Lecture 3: Data Encoding (1)

Statistical Charts: what would you like to show
* Comparison
* Relationship
* Composition
* Distribution

Comparison: How to compare two or more attributes?
* The most basic way is **bar chart**: For comparing categorical attributes
* **Line chart:** for showing trends over time
    * Lines imply for connections. Do not use for categorical data
    * The line suggests that there are data points in between.
    * Logarithmic scale: the biggest risk is that someone is not able to interpret, so only use logarithmic scale where you would just see the outlier and the rest of the data would be very small and very compressed.

Relationship: How to show relationships between attributes?
* if you have two attributes, the most simple thing to do is to pick a **scatterplot**: Use spatial position to encode values of two quantitative attributes
    * Problem: Over-plotting, too many items can result in visual clutter.
    * Solutions
        * Decrease opacity
        * Change scale
        * Increase screen size/resolution
        * Sample data
    * Problem with the transparency solution: losing the outliers
        * Logarithmic transparency
        * Data mapping onto the visual channel
        * Pre-processing the outlier, and save them in some different way

Composition: How to show composition of an attribute?
* **Pie chart** vs. **stacked bar chart**
    * Pie charts encode the data using angle and area.
    * Bar charts encode the data using the length of a line.
    * Both of them are not very effective
    * Stacked bar charts are usually more effective than pie charts.
    * More people have a preference for pie charts like a visual preference
    * Pie chart could only show the composition information about one item, but stacked bar chart can do it for multiple items at the same time
* Donut Chart
    * Even harder to interpret than regular pie charts.
    * It makes the judgment of angles harder because you only see a tiny fraction of the angle and it is harder to estimate that
* Stacked bar chart, Layered Bar Chart and Grouped Bar Chart
    * Example: different students(items) and how they perform for different questions(category)
        * If you want to compare the overall performance of students, you would use the **stack bar chart** because the most important information you want to bring across is the sum of the different points
            * But it is hard to compare the performance of students in one individual category

    * If you want to compare the performance of students in each questions, **layered bar chart** is more appropriate because you have the same baseline for each category
    * If you wan to look at which questions the students performed the best, than maybe a **group bar chart** is the best choice

Composition: How to show compositions over time?

* If you have a composition that changes over time, then you can use so-called **stacked area chart**
  * Version 1: stacking the data on top of each other
  * Version 2: show the proportions of hundred percent. Could be used to too the market share
* Stacked area vs. Line graph

Statistical Charts: How to show distributions?

* **Histograms**: Binning can hide patterns.
* **Density plot**
  * Difference between the density plot and scatterplot:
    * Density plot: accumulation of the actual raw data
    * Scatterplot: raw data collected as individual points
* **Box plot/Box-and-whiskers Plot**:
  * Vertical line defines the median
  * Box contains 50% of the data
  * Whiskers can be defined in various ways
    * Going to the minimum or the maximum
    * Q1 - Q3
* **Bar Charts with Confidence Intervals** vs. **Box Plots**
  * Bar charts are appropriate for counts
    * But it might touch the parts of the skew that where you might not even have a single data item
  * Box plots should be used to represent the chara`cteristics of a distribution.
* Violin Plot = Boxplot + Probability Density Function
  * Advantage: you see a little bit more of the actual shape of the curve that could be hidden by the box
  * Boxplot is reducing information to median, Q1, Q2, Q3 and the extent of the whiskers

Multi-Dimensional Data

* There are two keys in Multi-Dimensional table
* Multiple slices of that cube for different points in time where you measure something

Univariate Data

* Create histogram

Bivariate Data

* create scatterplot

Trivariate Data

* Rule of thumb: should not use 3D visualisation if you don’t have a spatial phenomenon
  * For the three dimensional protein structure, we could use three dimension visualisation
  * For abstract data, three dimension is not capable
* Alternative: map the third attribute to some other visual channel
  * Three different hues

Heat map:

* used where you have matrix data, two keys and one value for this combination
* Maps attribute value to color
* Matrix data: two keys, one value
  * Example: gene expression dataset
    * [key1: patient] x [key2: gene]
    * value = activity of gene
    * Each cell is a "pixel", value encoded using color
* Meaning derived from ordering
  * Order by time, frequency, etc.
  * If no ordering inherent, clustering is used
* Strengths
  * Scales to large matrices
  * Good for homogeneous data
  * All columns share same semantics

Scatterplot Matrix (SPLOM)

* When there are more than two quantitative attributes, like N attributes
* N^2 scatterplots
* Every pairwise plot is shown twice
* Symmetry along diagonal
* Diagonal can be used
  * for showing labels
  * for showing histograms of attributes
* Downsides:
  * Scalability: for each of the attributes, you need to get a complete row and complete column for plots

Spatial Axis Orientation:

* Rectilinear layouts
  * Scatterplots, bar charts, etc.
* Parallel layouts
  * Parallel Coordinates
* Radial layouts
  * Radial Bar Chart

Parallel Coordinates:

* Axes represent attributes
* Lines connecting axes represent items
* Limitation:
  * Scalability too Many Dimensions
  * Scalability too Many Items
    * Solution:
      * Reduce opacity
      * Increase the transparency
      * Bundling, clustering
      * Sampling

  * Correlations Only Between Adjacent Axes
    * Solution:
      * Let the user interactively change the order
      * Coming up with an ideal order or when a meaningful order
      * Brushing: let the user select a couple of lines
  * Ambiguity
    * Solution:
      * Interactive highlighting
      * Curves
  * Hard to work with missing data
  * Only work well with quantitative data

Radial layouts

* all exits meet at the centre point of the coordinate of the coordinate system
* Strength:
  * one continuous line between years
  * Visualising cyclic patterns
* Downsides
  * Distortion: the part of the data that is very close to the centre of the visualisation is very compressed
* **Superimposed line chart**: put the curves top pf each other in the same coordinate system on the same scale
  * Advantage: peaks can be compared