Title: Histogram design
Christoph Heinzl, Bernhard Fröhler
Type: PR/DA
Persons: 1-3
Workgroup: Visualization Group

Description

Histograms are often used as the first method to gain a quick overview over the statistical distribution of a collection of values, such as the pixel intensities in an image.
Depending for example on the datatype of the underlying data (categorical, ordinal or continuous) and the number of data values that are available, several visualization parameters can be considered in constructing a histogram:

- **Bin width** if a too large bin width is chosen, finer details in the distribution of the data are lost; if a too small bin width is chosen, the histogram might show sampling artefacts, or in the extreme case, when each data value is assigned its own bin, will look like a step function).
- **Aspect ratio** (this is the ratio between the width and the height). Depending on what aspect ratio is chosen, the perception of the slope of the histogram curve might be affected
- Chosen **visual encoding** (bar chart, line graph, ...)
    - Bar graph specific: How much space should there be between the single bars?
    - Line graph specific: How to handle borders (The line graph would typically have a vertex with the data value at the middle of each bin. Should the graph be cut off at beginning and end at the middle bin? Or should the line be continued to the edge of the bin at the same value as the first/last bin, or should the line be continued down to a value of zero (down to the x-axis)?)
- **Tick mark** positions & count: How many ticks to place? How to make sure they don't overlap? Where to place the labels exactly? How to make sure the labels are not too small / large? (Optional: all previous questions for tick marks for logarithmically scaled axis)

The perception of a histogram might vary quite a bit depending on the exact parameters chosen, and this might also influence the interpretation. On some of the above points, you should be able to find literature already.

Tasks

- Create a web application (e.g. in d3) that allows to enter data in a tabular format, and creates different histograms based on these values.
- At least the parameters mentioned above should be adaptable by the user.
- Search for rules for determining above parameters automatically from the data, and implement a few
- Research the variety of tasks that histograms are used for, for instance understanding distributions, filtering of data, finding modes in distribution (number and count)
- Evaluate the different encodings regarding their effect on the found tasks

Requirements

- Good skills in C++ and software engineering
- Good knowledge of English language (source code comments and final report will be in English)
- Interest in data visualization

Environment

The project should be implemented as a part of the open_iA framework
(https://github.com/3dct/open_iA)

Starting literature

Talbot et al., 2010: An Extension of Wilkinson’s Algorithm for Positioning Tick Labels on Axes

Contact

For more information please contact: Christoph Heinzl (christoph.heinzl@fh-wels.at) or Bernhard Fröhler (bernhard.froehler@fh-wels.at).