Reshape

Flexible data restructuring with R

Hadley Wickham
Statistics, Iowa State University
Introduction

• What is reshaping?
• Why it’s not easy at the moment
• Example
• Details
• Future work
Aggregating

- Reduce big table to small table
- (must lose information)
- Each cell in the new table corresponds to multiple cells in the old table
Reshaping

• Like aggregating, but each new cell corresponds to one old cell
• Useful when investigating relationships between different aspects of your data (and especially when using lattice graphics)
• Similar to transposing a matrix (but what happens to ragged data?)
Which one do I use?
Motivation

- Many different tools in R
- Tend to be rather specialised/limited
- Can be difficult to figure out which one to use
- Reshape handles many different needs within one framework
Example
Conceptual framework

- Id vs. measured variables
- random variables vs. their indices
- categorical vs. continuous
- Gives a more flexible data format
- Deals with very ragged data
- Missing values implicit
<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>20</td>
<td>1.95</td>
<td>100</td>
</tr>
<tr>
<td>John</td>
<td>21</td>
<td>1.96</td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>20</td>
<td>Height</td>
<td>1.95</td>
</tr>
<tr>
<td>John</td>
<td>20</td>
<td>Weight</td>
<td>100</td>
</tr>
<tr>
<td>John</td>
<td>21</td>
<td>Height</td>
<td>1.96</td>
</tr>
</tbody>
</table>
Mechanics

• Specifying the output format
• Adding margins
• Functions that return multiple values
• Row and column names
Output specification

- How do you specify what output you want?
- I’ve used a formula type interface
- Formatting output
- What are other alternatives?
Margins

- Multiple levels of aggregation at the same time
- Useful for summaries
- (Pivot table inspired)
Row & column names

- Explicit vs implicit
- Most inbuilt functions store implicitly (frustrating when trying to plot!)
- Reshape stores explicitly (but makes it easy to get rid of them)
Efficiency

- Size (limited by memory)
- Multiple copies of data
- 150,000 x 5
- Speed
  - merge is slow
- RDMS?
Future

- Aggregate function on data frame
- Non-numeric data and summaries
- Built-in graphical display of output
- Larger data/database integration
http://had.co.nz/reshape