Big data exploration with tabplot

Martijn Tennekes, Edwin de Jonge

July 11, 2013

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:
     - numeric variable: mean value
     - categorical variable: category fractions
  4. Plot:
     - numeric variable: bar chart
     - categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:
     - numeric variable: mean value
     - categorical variable: category fractions
  4. Plot:
     - numeric variable: bar chart
     - categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:
     - numeric variable: mean value
     - categorical variable: category fractions
  4. Plot:
     - numeric variable: bar chart
     - categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value

     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart

     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value

     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart

     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value

     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart

     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value

     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart

     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value

     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart

     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value
     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart
     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value
     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart
     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value
     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart
     categorical variable: stacked bar chart
- Implementation: R package tabplot

Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value
     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart
     categorical variable: stacked bar chart
- Implementation: R package tabplot

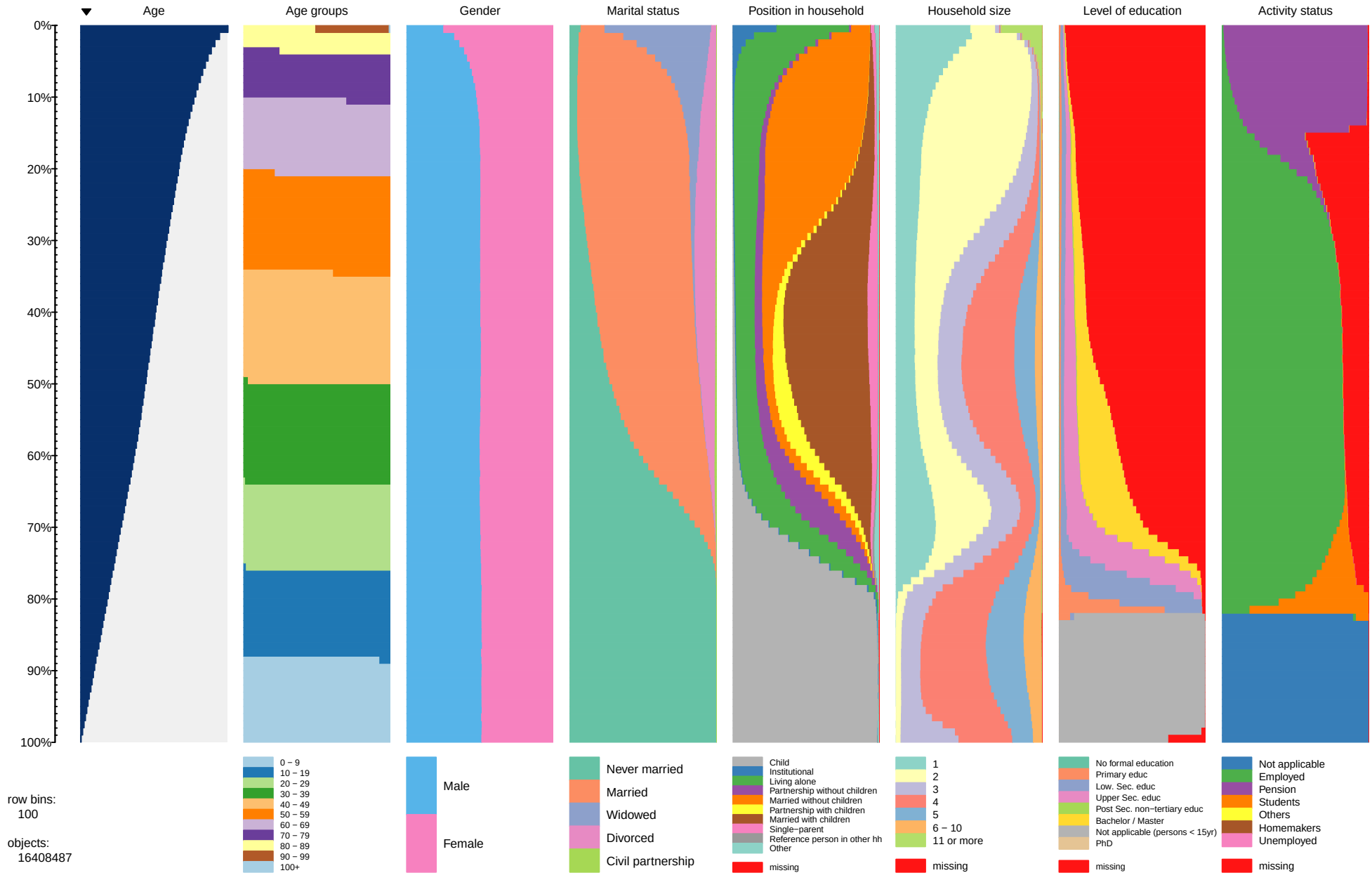
Tableplot

- Innovative data visualisation method
- One picture of a multivariate (big) data source
- Bottom up method:
  1. Sort records on a key variable
  2. Group records into equally sized bins
  3. Calculate per group:

     numeric variable: mean value
     categorical variable: category fractions
  4. Plot:

     numeric variable: bar chart
     categorical variable: stacked bar chart
- Implementation: R package tabplot

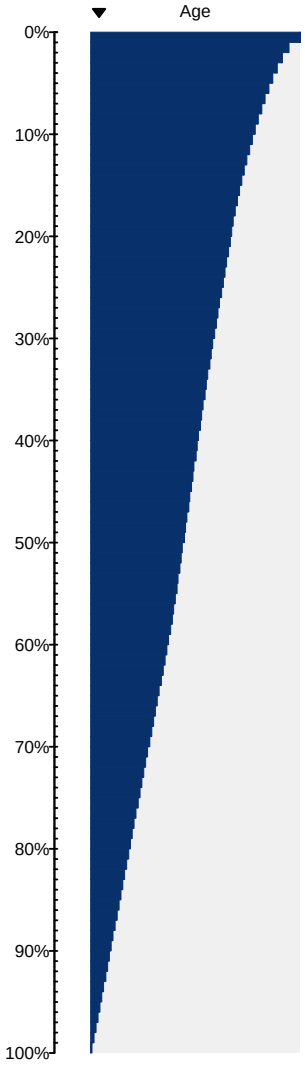
Tableplot: Virtual Census



row bins:
100

objects:
16408487

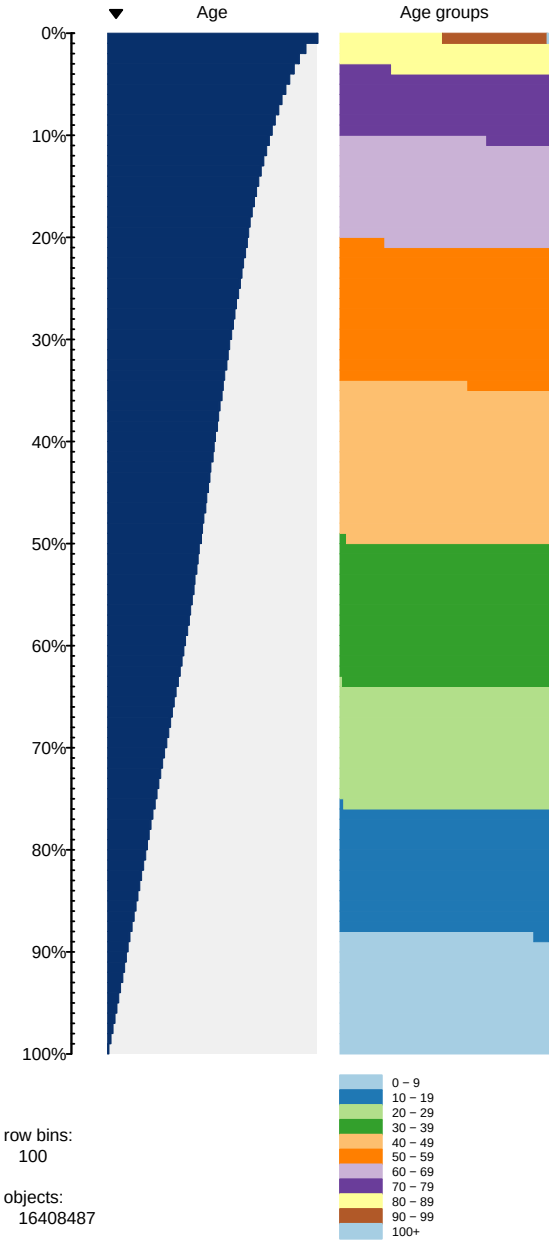
Tableplot: Virtual Census



row bins:
100

objects:
16408487

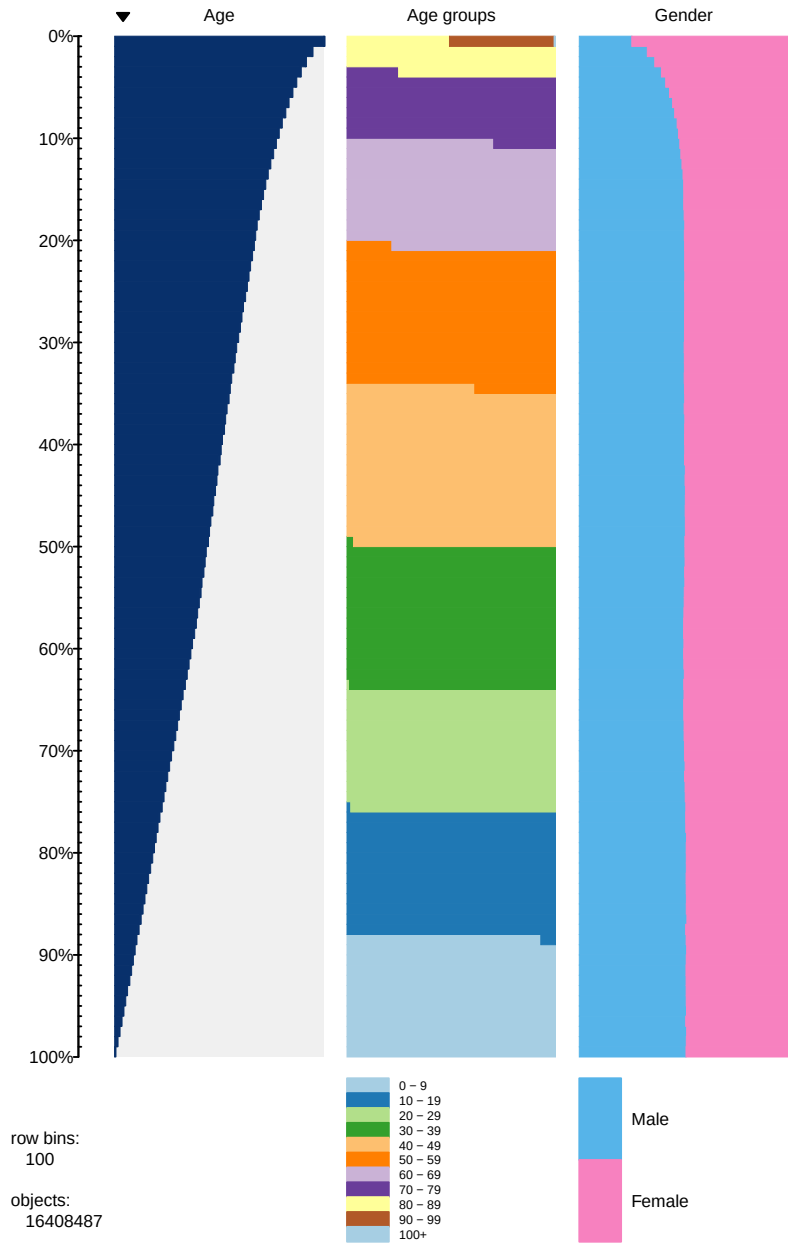
Tableplot: Virtual Census



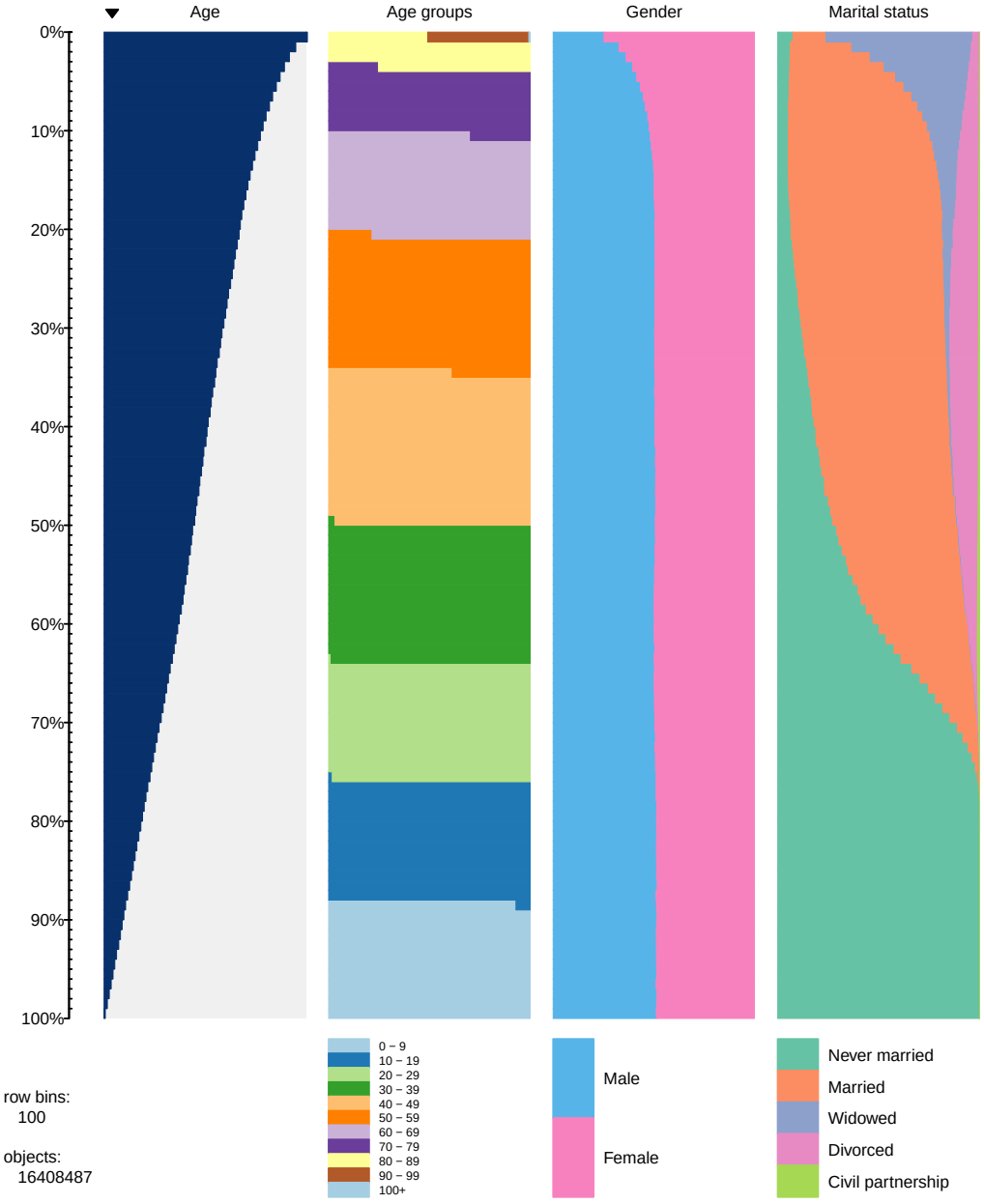
row bins:
100

objects:
16408487

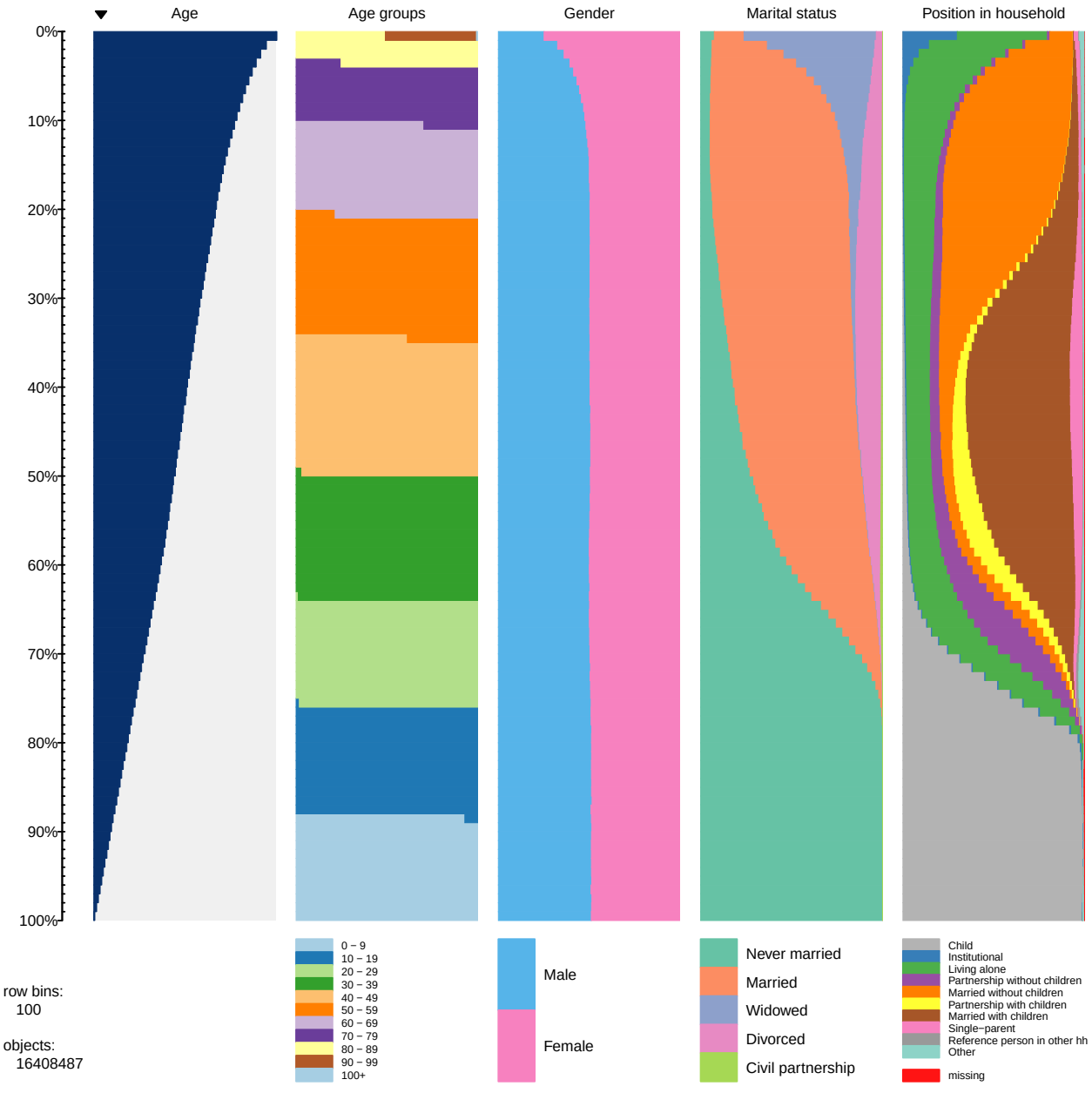
Tableplot: Virtual Census



Tableplot: Virtual Census



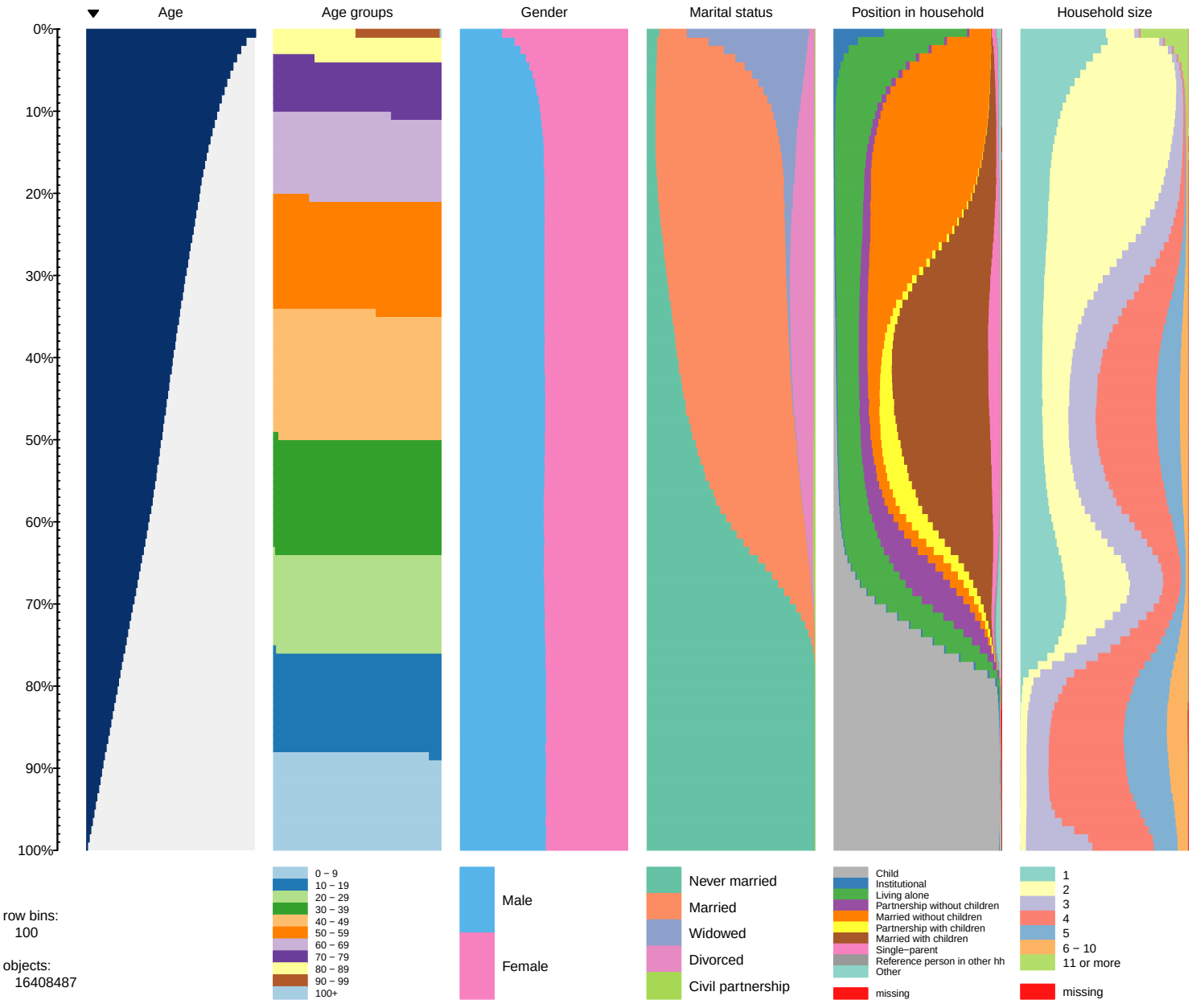
Tableplot: Virtual Census



row bins:
100

objects:
16408487

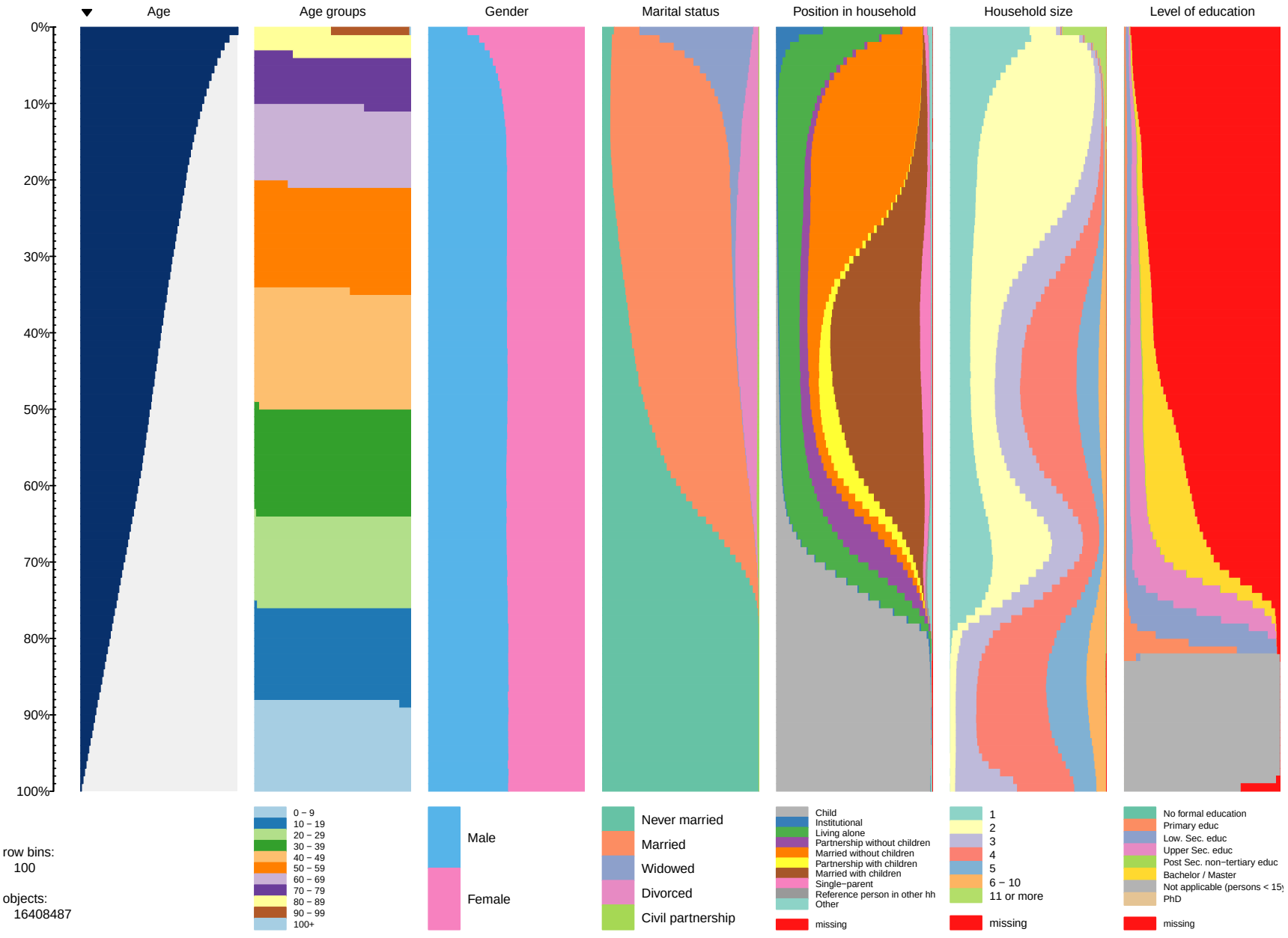
Tableplot: Virtual Census



row bins:
100

objects:
16408487

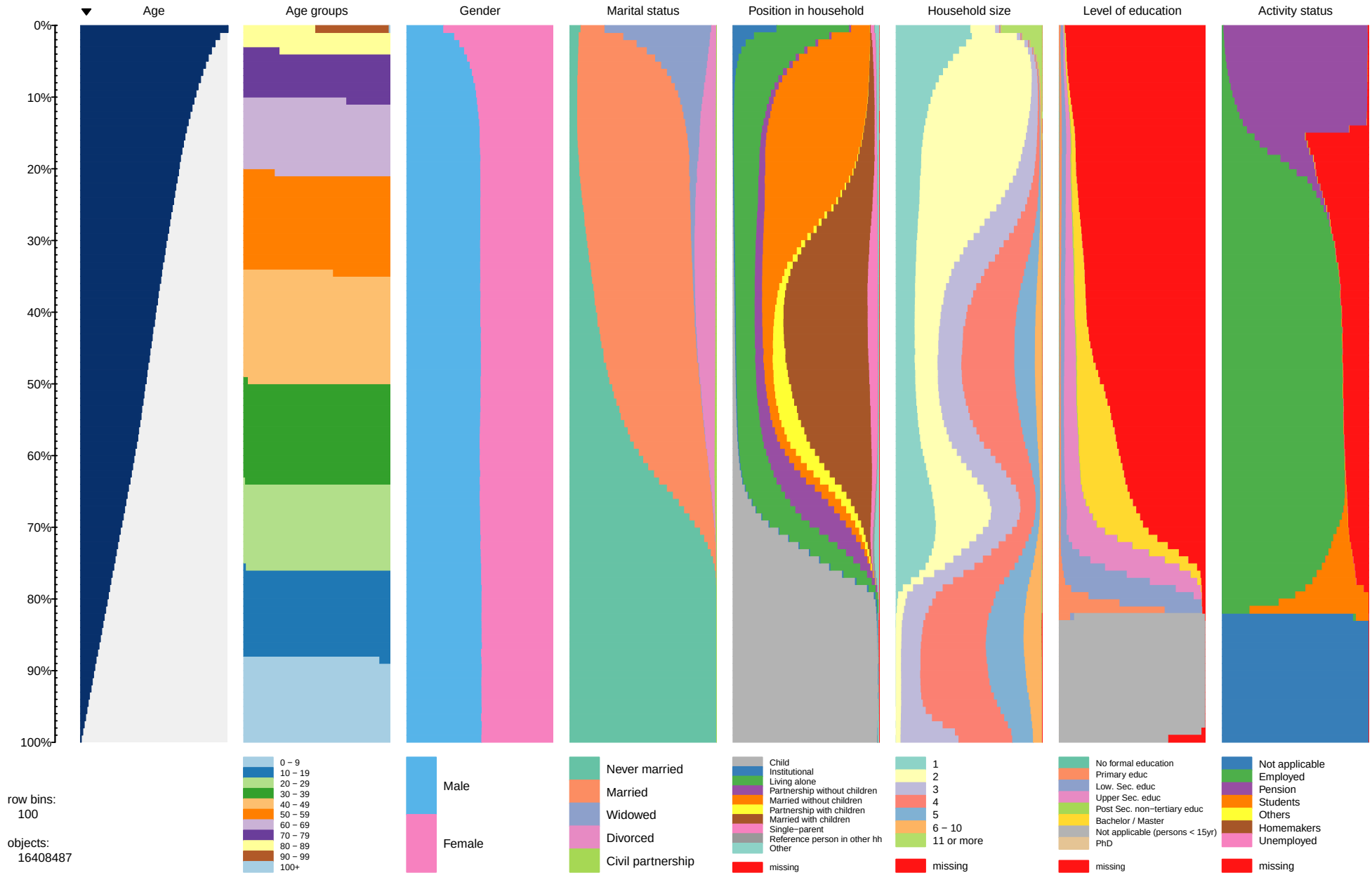
Tableplot: Virtual Census



row bins:
100

objects:
16408487

Tableplot: Virtual Census



row bins:
100

objects:
16408487

R package tabplot

- Core function: tableplot

  Main arguments:

dat	data.frame
select	columns
sortCol	sorted column
nBins	number of bins

- Shiny interface: itableplot

R package tabplot

- Core function: tableplot
  Main arguments:

dat	data.frame
select	columns
sortCol	sorted column
nBins	number of bins

- Shiny interface: itableplot

R package tabplot

- Core function: tableplot
  Main arguments:

dat	data.frame
select	columns
sortCol	sorted column
nBins	number of bins

- Shiny interface: itableplot

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared ← tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared ← tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Performance

- Under the engine: ffbase
- Data preprocessing (only once):
  - per column, the rank order is determined
  - prepared $\leftarrow$ tablePrepare(dat)
- Interactive tableplotting:
  - tableplot(prepared)
  - tableplot(prepared, sortCol=x, nBins=200)
  - Still too slow? Sample with argument maxN

Scales and layout options

- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

Scales and layout options

- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

Scales and layout options

- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

Scales and layout options

- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

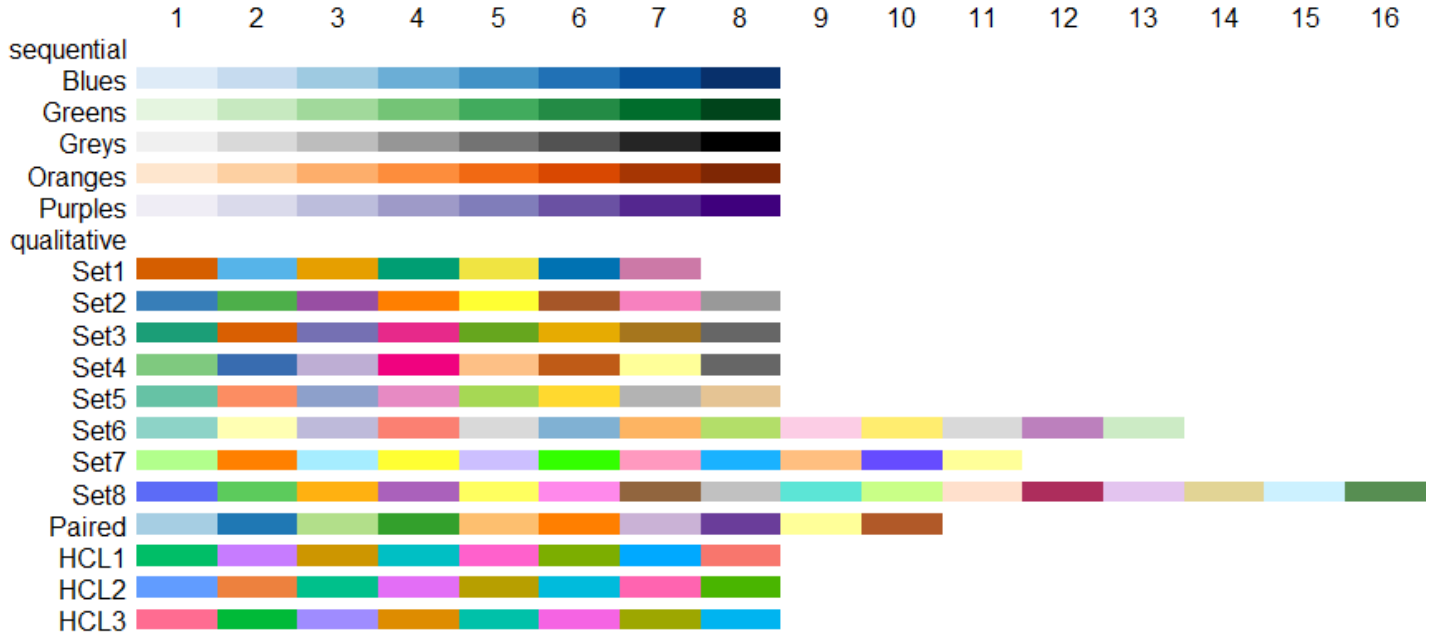
Scales and layout options

- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

Scales and layout options

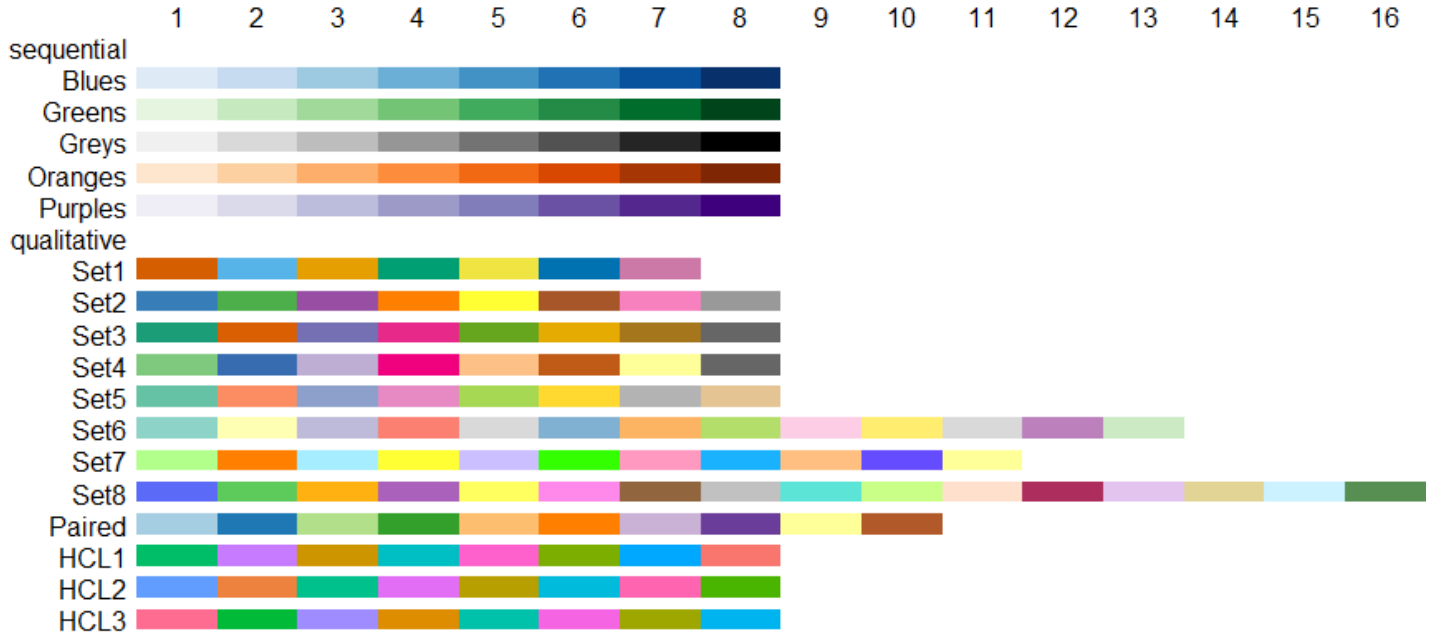
- tableplot(dat, ...)
- ...: arguments regarding scales and palettes
- tab $\leftarrow$ tableplot(dat)
- plot(tab, ...)
- ...: arguments regarding fontsize, size of legend, title, etc.

Categorical data



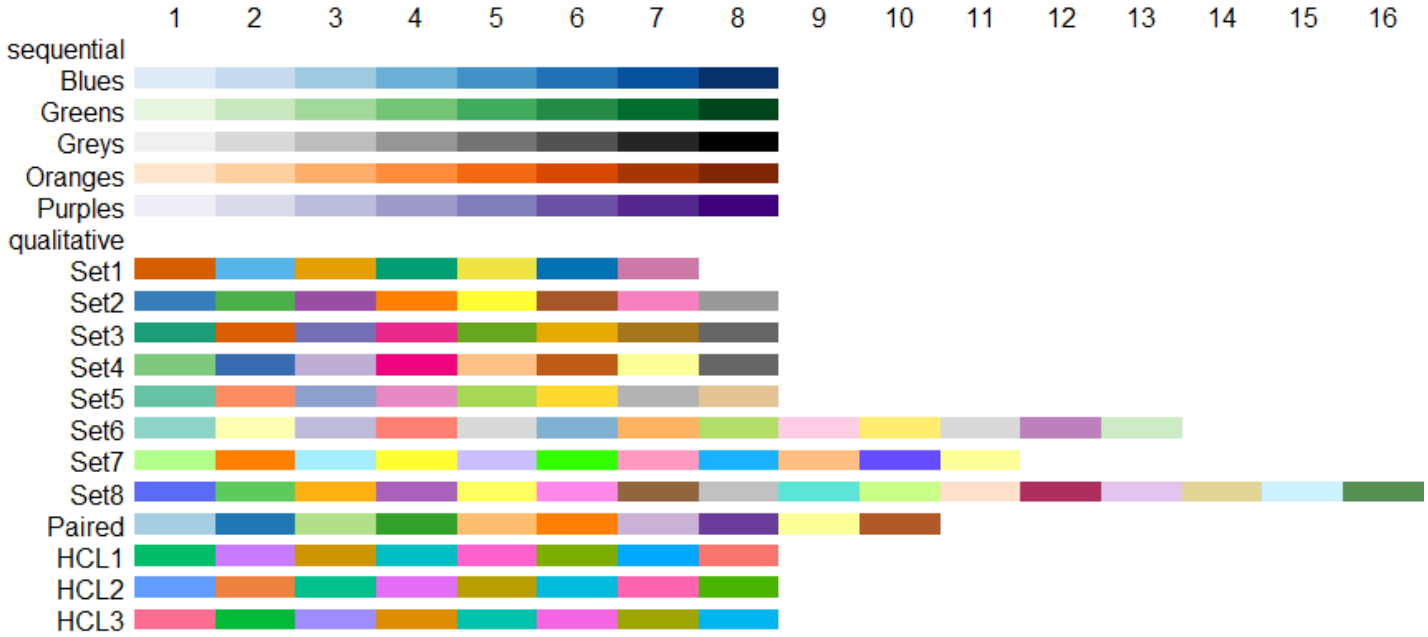
- Over 16 categories?
- Rainbow colour palette:

Categorical data



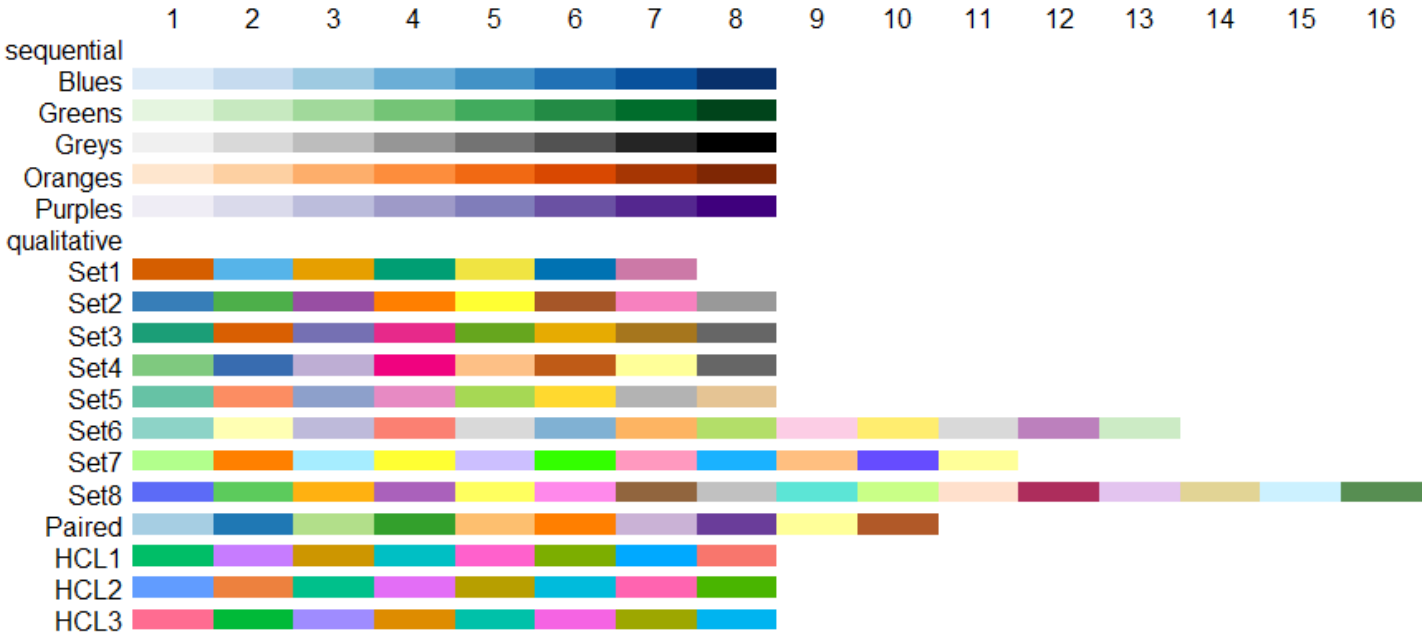
- Over 16 categories?
- Rainbow colour palette:

Categorical data

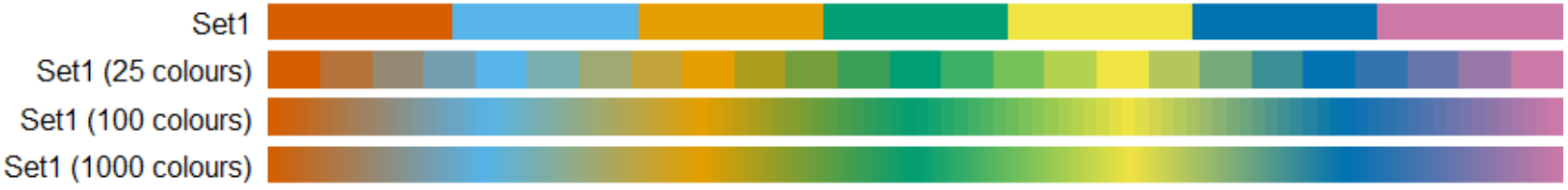


- Over 16 categories?
- Rainbow colour palette:

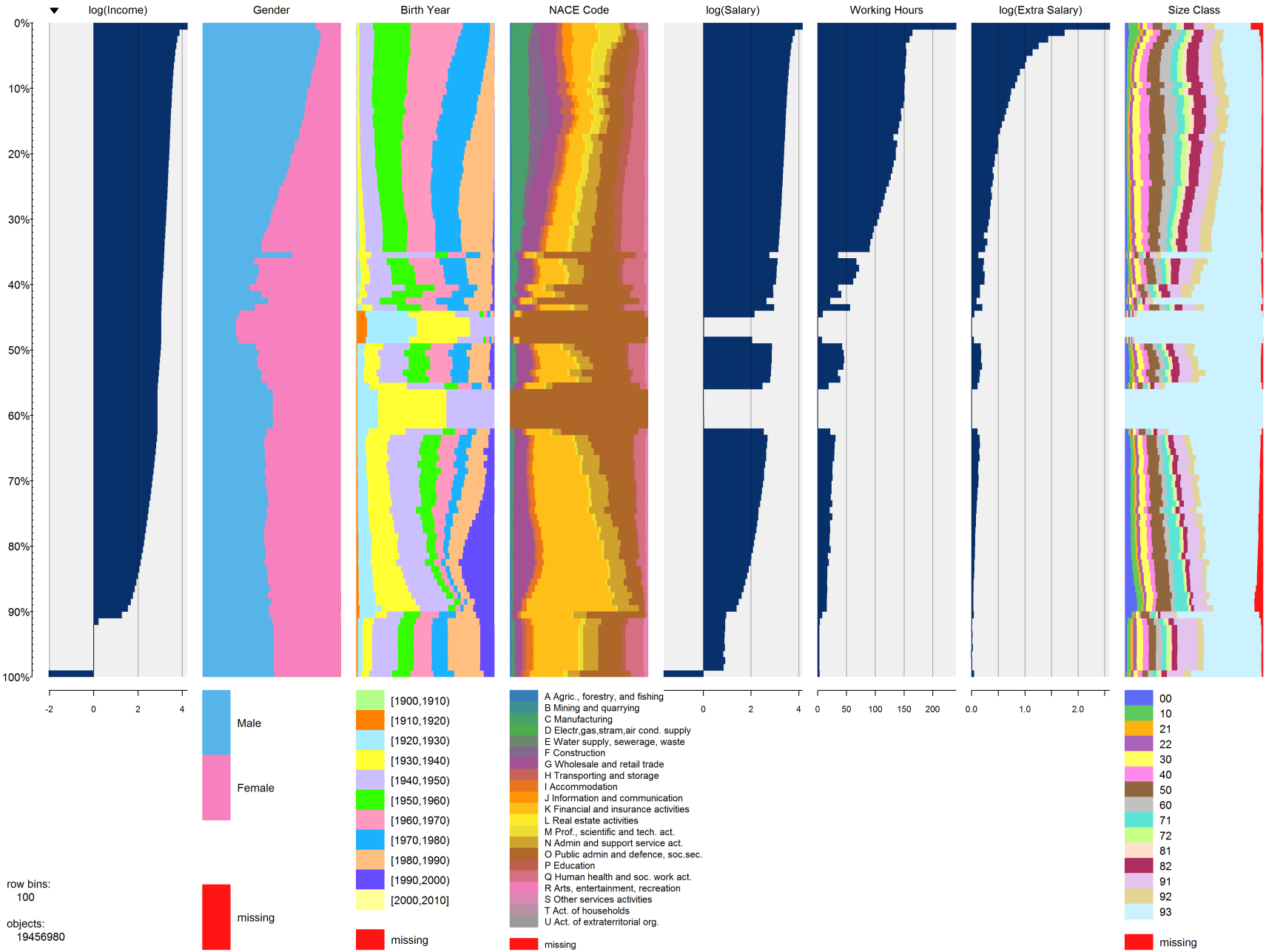
Categorical data



- Over 16 categories?
- Rainbow colour palette:



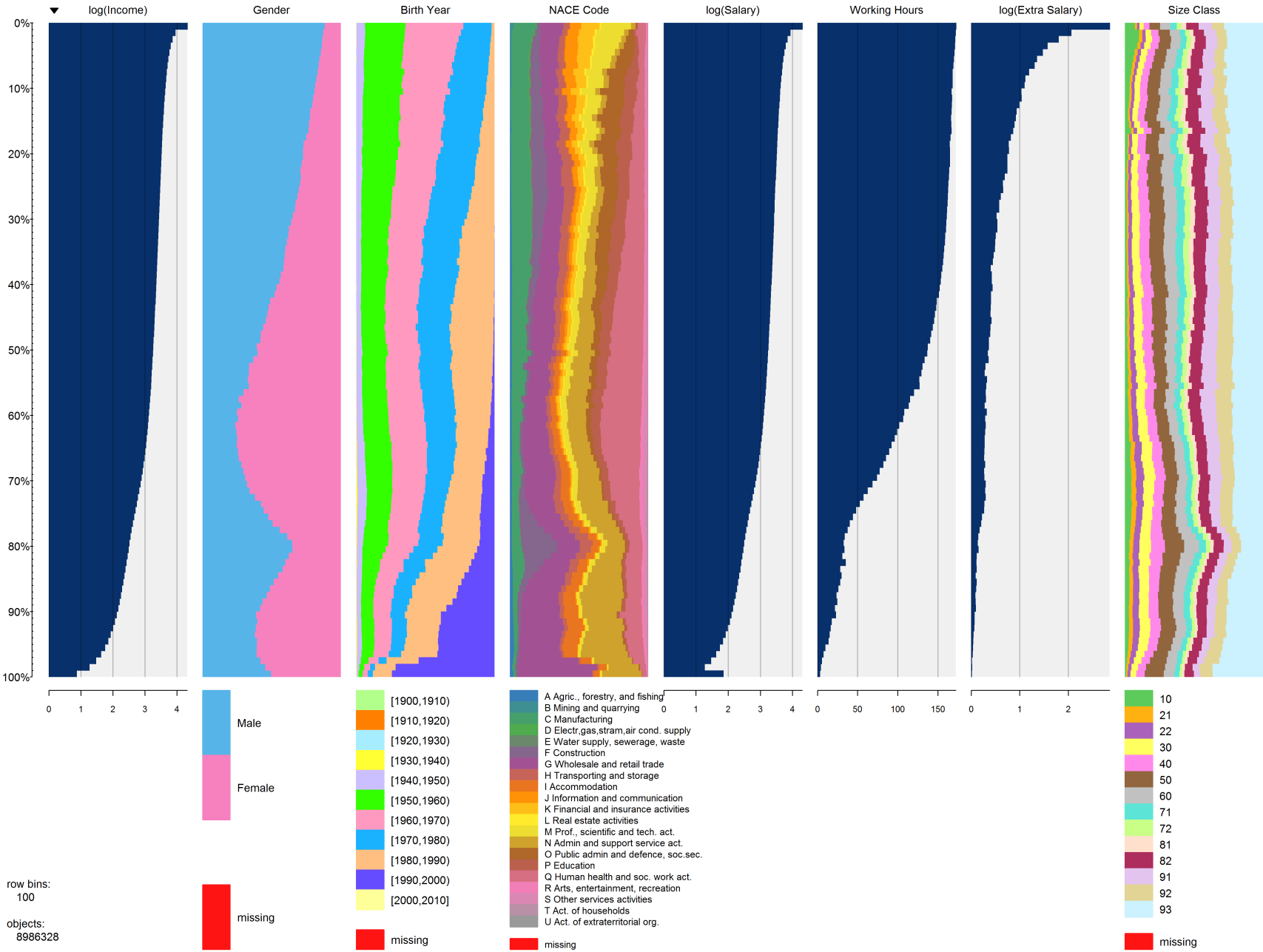
Dutch Labour and Benefits Database



row bins:
100

objects:
19456980

Dutch Labour and Benefits Database



row bins:
100

objects:
8986328

Future plans

- Highly interactice plot in itableplot (with javascript d3)
- Speed up performance even more
- Numeric variable alternatives (e.g. boxplots)
- Hierarchical color palettes (already implemented in treemap)

Future plans

- Highly interactice plot in itableplot (with javascript d3)
- Speed up performance even more
- Numeric variable alternatives (e.g. boxplots)
- Hierarchical color palettes (already implemented in treemap)

Future plans

- Highly interactice plot in itableplot (with javascript d3)
- Speed up performance even more
- Numeric variable alternatives (e.g. boxplots)
- Hierarchical color palettes (already implemented in treemap)

Future plans

- Highly interactice plot in itableplot (with javascript d3)
- Speed up performance even more
- Numeric variable alternatives (e.g. boxplots)
- Hierarchical color palettes (already implemented in treemap)

Future plans

- Highly interactice plot in itableplot (with javascript d3)
- Speed up performance even more
- Numeric variable alternatives (e.g. boxplots)
- Hierarchical color palettes (already implemented in treemap)

References

- Tennekes, M., Jonge, E. de, Daas, P.J.H. (2011) Visual profiling of large statistical datasets. Paper presented at the NTTS 2011
- 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.
- Tennekes, M., Jonge, E. de (2013) On the exploration of high cardinality categorical data. Paper presented at the NTTS 2013
- R package *tabplot* 1.0 is available on CRAN. Development site: https://github.com/mtennekes/tabplot. Version 1.1 (unstable) can be installed from here.