On the exploration of high cardinality categorical data

Martijn Tennekes, Edwin de Jonge

March 6, 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*

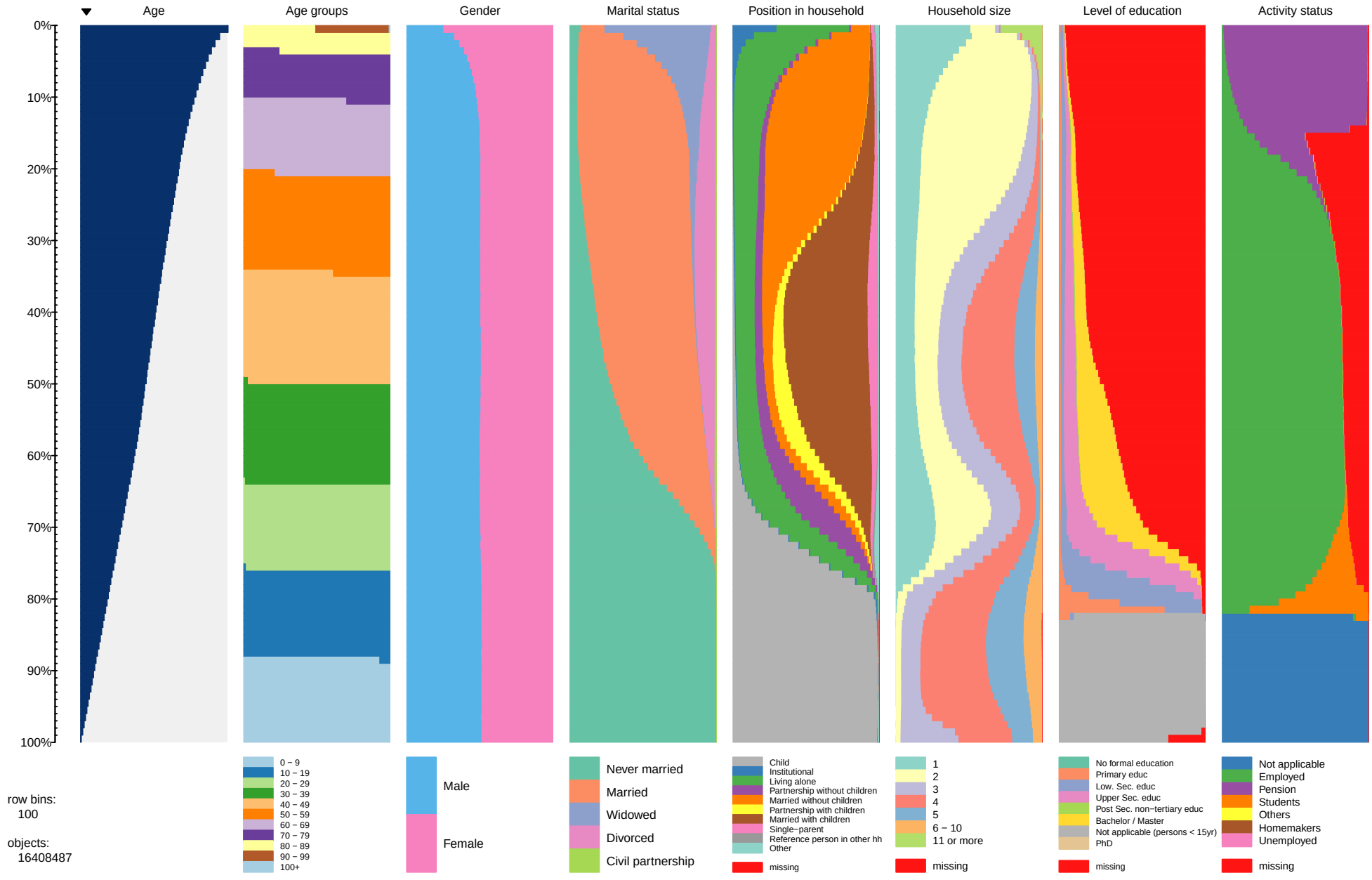
Tableplot

- 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*

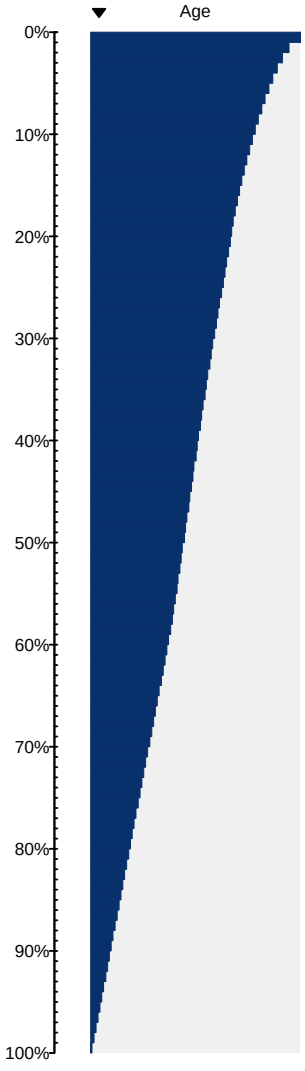
Tableplot: Virtual Census



row bins:
100

objects:
16408487

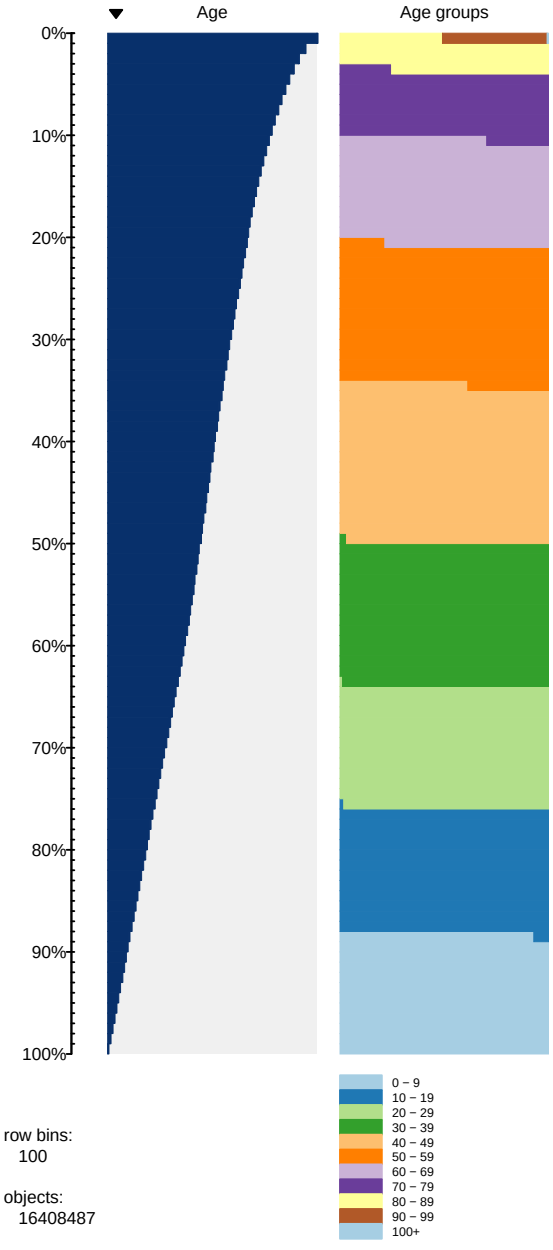
Tableplot: Virtual Census



row bins:
100

objects:
16408487

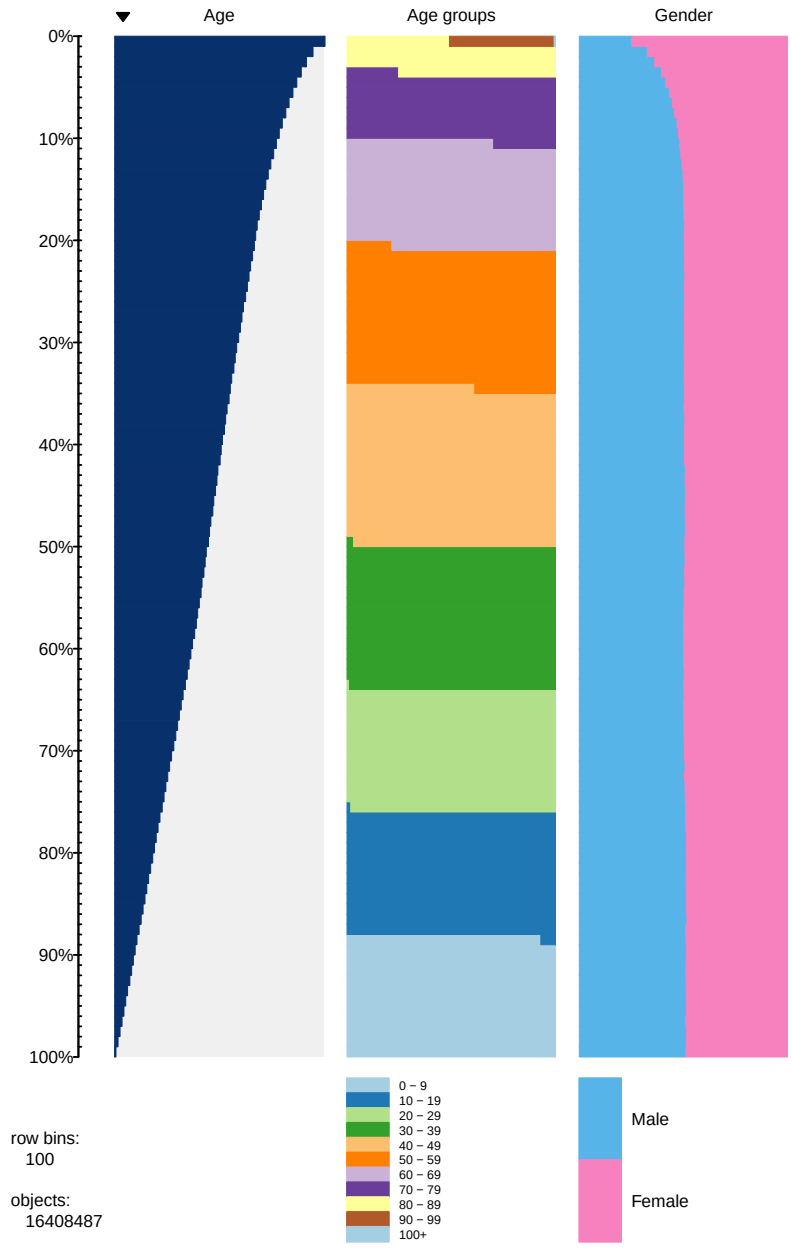
Tableplot: Virtual Census



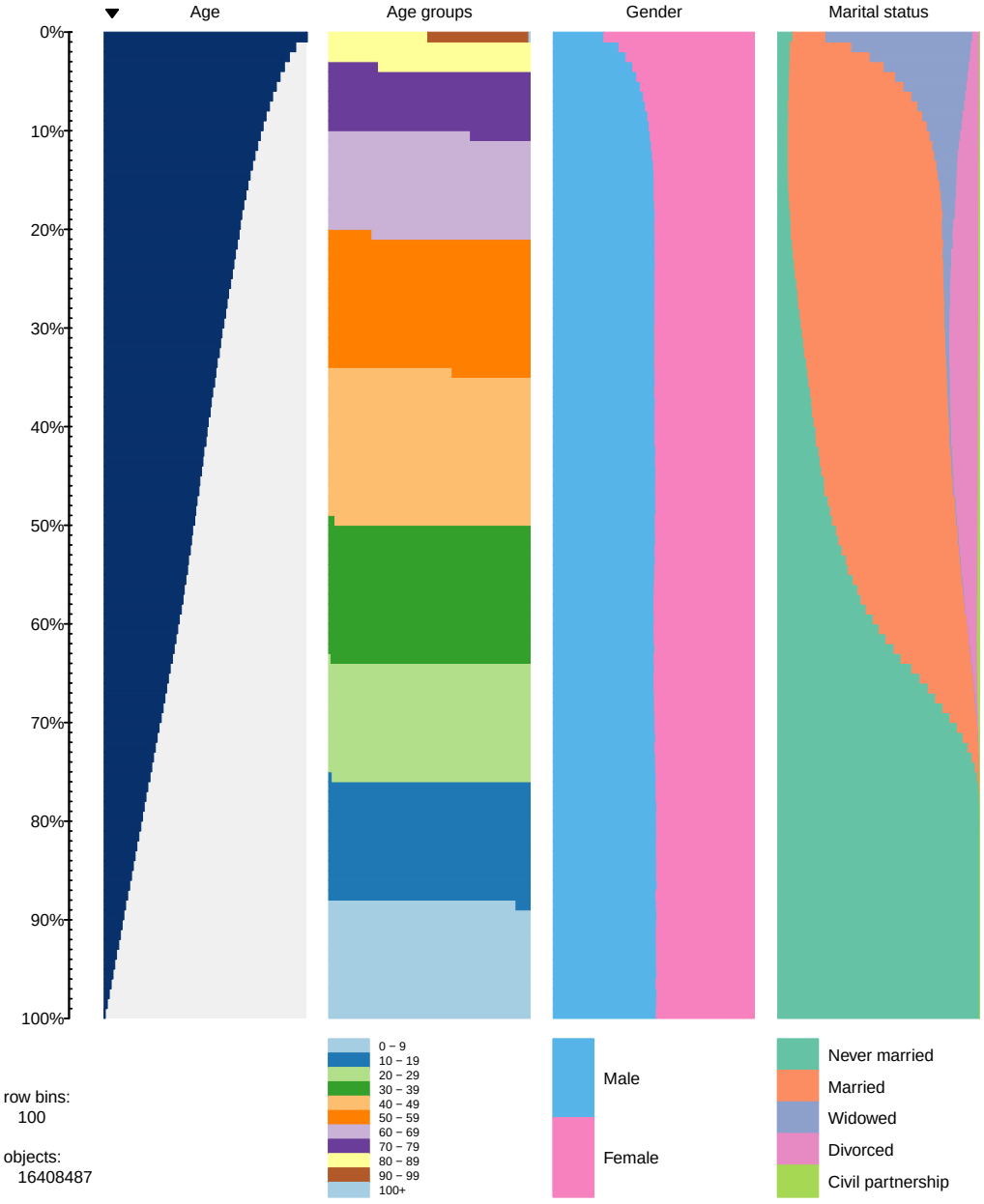
row bins:
100

objects:
16408487

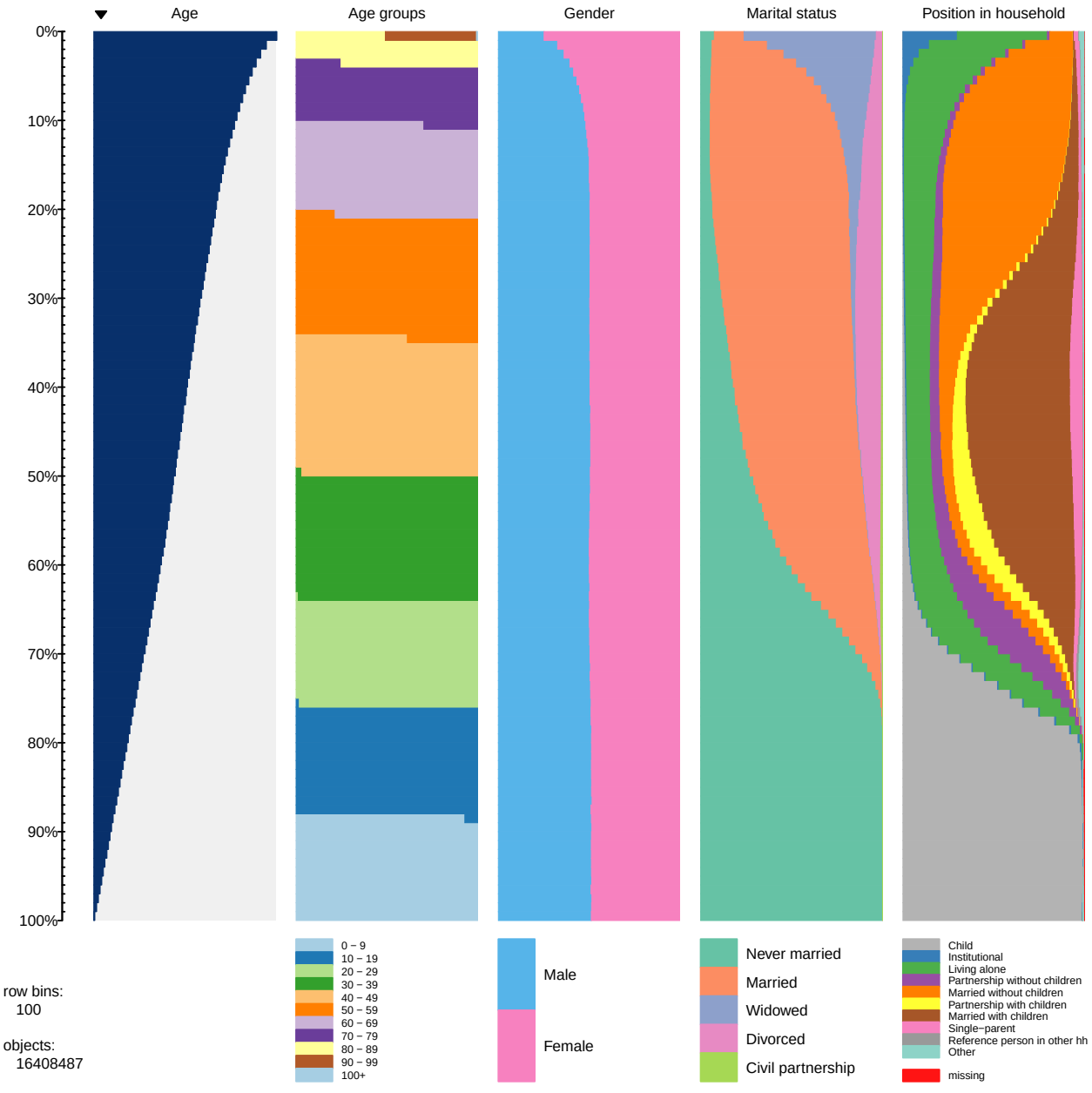
Tableplot: Virtual Census



Tableplot: Virtual Census



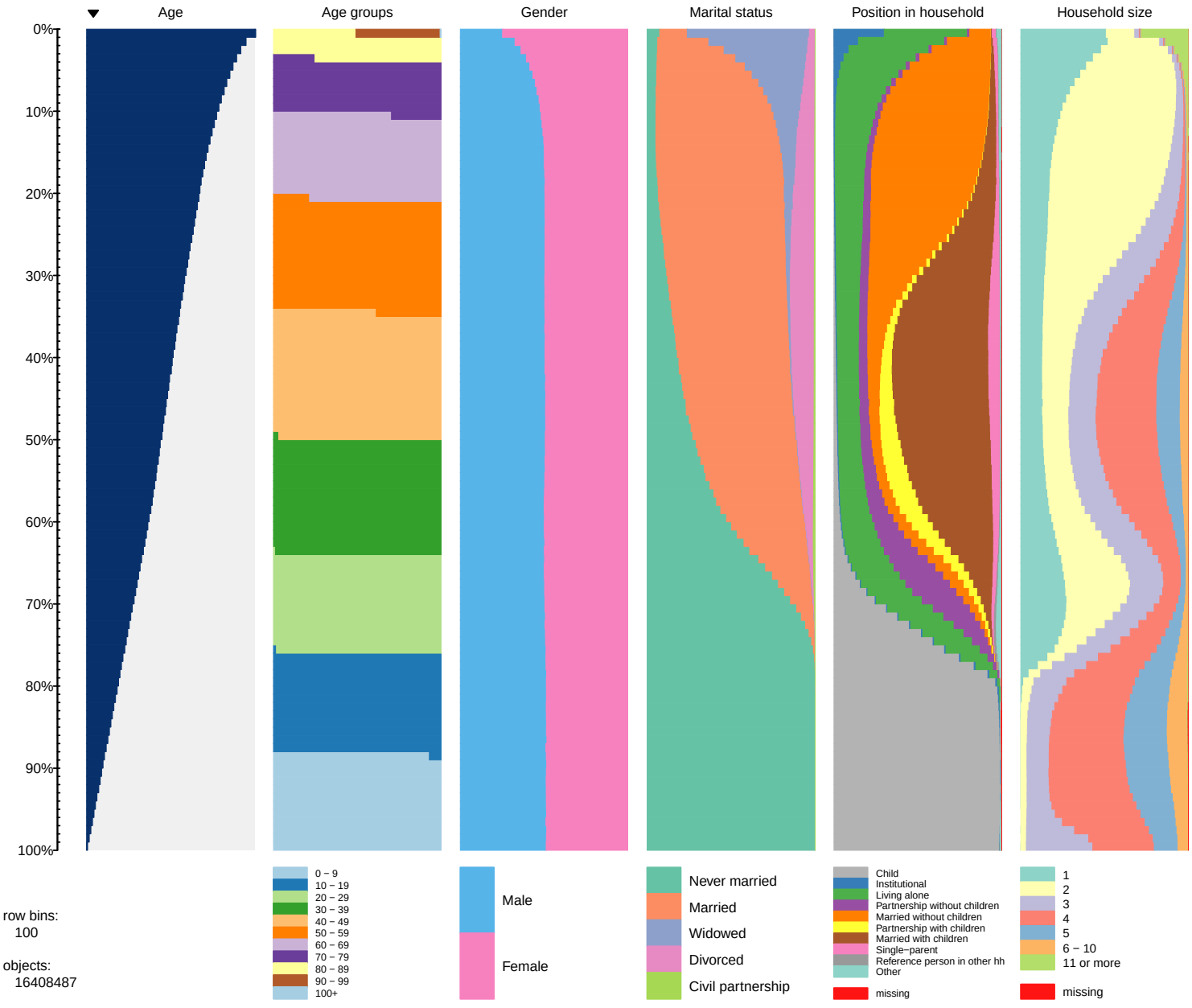
Tableplot: Virtual Census



row bins:
100

objects:
16408487

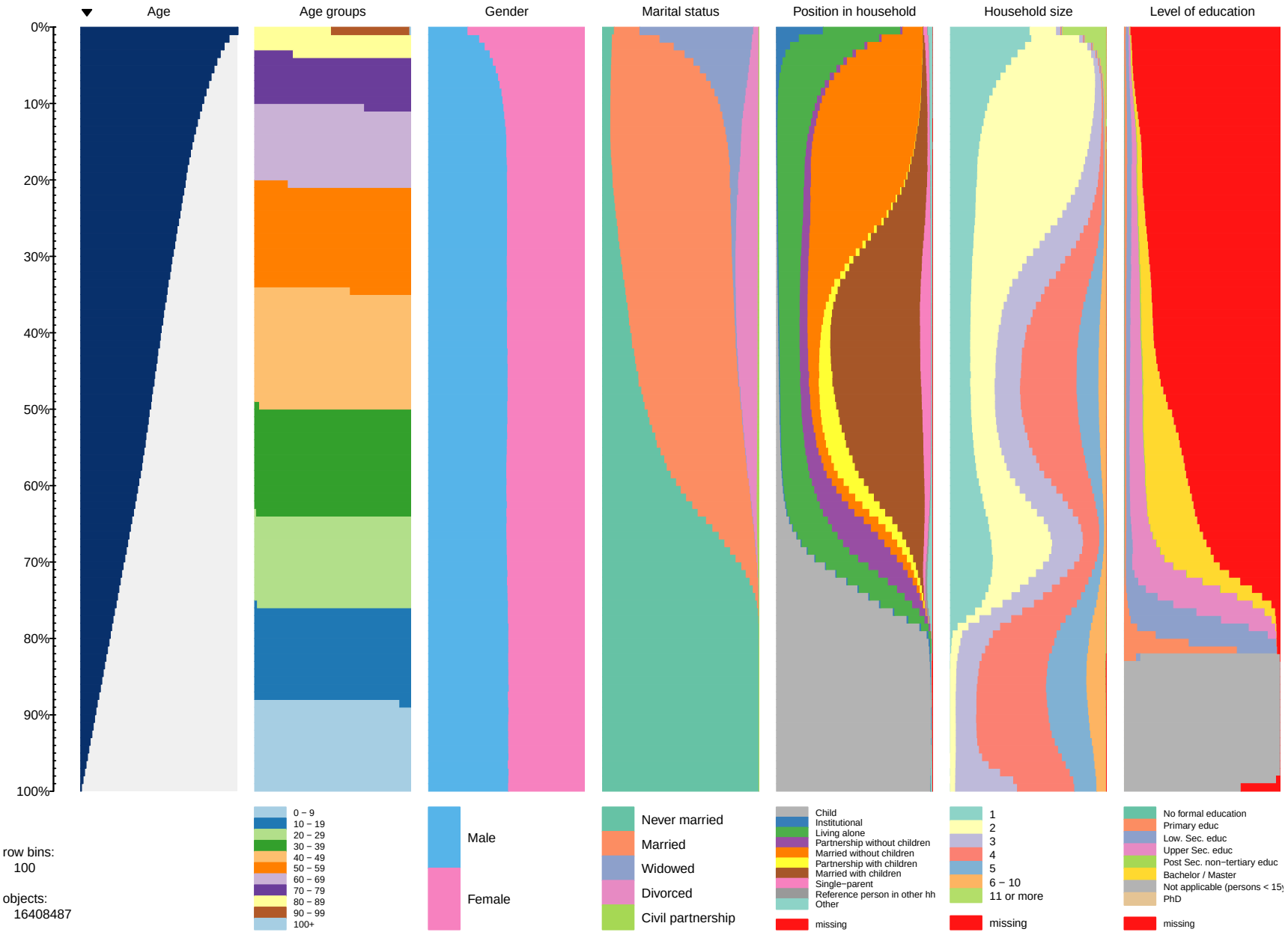
Tableplot: Virtual Census



row bins:
100

objects:
16408487

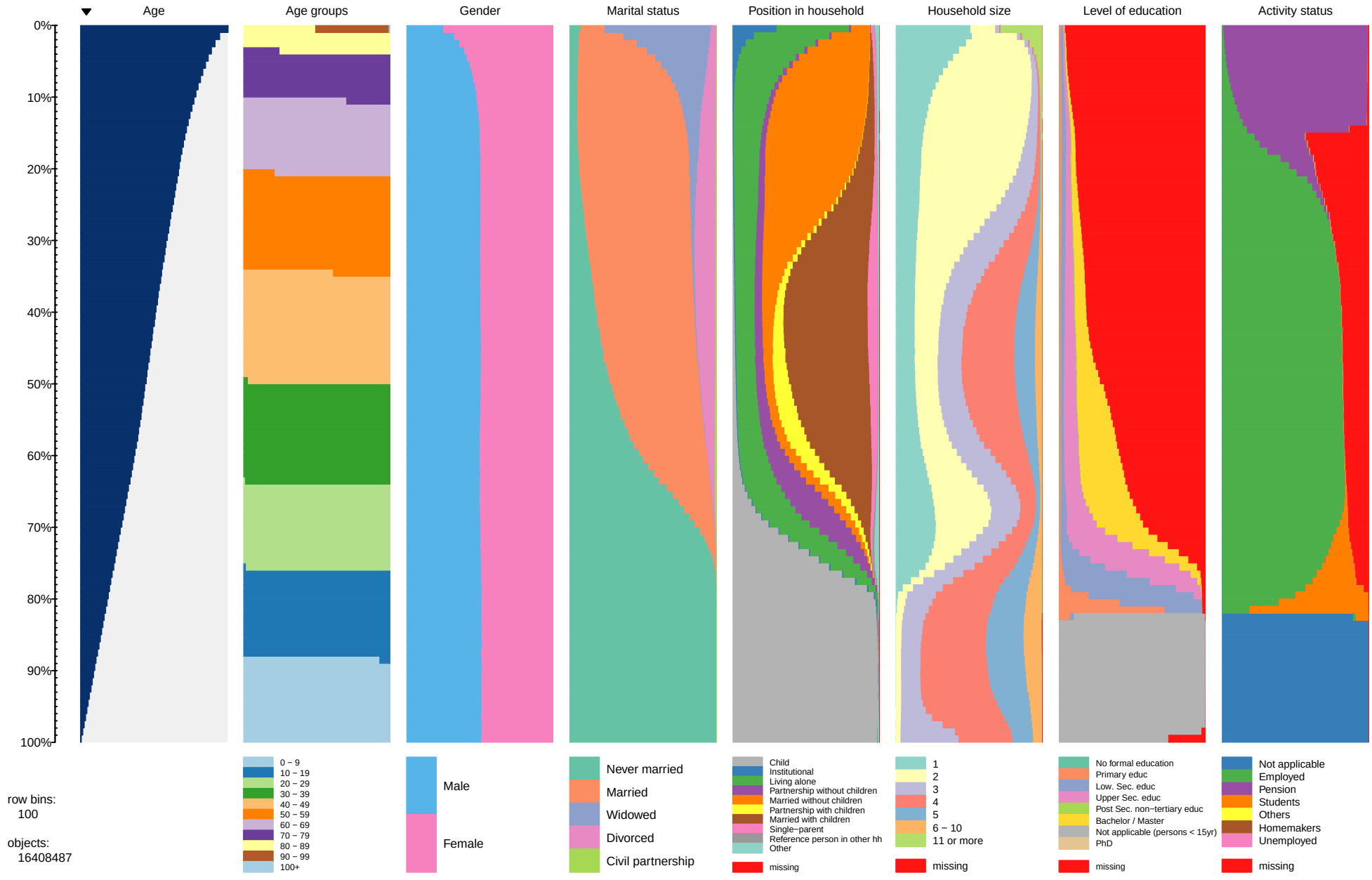
Tableplot: Virtual Census



row bins:
100

objects:
16408487

Tableplot: Virtual Census



row bins:
100

objects:
16408487

Categorical data

Case 1: a couple of categories

- Determine frequencies of each category
- Assign the categories to a qualitative colour palette
- Plot a stacked bar chart column

Categorical data

Case 1: a couple of categories

- Determine frequencies of each category
- Assign the categories to a qualitative colour palette
- Plot a stacked bar chart column

Categorical data

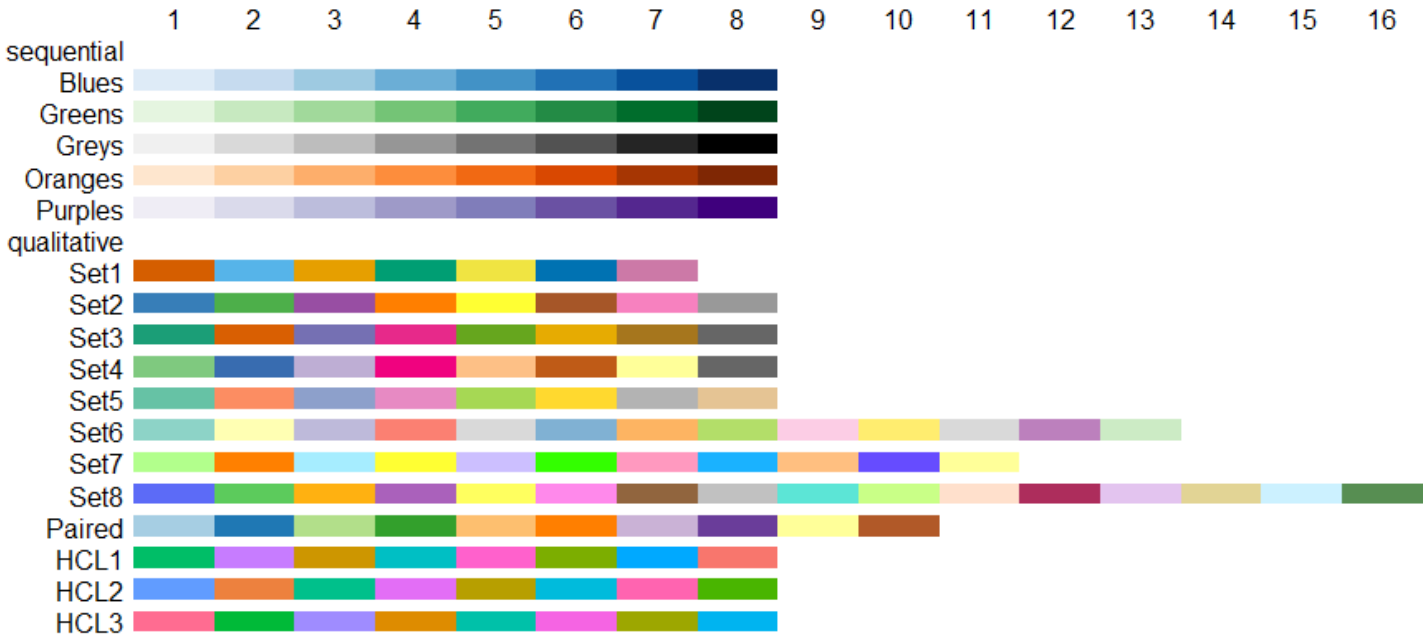
Case 1: a couple of categories

- Determine frequencies of each category
- Assign the categories to a qualitative colour palette
- Plot a stacked bar chart column

Categorical data

Case 1: a couple of categories

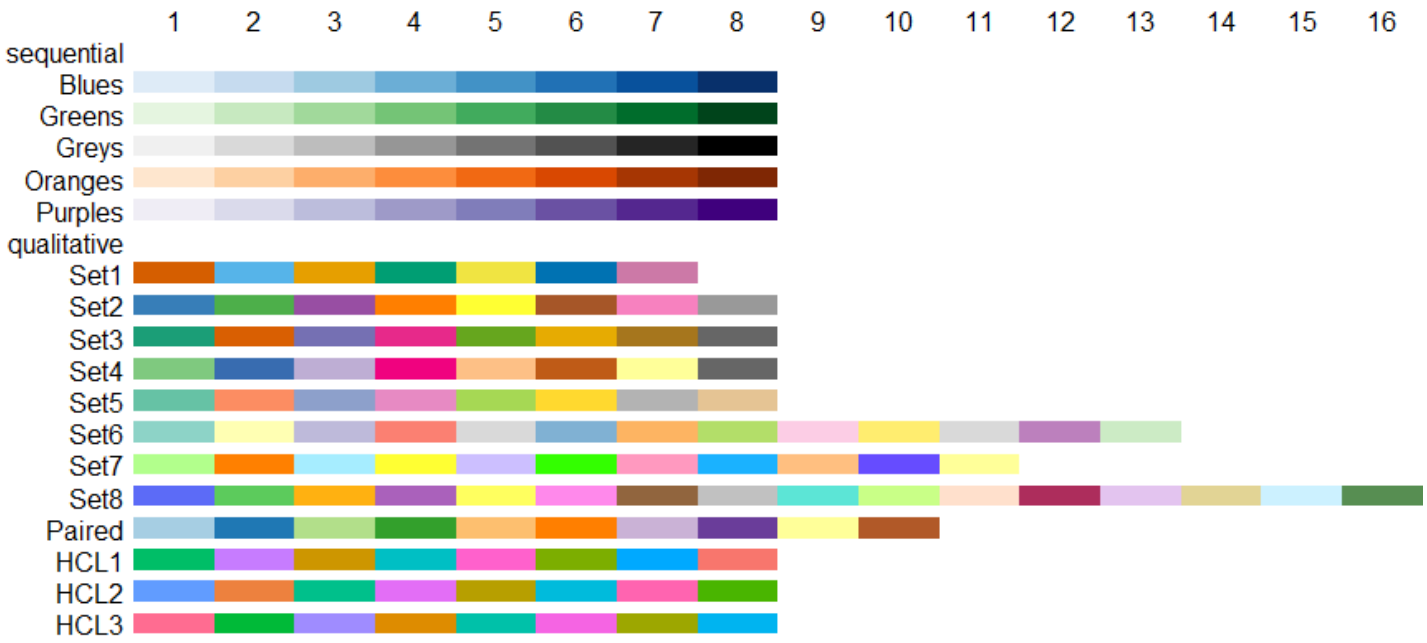
- Determine frequencies of each category
- Assign the categories to a qualitative colour palette
- Plot a stacked bar chart column



Categorical data

Case 1: a couple of categories

- Determine frequencies of each category
- Assign the categories to a qualitative colour palette
- Plot a stacked bar chart column



Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Categorical data

Case 2: many categories (high cardinality)

- Examples: NACE code, education classification
- Not straightforward to plot:
  - Plot all categories? Or cluster them?
  - How to assign them to a colour palette?
  - Legend space is limited

Clustering categories

- Needed: a proper aggregation scheme
  - Ordered variable: bin the categories
  - Hierarchical variable: highest level
- Be careful! Important details may not be noticed

Clustering categories

- Needed: a proper aggregation scheme
  - Ordered variable: bin the categories
  - Hierarchical variable: highest level
- Be careful! Important details may not be noticed

Clustering categories

- Needed: a proper aggregation scheme
  - Ordered variable: bin the categories
  - Hierarchical variable: highest level
- Be careful! Important details may not be noticed

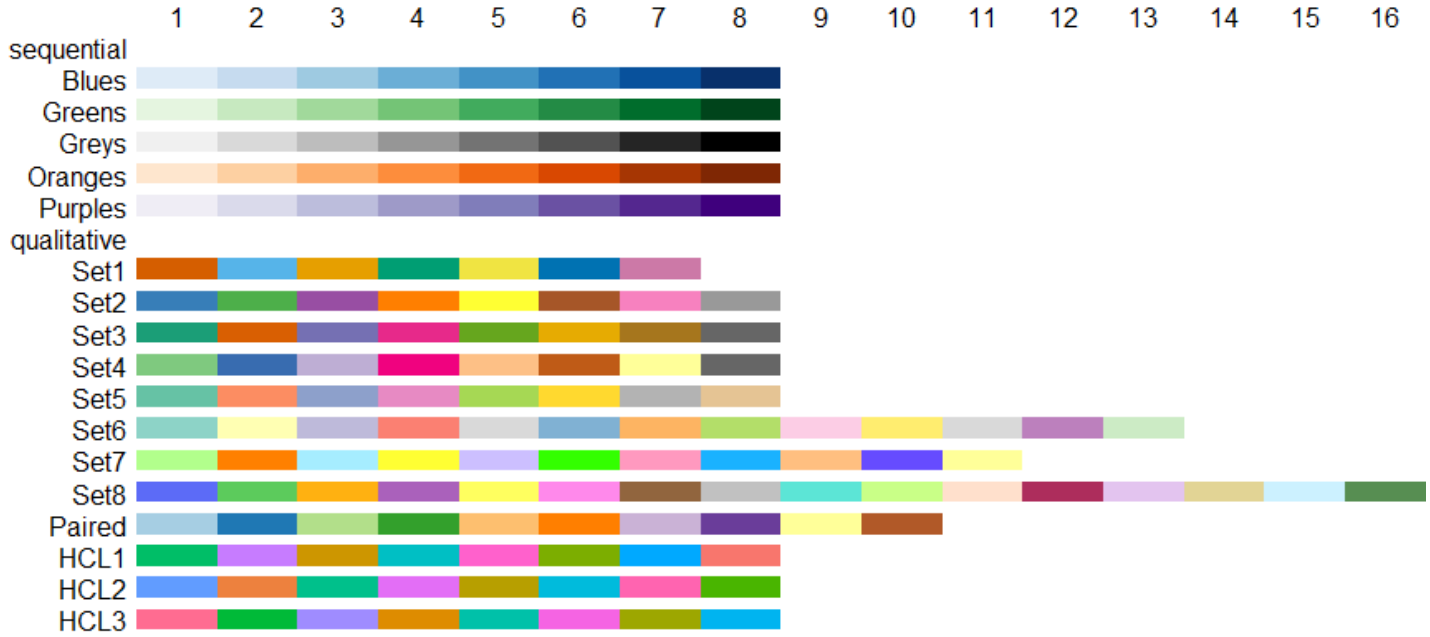
Clustering categories

- Needed: a proper aggregation scheme
  - Ordered variable: bin the categories
  - Hierarchical variable: highest level
- Be careful! Important details may not be noticed

Clustering categories

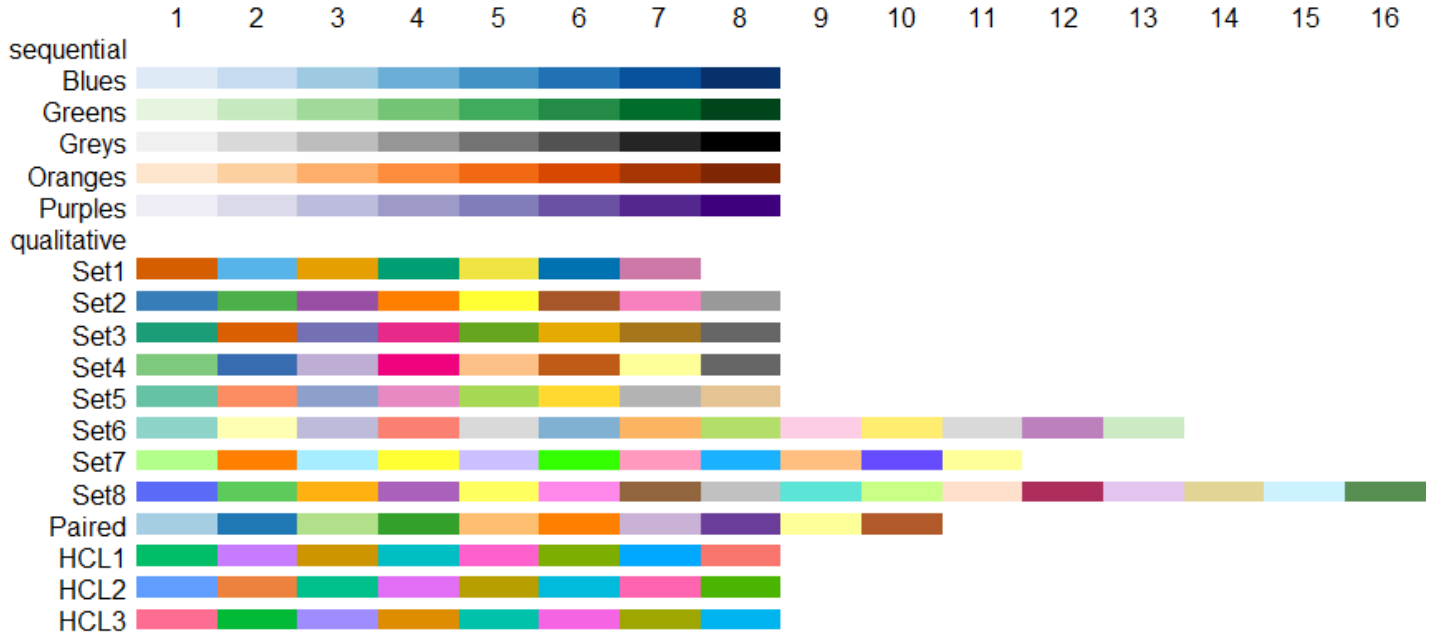
- Needed: a proper aggregation scheme
  - Ordered variable: bin the categories
  - Hierarchical variable: highest level
- Be careful! Important details may not be noticed

Category colours



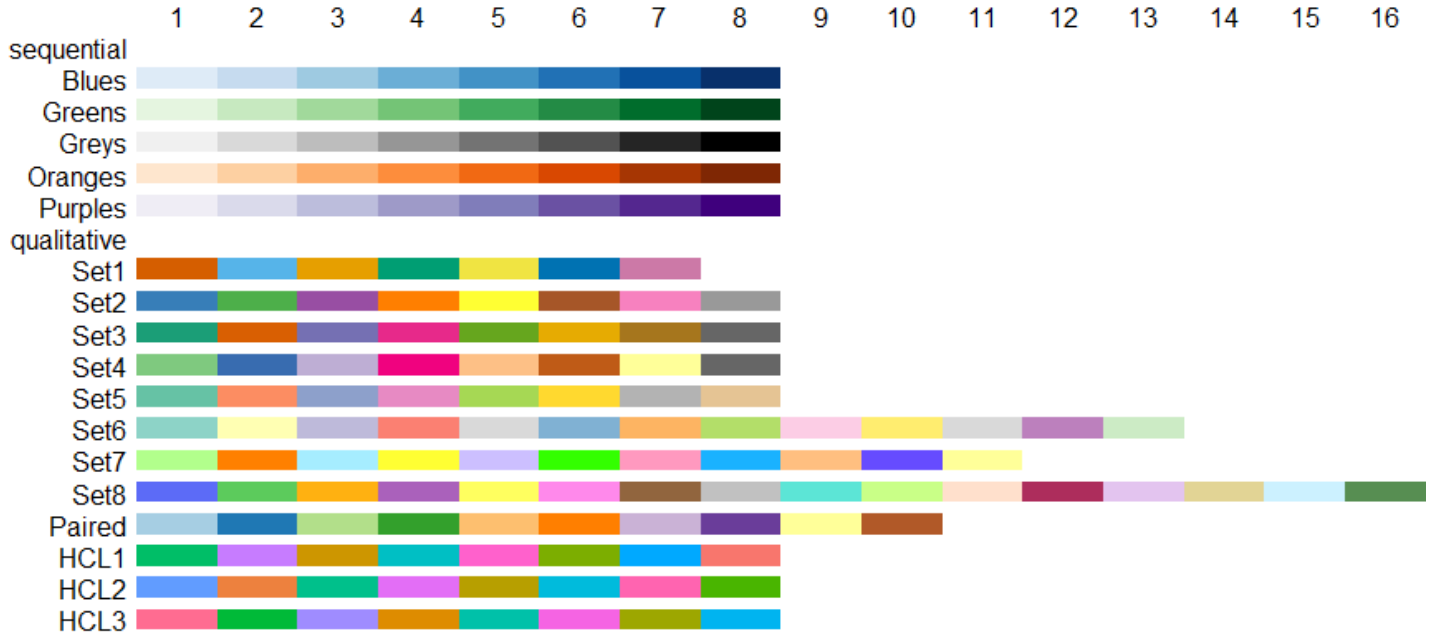
- Not enough colours for > 16 categories
- Create rainbow colour palette:

Category colours



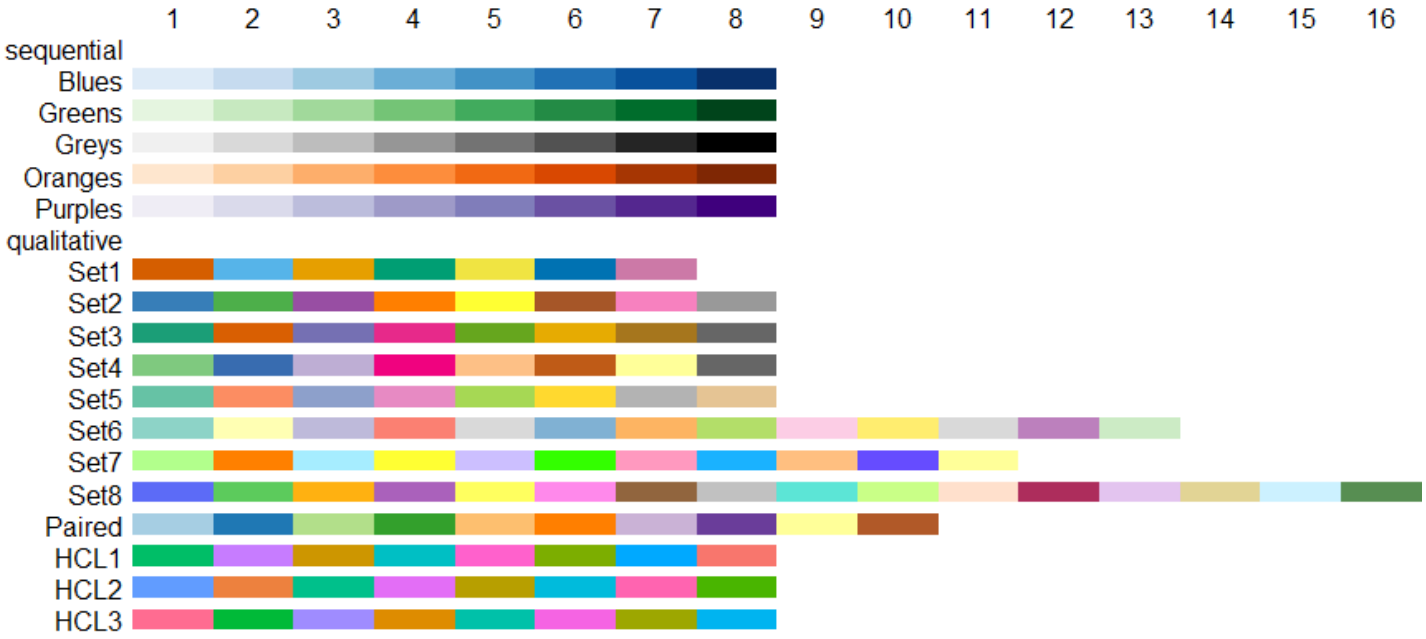
- Not enough colours for > 16 categories
- Create rainbow colour palette:

Category colours

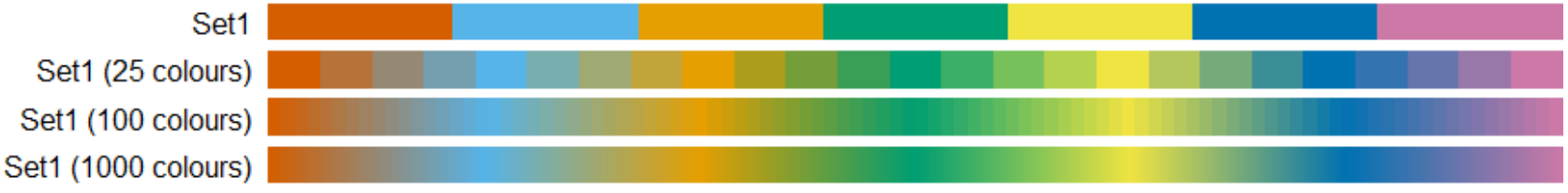


- Not enough colours for > 16 categories
- Create rainbow colour palette:

Category colours



- Not enough colours for $>$ 16 categories
- Create rainbow colour palette:



Method

Two techniques:

i. Clustering categories

ii. Rainbow palette

Can be used together

Method

Two techniques:

i Clustering categories

ii Rainbow palette

Can be used together

Method

Two techniques:

i Clustering categories

ii Rainbow palette

Can be used together

Method

Two techniques:

i Clustering categories

ii Rainbow palette

Can be used together

R package tabplot

Arguments regading categorical data:

- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

R package tabplot

Arguments regading categorical data:

- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

R package tabplot

Arguments regading categorical data:

- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

R package tabplot

Arguments regading categorical data:

- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

R package tabplot

Arguments regading categorical data:

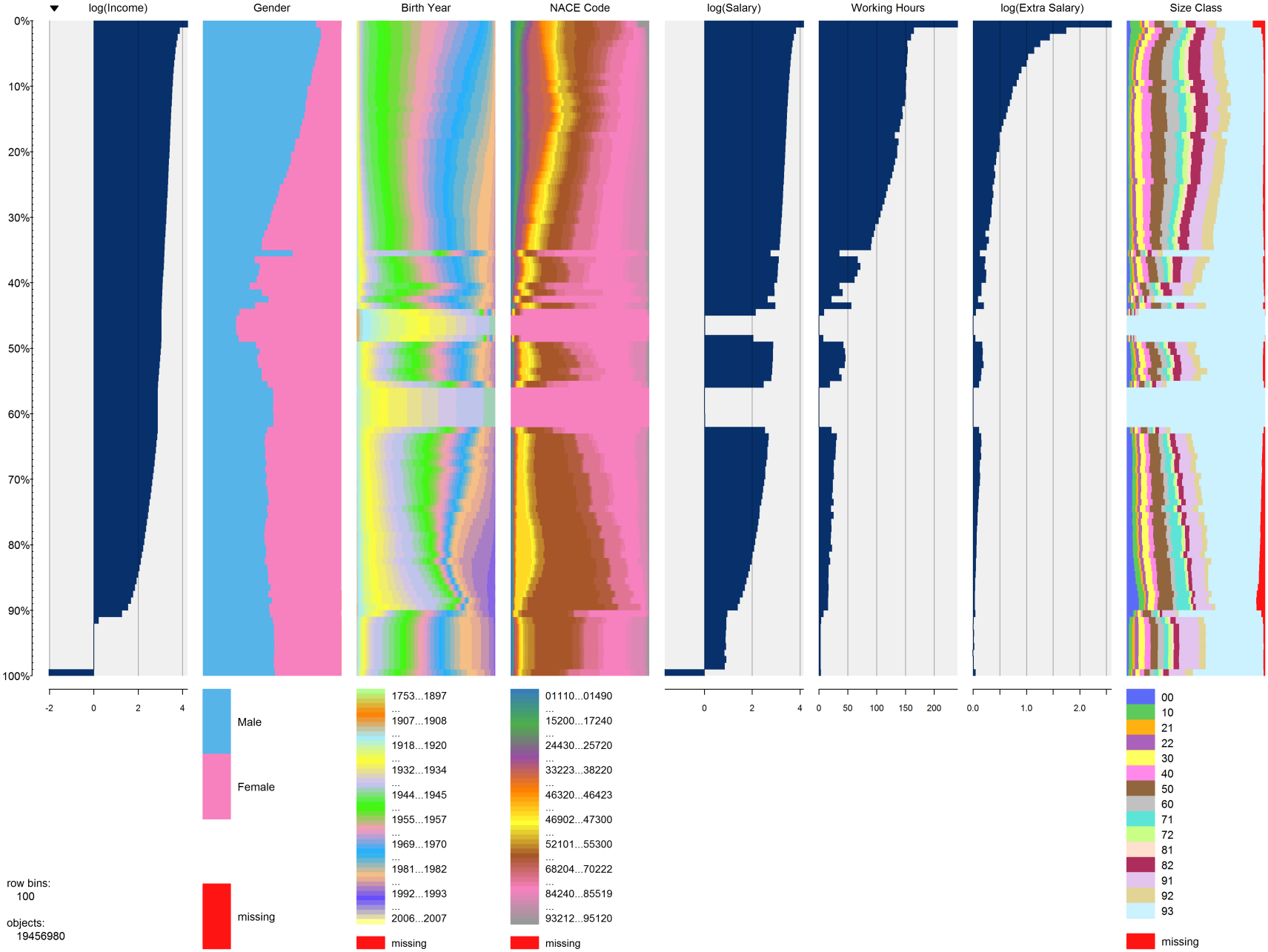
- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

R package tabplot

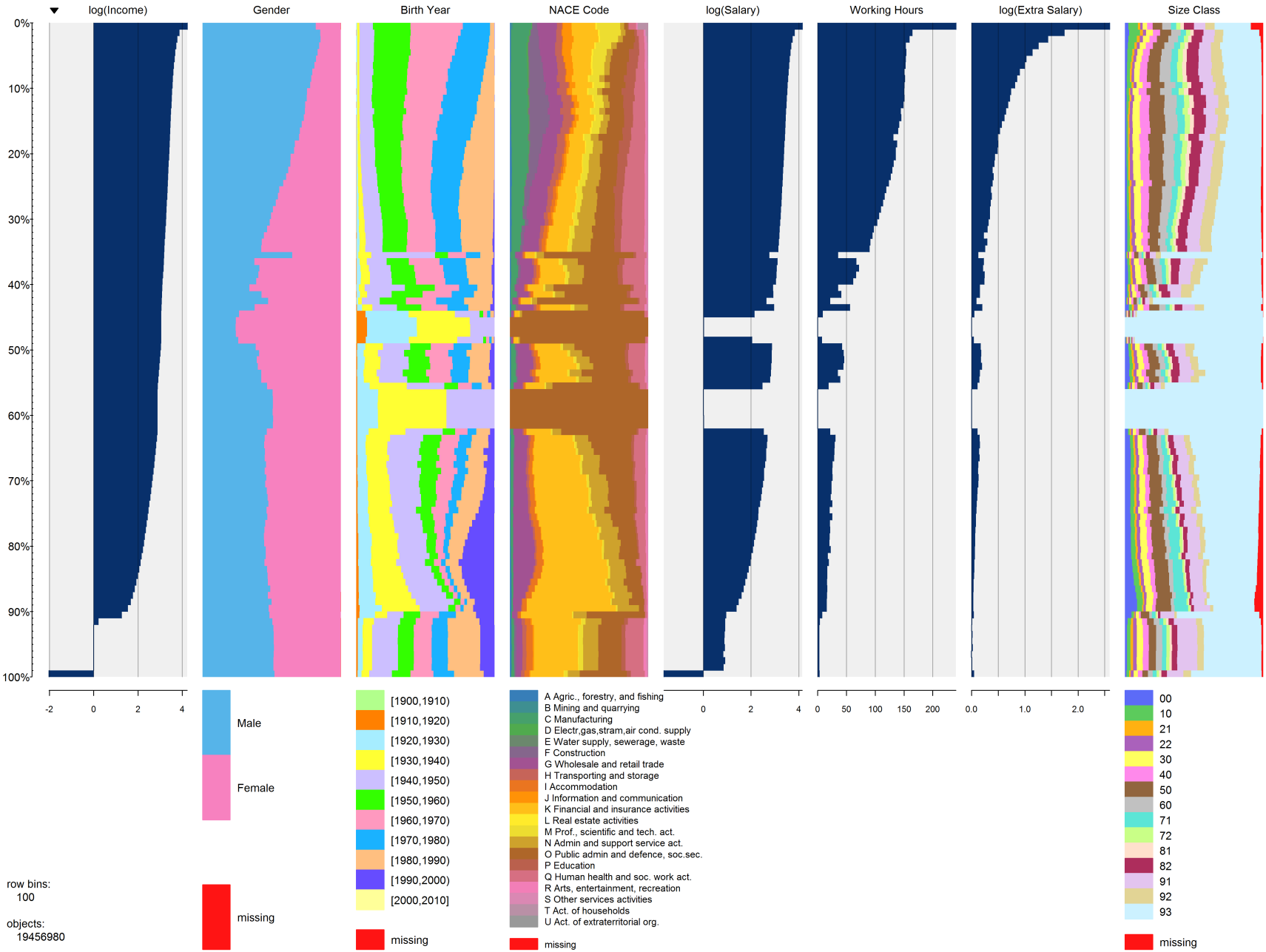
Arguments regading categorical data:

- max_levels (default=50): maximum number of categories. If number of categories $>$ max_levels, then they are automatically clustered to max_levels groups.
- pals: list of palettes
- change_palette_type_at (default=20): determines when a palette is repeated or when a rainbow palette is made
- legend.lines (default=8): number of lines available in the legend

Dutch Labour and Benefits Database



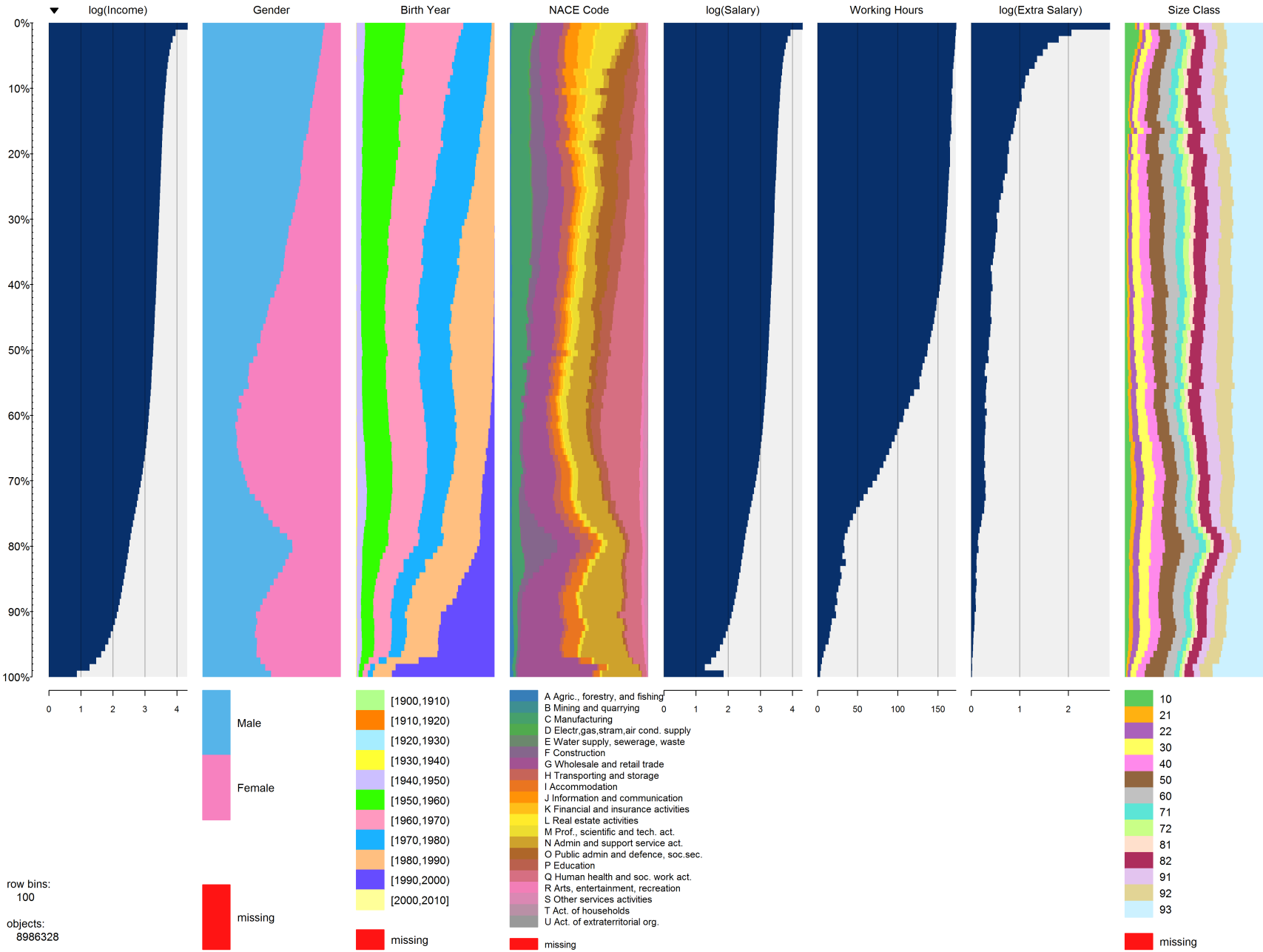
Dutch Labour and Benefits Database



row bins:
100

objects:
19456980

Dutch Labour and Benefits Database



row bins:
100

objects:
8986328

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

Concluding remarks

- Proposed method is useful for the visualisation of high cardinality categorical data
- A proper aggregation scheme is preferred
- Rainbow palette
  - Is applicable to ordinal variables
  - Needs further research for hierarchical variables
- R package *tabplot*: includes a basic interface that will be improved (prototype: package *tabplotd3*)

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.
- R package *tabplot* is available on CRAN. Development site: code.google.com/p/tableplot/