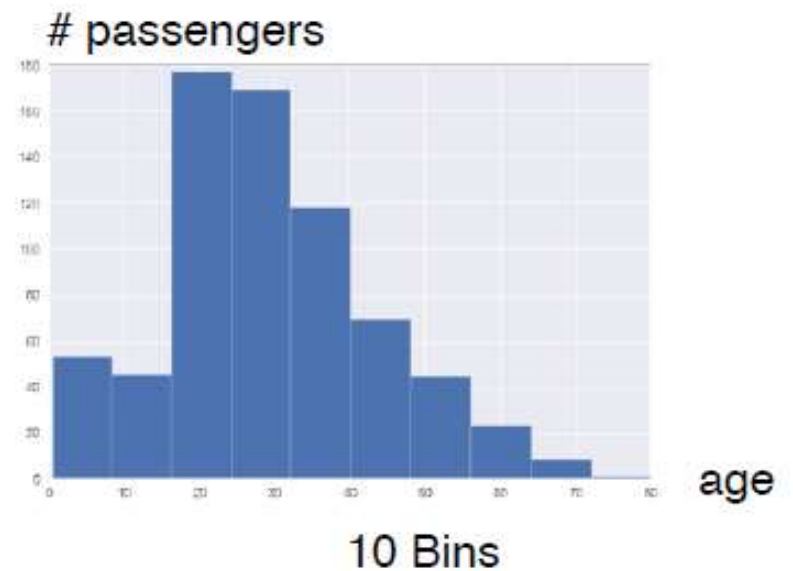
- A Bump Chart is used to compare two dimensions against each other using one of the Measure value.
- They are very useful for exploring the changes in Rank of a value over a time dimension or place dimension or some other dimension relevant to the analysis.

Visualizing Tables

- Introduction
- Comparisons
- Compositions
- Correlations
- Time series data
- **Distributions**

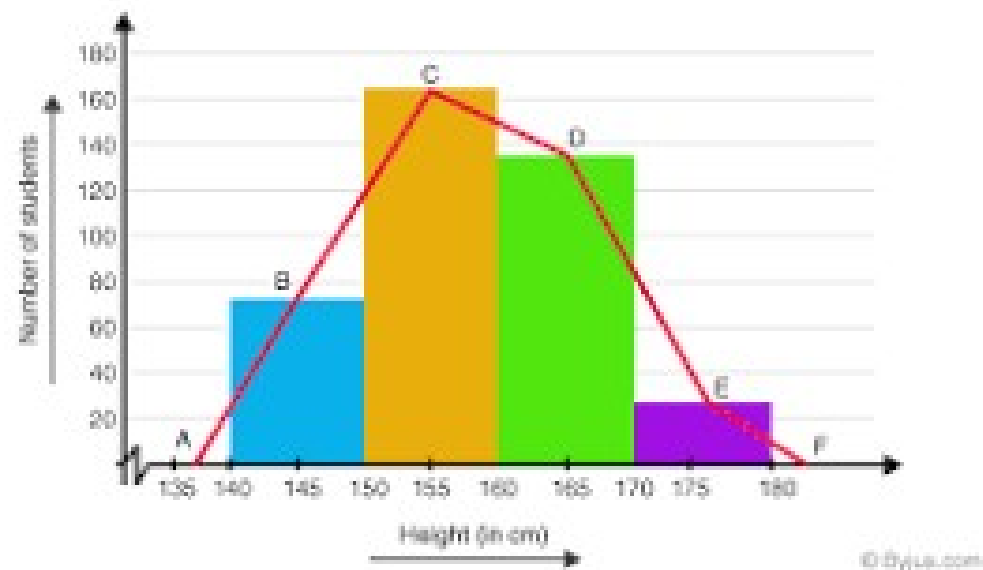
Histograms

- Good for showing distributions
- #bins hard to predict
- rule of thumb: $\#bins = \sqrt{n}$
- Make interactive!



Frequency Polygon

- A frequency polygon is a line graph of class frequency plotted against class midpoint.
- It can be obtained by joining the midpoints of the tops of the rectangles in the histogram



Stem-and-Leaf Plot

- A stem-and-leaf display or stem-and-leaf plot is a device for presenting [quantitative data](#) in a [graphical](#) format, similar to a [histogram](#), to assist in visualizing the [shape](#) of a [distribution](#).
- Unlike histograms, stem-and-leaf displays retain the original data to at least two significant digits, and put the data in order

```

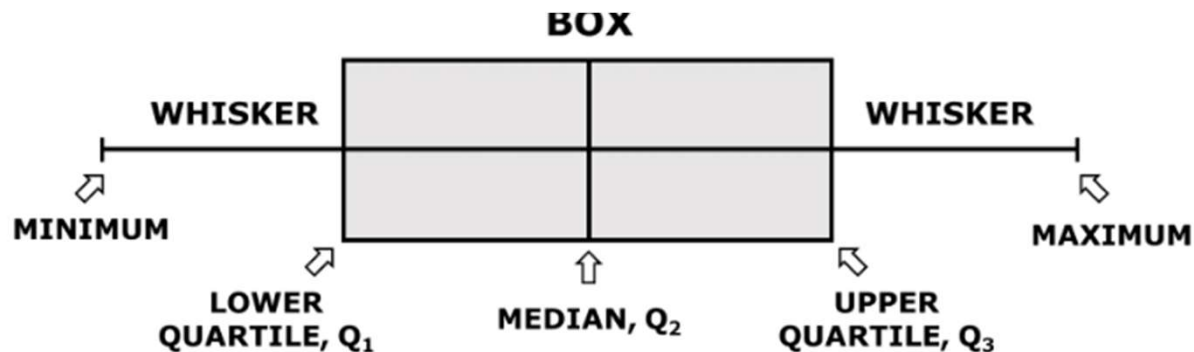
2 | 0
3 | 025
4 | 11378
5 | 133467889
6 | 024559
7 | 147
8 | 8
9 |
10 | 2

```

A stem-and-leaf plot of the values 20, 30, 32, 35, 41, 41, 43, 47, 48, 51, 53, 53, 54, 56, 57, 58, 58, 59, 60, 62, 64, 65, 65, 69, 71, 74, 77, 88 and 102

Box Plot

A box plot) is **method of displaying data**, invented by J. Tukey. A visual display that summarizes data using a 'box and whiskers' format to show **the minimum and maximum values (ends of the whiskers)**, **interquartile range (length of the box)**, and **median (line through the box)**



Reading

- S. Few, **Effectively Communicating Numbers - Selecting the Best Means and Manner of Display.**

http://www.perceptualedge.com/articles/Whitepapers/Communicating_Numbers.pdf