Introduction

Visual tools for data editing
Introduction

Visual tools for data editing

<table>
<thead>
<tr>
<th>Traditional tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheets</td>
</tr>
<tr>
<td>Frequency and cross tables</td>
</tr>
<tr>
<td>Bar charts</td>
</tr>
<tr>
<td>Scatter plots</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
## Introduction

Visual tools for data editing

<table>
<thead>
<tr>
<th>Traditional tools</th>
<th>Not always suitable for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheets</td>
<td>Hierarchical data</td>
</tr>
<tr>
<td>Frequency and cross tables</td>
<td>Multivariate data</td>
</tr>
<tr>
<td>Bar charts</td>
<td>BIG data</td>
</tr>
<tr>
<td>Scatter plots</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
# Introduction

Visual tools for data editing

<table>
<thead>
<tr>
<th>Traditional tools</th>
<th>Not always suitable for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheets</td>
<td>Hierarchical data</td>
</tr>
<tr>
<td>Frequency and cross tables</td>
<td>Multivariate data</td>
</tr>
<tr>
<td>Bar charts</td>
<td>BIG data</td>
</tr>
<tr>
<td>Scatter plots</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
**Introduction**

Visual tools for data editing

<table>
<thead>
<tr>
<th>Traditional tools</th>
<th>Not always suitable for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheets</td>
<td>Hierarchical data</td>
</tr>
<tr>
<td>Frequency and cross tables</td>
<td>Multivariate data</td>
</tr>
<tr>
<td>Bar charts</td>
<td></td>
</tr>
<tr>
<td>Scatter plots</td>
<td>BIG data</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

Visual tools for data editing

<table>
<thead>
<tr>
<th>Traditional tools</th>
<th>Not always suitable for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheets</td>
<td>Hierarchical data</td>
</tr>
<tr>
<td>Frequency and cross tables</td>
<td>Multivariate data</td>
</tr>
<tr>
<td>Bar charts</td>
<td>BIG data</td>
</tr>
<tr>
<td>Scatter plots</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

Visual tools for data editing

**Traditional tools**
- Spreadsheets
- Frequency and cross tables
- Bar charts
- Scatter plots
- ...

**Not always suitable for**
- Hierarchical data
- Multivariate data
- BIG data

**Novel tools**
- Treemap
- Tableplot
# Introduction

Visual tools for data editing

## Traditional tools

- Spreadsheets
- Frequency and cross tables
- Bar charts
- Scatter plots
- ...

## Novel tools

- Treemap
- Tableplot

## Not always suitable for

- Hierarchical data
- Multivariate data
- BIG data
Introduction

Visual tools for data editing

Traditional tools
- Spreadsheets
- Frequency and cross tables
- Bar charts
- Scatter plots
- ...

Not always suitable for
- Hierarchical data
- Multivariate data
- BIG data

Novel tools
- Treemap
- Tableplot

![Graph showing growth w.r.t. last year and total value added in the sector Manufacturing](image)

- Food, beverages, and tobacco
- Textiles and textile products
- Wood and wood products
- Pulp, paper, publishing, and printing
- Coke, petroleum products, and nuclear fuel
- Chemicals, chemical products and rubber and plastic products
- Other non-metallic mineral products
- Metals and metal products
- Machinery and equipment n.e.c.
- Electrical and optical equipment
- Transport equipment N.e.c.

![Graph showing log of employees and legal form](image)
Treemap (SBS data)

Total value added
Treemap (SBS data)

Total value added

- Manufacturing
- Wholesale and retail trade
- Construction
- Real estate, renting, business activities
- Transportation, storage, communication
- Agriculture
- Mining and quarrying
- Electricity, gas, water supply
- Hotels and restaurants
- Other community, social and personal services
- Health and social work
- Total value added
Treemap: trade statistics

Relative difference in % with import XX from NL

Export from NL to EU countries
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package `treemap`
  - Interactive use is very limited
  - Can be used with other tools
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package `treemap`
  - Interactive use is very limited
  - Can be used with other tools
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package `treemap`
  - Interactive use is very limited
  - Can be used with other tools
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package *treemap*
    - Interactive use is very limited
    - Can be used with other tools
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package *treemap*
  - Interactive use is very limited
  - Can be used with other tools
Treemap

- Useful to explore hierarchical data
- Follows the top-down approach of data editing
- Implementation in R:
  - Package *treemap*
  - Interactive use is very limited
  - Can be used with other tools
<table>
<thead>
<tr>
<th>turnover</th>
<th>turnover_cat</th>
<th>employees</th>
<th>employees_cat</th>
<th>legal_form</th>
<th>sector</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1758</td>
<td>[1,000, 10,000)</td>
<td>17</td>
<td>10-19</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>20-49</td>
</tr>
<tr>
<td>4</td>
<td>[10, 100)</td>
<td>3</td>
<td>2-4</td>
<td>Natural person</td>
<td>Wholesale and retail trade</td>
<td>2-4</td>
</tr>
<tr>
<td>84</td>
<td>[10, 100)</td>
<td>1</td>
<td>1</td>
<td>Natural person</td>
<td>Manufacturing</td>
<td>2-4</td>
</tr>
<tr>
<td>11778</td>
<td>[10,000, 100,000)</td>
<td>69</td>
<td>50-99</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>50-99</td>
</tr>
<tr>
<td>10</td>
<td>[10, 100)</td>
<td>7</td>
<td>5-9</td>
<td>Legal person</td>
<td>Health and social work</td>
<td>5-9</td>
</tr>
<tr>
<td>438</td>
<td>[100, 1,000)</td>
<td>4</td>
<td>5-9</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>2-4</td>
</tr>
<tr>
<td>134</td>
<td>[100, 1,000)</td>
<td>3</td>
<td>1</td>
<td>Natural person</td>
<td>Manufacturing</td>
<td>2-4</td>
</tr>
<tr>
<td>1394</td>
<td>[1,000, 10,000)</td>
<td>20</td>
<td>10-19</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>10-19</td>
</tr>
<tr>
<td>164</td>
<td>[100, 1,000)</td>
<td>3</td>
<td>2-4</td>
<td>Natural person</td>
<td>Real est., renting, bus. act.</td>
<td>2-4</td>
</tr>
<tr>
<td>21700</td>
<td>[10,000, 100,000)</td>
<td>76</td>
<td>50-99</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>10-19</td>
</tr>
<tr>
<td>872</td>
<td>[100, 1,000)</td>
<td>14</td>
<td>10-19</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>10-19</td>
</tr>
<tr>
<td>158</td>
<td>[100, 1,000)</td>
<td>4</td>
<td>2-4</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>2-4</td>
</tr>
<tr>
<td>57120</td>
<td>[10,000, 100,000)</td>
<td>211</td>
<td>200-499</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>100-199</td>
</tr>
<tr>
<td>9</td>
<td>[10, 100)</td>
<td>0</td>
<td>1</td>
<td>Natural person</td>
<td>Wholesale and retail trade</td>
<td>1</td>
</tr>
<tr>
<td>99296</td>
<td>[10,000, 100,000)</td>
<td>127</td>
<td>100-199</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>50-99</td>
</tr>
<tr>
<td>695</td>
<td>[100, 1,000)</td>
<td>5</td>
<td>2-4</td>
<td>Natural person</td>
<td>Real est., renting, bus. act.</td>
<td>2-4</td>
</tr>
<tr>
<td>55618</td>
<td>[10,000, 100,000)</td>
<td>362</td>
<td>200-499</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>200-499</td>
</tr>
<tr>
<td>1113</td>
<td>[1,000, 10,000)</td>
<td>17</td>
<td>10-19</td>
<td>Legal person</td>
<td>Hotels and restaurants</td>
<td>10-19</td>
</tr>
<tr>
<td>7478</td>
<td>[1,000, 10,000)</td>
<td>3</td>
<td>5-9</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>10-19</td>
</tr>
<tr>
<td>8662</td>
<td>[1,000, 10,000)</td>
<td>75</td>
<td>50-99</td>
<td>Legal person</td>
<td>Transp., storage, comm.</td>
<td>100-199</td>
</tr>
<tr>
<td>8633</td>
<td>[1,000, 10,000)</td>
<td>36.08</td>
<td>20-49</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>20-49</td>
</tr>
<tr>
<td>1028</td>
<td>[1,000, 10,000)</td>
<td>1</td>
<td>5-9</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>20-49</td>
</tr>
<tr>
<td>183</td>
<td>[100, 1,000)</td>
<td>0.2</td>
<td>2-4</td>
<td>Natural person</td>
<td>Wholesale and retail trade</td>
<td>2-4</td>
</tr>
<tr>
<td>25</td>
<td>[10, 100)</td>
<td>1.6</td>
<td>1</td>
<td>Natural person</td>
<td>Real est., renting, bus. act.</td>
<td>1</td>
</tr>
<tr>
<td>7182</td>
<td>[1,000, 10,000)</td>
<td>82</td>
<td>50-99</td>
<td>Legal person</td>
<td>Construction</td>
<td>10-19</td>
</tr>
<tr>
<td>2005</td>
<td>[1,000, 10,000)</td>
<td>3</td>
<td>0</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>20-49</td>
</tr>
<tr>
<td>10398</td>
<td>[10,000, 100,000)</td>
<td>8</td>
<td>10-19</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>100-199</td>
</tr>
<tr>
<td>2245</td>
<td>[1,000, 10,000)</td>
<td>34</td>
<td>20-49</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>20-49</td>
</tr>
<tr>
<td>4903</td>
<td>[1,000, 10,000)</td>
<td>18</td>
<td>10-19</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>20-49</td>
</tr>
<tr>
<td>68</td>
<td>[10, 100)</td>
<td>-1</td>
<td>1</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>2-4</td>
</tr>
</tbody>
</table>

objects: 51621
Tableplot: SBS data (unprocessed)

<table>
<thead>
<tr>
<th>turnover</th>
<th>turnover_cat</th>
<th>employees</th>
<th>employees_cat</th>
<th>legal_form</th>
<th>sector</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>8124110</td>
<td>[1 mln, 10 mln)</td>
<td>17,172</td>
<td>500+</td>
<td>Legal person</td>
<td>Transp., storage, comm.</td>
<td>500+</td>
</tr>
<tr>
<td>7375977</td>
<td>[1 mln, 10 mln)</td>
<td>1865</td>
<td>500+</td>
<td>Legal person</td>
<td>Mining and quarrying</td>
<td>500+</td>
</tr>
<tr>
<td>5496033</td>
<td>[1 mln, 10 mln)</td>
<td>1366</td>
<td>500+</td>
<td>Legal person</td>
<td>Electricity, gas, water supply</td>
<td>500+</td>
</tr>
<tr>
<td>4421226</td>
<td>[1 mln, 10 mln)</td>
<td>431</td>
<td>200-499</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>200-499</td>
</tr>
<tr>
<td>4386360</td>
<td>[1 mln, 10 mln)</td>
<td>136</td>
<td>100-199</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>4012386</td>
<td>[1 mln, 10 mln)</td>
<td>431</td>
<td>200-499</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>200-499</td>
</tr>
<tr>
<td>3909439</td>
<td>[1 mln, 10 mln)</td>
<td>9187</td>
<td>500+</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>3133380</td>
<td>[1 mln, 10 mln)</td>
<td>1004</td>
<td>500+</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>500+</td>
</tr>
<tr>
<td>3095844</td>
<td>[1 mln, 10 mln)</td>
<td>1036</td>
<td>500+</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>2931738</td>
<td>[1 mln, 10 mln)</td>
<td>1366</td>
<td>500+</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>200-499</td>
</tr>
<tr>
<td>2872912</td>
<td>[1 mln, 10 mln)</td>
<td>1695</td>
<td>500+</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>2765533</td>
<td>[1 mln, 10 mln)</td>
<td>12</td>
<td>10-19</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>20-49</td>
</tr>
<tr>
<td>2629875</td>
<td>[1 mln, 10 mln)</td>
<td>229</td>
<td>200-499</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>50-99</td>
</tr>
<tr>
<td>2049669</td>
<td>[1 mln, 10 mln)</td>
<td>4124</td>
<td>500+</td>
<td>Legal person</td>
<td>Real est., renting, bus. act.</td>
<td>500+</td>
</tr>
<tr>
<td>1952200</td>
<td>[1 mln, 10 mln)</td>
<td>1309</td>
<td>500+</td>
<td>Legal person</td>
<td>Electricity, gas, water supply</td>
<td>500+</td>
</tr>
<tr>
<td>1945146</td>
<td>[1 mln, 10 mln)</td>
<td>336</td>
<td>200-499</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>1840955</td>
<td>[1 mln, 10 mln)</td>
<td>415</td>
<td>200-499</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>200-499</td>
</tr>
<tr>
<td>1724782</td>
<td>[1 mln, 10 mln)</td>
<td>168</td>
<td>100-199</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>100-199</td>
</tr>
<tr>
<td>1721028</td>
<td>[1 mln, 10 mln)</td>
<td>2787</td>
<td>500+</td>
<td>Legal person</td>
<td>Transp., storage, comm.</td>
<td>500+</td>
</tr>
<tr>
<td>1493974</td>
<td>[1 mln, 10 mln)</td>
<td>105</td>
<td>100-199</td>
<td>Legal person</td>
<td>Electricity, gas, water supply</td>
<td>200-499</td>
</tr>
<tr>
<td>1406432</td>
<td>[1 mln, 10 mln)</td>
<td>943</td>
<td>500+</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>500+</td>
</tr>
<tr>
<td>1361787</td>
<td>[1 mln, 10 mln)</td>
<td>710</td>
<td>500+</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>1329278</td>
<td>[1 mln, 10 mln)</td>
<td>5841</td>
<td>500+</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>500+</td>
</tr>
<tr>
<td>1320454</td>
<td>[1 mln, 10 mln)</td>
<td>5032</td>
<td>500+</td>
<td>Legal person</td>
<td>Manufacturing</td>
<td>500+</td>
</tr>
<tr>
<td>1318814</td>
<td>[1 mln, 10 mln)</td>
<td>30</td>
<td>20-49</td>
<td>Legal person</td>
<td>Wholesale and retail trade</td>
<td>20-49</td>
</tr>
<tr>
<td>1316711</td>
<td>[1 mln, 10 mln)</td>
<td>2055</td>
<td>500+</td>
<td>Legal person</td>
<td>Electricity, gas, water supply</td>
<td>500+</td>
</tr>
</tbody>
</table>

objects: 51621
Tableplot: SBS data (unprocessed)

<table>
<thead>
<tr>
<th>turnover</th>
<th>turnover_cat</th>
<th>employees</th>
<th>employees_cat</th>
<th>legal_form</th>
<th>sector</th>
<th>size</th>
</tr>
</thead>
</table>

objects: 51621
Tableplot: SBS data (unprocessed)

row bins: 100
objects: 51621
Tableplot: SBS data (unprocessed)

<table>
<thead>
<tr>
<th>log(turnover)</th>
<th>turnover_cat</th>
<th>log(employees)</th>
<th>employees_cat</th>
<th>legal_form</th>
<th>sector</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>0</td>
<td>(0, 10)</td>
<td>[10, 100)</td>
<td>(0, 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[10, 1000)</td>
<td>[1000, 10000)</td>
<td>[10000, 100000)</td>
<td>[10000, 100000)</td>
<td>[100000, 1 mln)</td>
<td>&gt;= 10 mln</td>
<td></td>
</tr>
<tr>
<td>&gt;= 10 mln</td>
<td>missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

row bins: 100
objects: 51621
Tableplot: SBS data (edited)

- **log(turnover)**
- **turnover_cat**: < 0, 0, (0, 10), [10, 100), [100, 1,000), [1,000, 10,000), [10,000, 100,000), [100,000, 1 mln), [1 mln, 10 mln), >= 10 mln
- **missing**

- **log(employees)**
- **employees_cat**: 0, 1, 2−4, 5−9, 10−19, 20−49, 50−99, 100−199, 200−499, 500+
- **missing**

- **legal_form**: Natural person, Legal person
- **sector**: Agriculture, Mining and quarrying, Manufacturing, Electricity, gas, water supply, Construction, Wholesale and retail trade, Hotels and restaurants, Transp., storage, comm., Real est., renting, bus. act., Health and social work, Oth. comm., soc., pers.serv.
- **size**: 0, 1, 2−4, 5−9, 10−19, 20−49, 50−99, 100−199, 200−499, 500+
- **missing**

Row bins: 100
Objects: 51621
Tableplot: virtual census

- **Age groups**
  - 0 − 9
  - 10 − 19
  - 20 − 29
  - 30 − 39
  - 40 − 49
  - 50 − 59
  - 60 − 69
  - 70 − 79
  - 80 − 89
  - 90 − 99
  - 100+

- **Gender**
  - Male
  - Female

- **Marital status**
  - Never married
  - Married
  - Widowed
  - Divorced
  - Single-parent
  - Civil partnership

- **Position in household**
  - Child
  - Institutional
  - Living alone
  - Partnership without children
  - Married without children
  - Partnership with children
  - Married with children
  - Reference person in other hh

- **Household size**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6 − 10
  - 11 or more

- **Level of education**
  - No formal education
  - Primary educ
  - Low. Sec. educ
  - Upper Sec. educ
  - Post Sec. non-tertiary educ
  - Bachelor / Master
  - PhD

- **Activity status**
  - Employed
  - Pension
  - Students
  - Others
  - Homemakers
  - Unemployed
  - Not applicable (persons < 15yr)
  - Not applicable

- **Row bins:** 100
- **Objects:** 16408487
Tableplot: policy record administration

- log(Income_tax)
- Birth_year
- Gender
- Sector
- Collective_agreement
- Risk_group
- Wages
- Travel_costs
- Size
- FTE

Row bins: 100
Objects: 20207271

- 1753
- ... 1923
- ... 1952
- ... 1981
- ... 2010

- 01110
- ... 27900
- ... 46732
- ... 69105
- ... 99000

- 0
- ... 552
- ... 1283
- ... 2066
- ... 9999

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- ... 93

- 00
- 10
- 21
- 32
- 43
- 54
- 65
- 76
- 87
- 98
- ... 99
Tableplot

- Useful to explore BIG multivariate data
  - Strange data patterns
  - Selectivity of missing values
  - Correlation between variables

- Implementation in R:
  - Package `tabplot`
  - Old graphical user interface: package `tabplotGTK`
  - Interactive web-based interface: package `tabplotd3`
Tableplot

- Useful to explore BIG multivariate data
  - Strange data patterns
  - Selectivity of missing values
  - Correlation between variables

- Implementation in R:
  - Package `tabplot`
  - Old graphical user interface: package `tabplotGTK`
  - Interactive web-based interface: package `tabplotd3`
Tableplot

- Useful to explore BIG multivariate data
  - Strange data patterns
  - Selectivity of missing values
  - Correlation between variables
- Implementation in R:
  - Package `tabplot`
  - Old graphical user interface: package `tabplotGTK`
  - Interactive web-based interface: package `tabplotd3`
Tableplot

- Useful to explore BIG multivariate data
  - Strange data patterns
  - Selectivity of missing values
  - Correlation between variables

- Implementation in R:
  - Package `tabplot`
    - Old graphical user interface: package `tabplotGTK`
    - Interactive web-based interface: package `tabplotd3`
Tableplot

Useful to explore BIG multivariate data
- Strange data patterns
- Selectivity of missing values
- Correlation between variables

Implementation in R:
- Package `tabplot`
- Old graphical user interface: package `tabplotGTK`
- Interactive web-based interface: package `tabplotd3`
Tableplot

- Useful to explore BIG multivariate data
  - Strange data patterns
  - Selectivity of missing values
  - Correlation between variables

- Implementation in R:
  - Package *tabplot*
  - Old graphical user interface: package *tabplotGTK*
  - Interactive web-based interface: package *tabplotd3*
References


- All R-packages are available on CRAN. Development sites:
  - treemap: code.google.com/p/treemap-package/
  - tabplot: code.google.com/p/tableplot/
  - tabplotd3: code.google.com/p/tabplotd3/