Data preparation

With LMDatasource module
Table of content

- Introduction to LMDataSource module
- Definition of the primary key
- Definition of attributes
- Frequency analysis
LMDDataSource and 4ft-Miner

Data Source

Metabase
- Definition of the data matrix
- Definition of the relevant rules set
- The set of simple rules

Analyzed data (read only)

4ft-Miner
- 4ftTask
- 4ftResult
Database table and data matrix

- An input for the 4ft-Miner procedure is a data matrix created by LMDataSource module.
- Data matrix is created from one database table.
- Each column of the data matrix is created by one column of the database table.
- The column of the data matrix is called attribute.
- Values of the attribute are called categories.
### Database table Loan

<table>
<thead>
<tr>
<th>Nr.</th>
<th>amount</th>
<th>birth_number</th>
<th>District</th>
<th>duration</th>
<th>loan_id</th>
<th>payments</th>
<th>Salary</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80962</td>
<td>450204</td>
<td>Praha</td>
<td>24</td>
<td>4959</td>
<td>3373</td>
<td>12541</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>30276</td>
<td>395423</td>
<td>Tabor</td>
<td>12</td>
<td>4961</td>
<td>2523</td>
<td>9104</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>30276</td>
<td>620209</td>
<td>Frydek - Mistek</td>
<td>12</td>
<td>4962</td>
<td>2523</td>
<td>9893</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>318480</td>
<td>520826</td>
<td>Strakonice</td>
<td>60</td>
<td>4967</td>
<td>5308</td>
<td>8547</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>110736</td>
<td>405130</td>
<td>Prachatice</td>
<td>48</td>
<td>4968</td>
<td>2307</td>
<td>8402</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td>165960</td>
<td>445613</td>
<td>Jindrichuv Hradec</td>
<td>24</td>
<td>4973</td>
<td>6915</td>
<td>8427</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>102876</td>
<td>420128</td>
<td>Ostrava - mesto</td>
<td>12</td>
<td>4986</td>
<td>8573</td>
<td>10673</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>265320</td>
<td>670921</td>
<td>Chrudim</td>
<td>36</td>
<td>4988</td>
<td>7370</td>
<td>8254</td>
<td>D</td>
</tr>
<tr>
<td>9</td>
<td>352704</td>
<td>506227</td>
<td>Tabor</td>
<td>48</td>
<td>4989</td>
<td>7348</td>
<td>9104</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>162576</td>
<td>530126</td>
<td>Liberec</td>
<td>36</td>
<td>4990</td>
<td>4516</td>
<td>9198</td>
<td>C</td>
</tr>
</tbody>
</table>

### Atributes in corresponding data matrix

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Family</th>
<th>Matrix</th>
<th>Column</th>
<th>Categories</th>
<th>X-Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Loans</td>
<td>Age</td>
<td>Age</td>
<td>47</td>
<td>No</td>
</tr>
<tr>
<td>Age_in_years</td>
<td>Loans</td>
<td>Age</td>
<td>Age</td>
<td>47</td>
<td>No</td>
</tr>
<tr>
<td>Amount</td>
<td>Loans</td>
<td>amount</td>
<td>amount</td>
<td>6</td>
<td>No</td>
</tr>
<tr>
<td>District</td>
<td>Loans</td>
<td>District</td>
<td>District</td>
<td>77</td>
<td>No</td>
</tr>
<tr>
<td>Duration</td>
<td>Loans</td>
<td>duration</td>
<td>duration</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>Quality</td>
<td>Loans</td>
<td>status</td>
<td>status</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Repayment</td>
<td>Loans</td>
<td>payments</td>
<td>payments</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Salary</td>
<td>Loans</td>
<td>Salary</td>
<td>Salary</td>
<td>3</td>
<td>No</td>
</tr>
<tr>
<td>Sex</td>
<td>Loans</td>
<td>Sex</td>
<td>Sex</td>
<td>2</td>
<td>No</td>
</tr>
</tbody>
</table>
Procesing of more database tables

If you want to analyze data spreaded to more database tables, it is necessary to create one summarizing table. For this purpose use SQL, LMDataSource module is not established for this.
Table of content

- Introduction to LMDataSource module
- Definition of the primary key
- Definition of attributes
- Frequency analysis
Starting with a new data

1) Preparation of a new metabase (see Tutorial First steps)
2) Definition of a primary key for the chosen table of analysed database.
Definition of the primary key, step 1

**LMDataSource.exe**

**LM LMExample.mdb** is a filename assigned by LMAadmin module

Select Attributes List
Definition of the primary key, step 2
Definition of the primary key, step 3

Select proper row and click Primary key
Definition of the primary key, step 3

Primary key is now assigned to column loan_id
Definition of the primary key, step 4

It is necessary to test if the primary key is unique value (by Check button).
Definition of the primary key, step 4

If no duplicities are found, it is possible to use this column as a primary key.
Table of content

- Introduction to LMDataSource module
- Definition of the primary key
- **Definition of attributes**
- Frequency analysis
Definition of attributes

- **How to define an attribute**
- Definition of categories – Each value – one category
- Definition of categories – equidistant intervals
- Definition of categories – equifrequency intervals
- Definition of categories – By values in associated table
How to define attributes

In LMDDataSource.exe, select Database, Attribute List
How to define attributes

**Step 1**: Data matrix selection (can be more than one as in this example)

Click Add and in the new windows, select proper data matrix.
How to define attributes

**Step 2**: Selection of the column, from which the attribute will be created.

Now select proper column and click Create attribute
How to define attributes

**Step 3:** Naming the attribute, default name = name of the selected column

![Attribute definition dialog box](image)

**Step 4:** Selection of the type of creation

![Type of creation options](image)
Definition of attributes

- How to define an attribute
- **Definition of categories** – Each value - one category
- Definition of categories – equidistant intervals
- Definition of categories – equifrequency intervals
- Definition of categories – By values in associated table
Definition of categories – Each value - one category

Each value – one category creates attribute based on the number of unique values in the proper column of database table. It is valuable for string values.
Category definition – Each value, one category

Step 1: choosing the type of creation

![Image of category definition interface with 'Each value - one category' selected]
Category definition – Each value, one category

Step 2: frequency analysis, to see count of individual values
Category definition – Each value, one category

**Step 2**: frequency analysis, to see count for individual values

There is possibility to order frequencies...
Definition of attributes

- How to define an attribute
- Definition of categories – Each value – one category
  - Definition of categories – equidistant intervals
  - Definition of categories – equifrequency intervals
  - Definition of categories – By values in associated table
Category definition – equidistant intervals

Equidistant interval enables to divide values into intervals with predefined length. Intervals can be either opened or closed from both sides.
Category definition – equidistant intervals

Example:

Let's assume we would like to create attribute „Salary“ which will be divided into intervals 500 long, from 8000. (8000;8500>, (8500;9000), ...
Category definition – equidistant intervals

Step 1: choosing the type of creation, setting the origin of interval, its length and closed from.
Category definition – equidistant intervals

Now we have 10 categories which covers all values in Salary column.
Category definition – equidistant intervals

Step 2: frequency analysis.

As we can see, there is 5 categories which have low or zero frequency. Let's join those intervals.
Category definition – equidistant intervals

Step 3: joining intervals.
Select five columns (from 10500 to 12500) and click Join button.
Category definition – equidistant intervals

**Step 4:** renaming categories.
We can rename newly created category by clicking Category button.
Category definition – equidistant intervals

Step 4: Renaming categories

Then select Edit category.

And type the new name.
Category definition – equidistant intervals

Now we can check again frequency analysis. Every category now contains significant number of values.
Definition of attributes

- How to define an attribute
- Definition of categories – Each value – one category
- Definition of categories – equidistant intervals
- **Definition of categories – equifrequency intervals**
- Definition of categories – By values in associated table
Category definition - equifrequency interval

Equifrequency interval divides values into entered number of intervals. These intervals have approximately same number of values.
Category definition - equifrequency interval

Example:

Let's assume we would like to create another attribute from salary. Now we want to have 5 categories, which contains approximately same number of clients.
Category definition - equifrequency interval

**Step 1**: choosing the type of creation, setting the number of interval we would like to create.
Category definition - equifrequency interval

As we can see, we have five intervals with different range.
Category definition - equifrequency interval

Step 2: frequency analysis.
From the frequency analysis it is clear that we have divided values into five intervals with approx. length.
Category definition - equifrequency interval

**Step 3:** renaming intervals.
We can rename intervals to have categories like this.
Definition of attributes

- How to define an attribute
- Definition of categories – Each value – one category
- Definition of categories – equidistant intervals
- Definition of categories – equifrequency intervals
- Definition of categories – By values in associated table
Category definition – by values in associated table

Special type of automatic categorization, usable for pre prepared values (e.g. answers from the questionnaire) and in the database there is a codebook table which converts this codes into text explanation.
Table of content

- Introduction to LMDataSource module
- Definition of the primary key
- Definition of attributes
- Frequency analysis
Frequency analysis - example

LISp-Miner offers some analysis tools, for example contingency tables.
Frequency analysis - example
Frequency analysis - example

For this frequency analysis (previous slide), there was created new attribute Age from the birth_number column of the Loan table.