Big data exploration with tabplot

Martijn Tennekes[1*], Edwin de Jonge[1]

1. Statistics Netherlands
*Contact author: m.tennekes@cbs.nl

Keywords: Big data, visualization

The tableplot is an innovative visualization to explore large datasets (Tennekes et al., 2013). A tableplot is created by 1) sorting the data, 2) binning the data, 3) calculating mean values or category fractions, 4) visualizing it column-wise by a bar chart of mean values and a stacked bar chart of category fractions. The tableplot has been implemented in the *R* package **tabplot**. It includes a graphical user interface that uses the **shiny** package as well as the *javascript* **d3** package.

The **tabplot** package was introduced at the useR 2012 by a poster presentation. Recent developments include the increase of processing speed, which is in particular needed for the interactive interface, and the visualization of high cardinality categorical data. Furthermore, the **tabplot** package has been applied on two big data sources at Statistics Netherlands. The Dutch Virtual Census contains demographic information like age, gender, and household status about all 16,5 million Dutch inhabitants. The Dutch Labour and Benefits Database is a data source of all salaries and social benefits of the Dutch population, and contains about 100 million records on an annual basis.

The processing time consists of two major parts: sorting and aggregating. Both parts are implemented using the *R* package **ffbase** which uses *C* code. To increase the interactive performance, the sorting part is executed only once as a data preparation step. For each variable, the order of values is determined and stored. With these sorting orders, aggregation is fast, independent of the number of row bins. This makes it possible to explore datasets in a fast, interactive way. When even more speed is required, a uniform sample of the ordered data is drawn, aggregated, and plotted. Visualizing a sample is sufficient for initial data exploration. When a deeper look at the data is required, a larger sample and eventually the full dataset is processed.

The case studies at Statistics Netherlands emphasized the importance of visualizing categorical data. It is often worthwhile to categorize numeric variables, because this provides a better understanding of the data distributions per bin, as well as on the distribution of missing values along the bins. The core function `tableplot` offers flexibility to specify color palettes for categorical data. High cardinality categorical variables are visualized by grouping the categories, either by a specific aggregation scheme at hand or by merging neighboring categories uniformly so that the merged categories represent an equal number of original categories.

References

Tennekes, M., Jonge, E. de, Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots, *Journal of Data Science 11 (1)*, 43–58.