Converting Tables to LongForm Dataframes

Jimmy Oh - Department of Statistics

1 Motivation

A lot of useful data is out there, data that could help you make better decisions through a better understanding of our world. Unfortunately, many data releases are still made for direct human consumption and are not directly machine readable - a significant barrier to effective utilisation of the data. One symptom of this is the release of data in tabular form (*Table*) that can only be understood after identifying patterns and discerning the structure of the Table, a task easy for a human brain but rather difficult for a computer.

	1	2	3	4	5	6	7	8	9	10	11
1	Labour Force Status by Sex by Sing/Comb Ethnic Group (Qrtly−Mar/Jun/Sep/Dec)										
2		Male									
3		European Only								Maori Only	
4		Persons Em	Persons Un	Not in Labo	Working Ag	Labour Forc	Unemploym	Employmen	Total Labou	Persons Em	Persons Un
5	2007Q4	855.8	20.0	280.0	1155.8	75.8	2.3	74.0	875.8	71.1	6.1
6	2008Q1	863.0	25.4	283.5	1171.9	75.8	2.9	73.6	888.5	69.1	7.5
7	2008Q2	850.1	26.0	280.7	1156.8	75.7	3.0	73.5	876.1	67.2	5.7
8	2008Q3	839.6	29.8	285.9	1155.3	75.2	3.4	72.7	869.4	71.7	8.7
9	2008Q4	854.8	29.5	274.7	1158.9	76.3	3.3	73.8	884.2	76.1	8.5
10	2009Q1	845.0	35.4	279.4	1159.8	75.9	4.0	72.9	880.4	75.4	8.4
11	2009Q2	831.6	34.9	279.7	1146.2	75.6	4.0	72.6	866.5	74.2	9.9
12	2009Q3	813.3	42.5	290.4	1146.2	74.7	5.0	71.0	855.8	70.9	10.9
13	2009Q4	831.1	40.1	277.0	1148.2	75.9	4.6	72.4	871.2	71.7	13.6
14	2010Q1	822.5	36.4	283.2	1142.1	75.2	4.2	72.0	858.9	71.8	11.3

Figure 1: An example of a hierarchical Table. The Table is of the Labour Force Status data[a] and in total spans 240 columns, making it suitable for neither man nor machine. Restructuring the data is a significant challenge, in turn making any serious statistical analyses very time consuming.

[a] Infoshare, Statistics New Zealand (2013). URL http://www.stats.govt.nz/infoshare/

LongForm is a simple alternative method of presenting the data that, due to its simplicity, is both easy to implement and is machine readable, greatly enhancing potential applications of the data. This is where TableToLongForm comes in, providing a way to automatically convert hierarchical Tables to a simple LongForm Dataframe, thus enabling much greater utilisation of the data.
No other research that attempts such a generalised conversion could be found, potentially making this a world first.

	1	2	3	4	5	6	7	8	9	10	11
1				Persons Em	Persons Un	Not in Labo	Working Ag	Labour Forc	Unemploym	Employmen	Total Labou
2	Male	European C	2007Q4	855.8	20	280	1155.8	75.8	2.3	74	875.8
3	Male	European C	2008Q1	863	25.4	283.5	1171.9	75.8	2.9	73.6	888.5
4	Male	European C	2008Q2	850.1	26	280.7	1156.8	75.7	3	73.5	876.1
5	Male	European C	2008Q3	839.6	29.8	285.9	1155.3	75.2	3.4	72.7	869.4
6	Male	European C	2008Q4	854.8	29.5	274.7	1158.9	76.3	3.3	73.8	884.2
7	Male	European C	2009Q1	845	35.4	279.4	1159.8	75.9	4	72.9	880.4
8	Male	European C	2009Q2	831.6	34.9	279.7	1146.2	75.6	4	72.6	866.5
9	Male	European C	2009Q3	813.3	42.5	290.4	1146.2	74.7	5	71	855.8
10	Male	European C	2009Q4	831.1	40.1	277	1148.2	75.9	4.6	72.4	871.2
11	Male	European C	2010Q1	822.5	36.4	283.2	1142.1	75.2	4.2	72	858.9
12	Male	European C	2010Q2	825.3	39.9	290.1	1155.3	74.9	4.6	71.4	865.2
13	Male	European C	2010Q3	836.9	31	287.1	1155.1	75.1	3.6	72.5	867.9
14	Male	European C	2010Q4	838.1	39.6	277.1	1154.8	76	4.5	72.6	877.7

Figure 2: An example of a LongForm Dataframe. This is the Table in Figure 1 above after automatic conversion with TableToLongForm. While it spans 660 rows, it is easily manipulated by computer software, making it possible to use the data in complex ways efficiently.

2 Writing the Algorithms

TableToLongForm consists of a number of algorithms written in the Statistical Software **R** that can collectively process a variety of so-called *Recognised Patterns* of hierarchical structure. Any Table that consists of some combination of the Recognised Patterns can be automatically converted with TableToLongForm. Here we outline how the algorithms that process the Recognised Patterns are written.

2.1 Find a Problem Table

	1	2
1	Accounting	
2	NZ Maori	Male
3		Female
4	NZ European	Male
5		Female
6	Biology	
7	NZ Maori	Male
8		Female
9	NZ European	Male
10		Female

This is an excerpt from a table[b], showing only the table's row labels. From this information the computer must understand that Accounting and Biology are the top-most *Headings*, followed by the Ethnicities as *Sub-Headings*, followed lastly by Gender.

2.2 Identify Tell-tale Features

	1	2
1	Accounting	
2	NZ Maori	Male
3		Female
4	NZ European	Male
5		Female
6	Biology	
7	NZ Maori	Male
8		Female
9	NZ European	Male
10		Female

We look for tell-tale features that give away the relative hierarchies of these headings. For instance, we notice that the Subject headings have empty cells to the right (highlighted in red), and that the Ethnic sub-headings have empty cells below (highlighted in green).

2.3 Sub-divide into Smaller Pieces

	1	2
1	Accounting	
2	NZ Maori	Male
3		Female
4	NZ European	Male
5		Female
6	Biology	
7	NZ Maori	Male
8		Female
9	NZ European	Male
10		Female

	1	2
1	Accounting	
2	NZ Maori	Male
3		Female
4	NZ European	Male
5		Female
6	Biology	
7	NZ Maori	Male
8		Female
9	NZ European	Male
10		Female

	1	2
1	Accounting	
2	NZ Maori	Male
3		Female
4	NZ European	Male
5		Female
6	Biology	
7	NZ Maori	Male
8		Female
9	NZ European	Male
10		Female

We can use these tell-tale features to sub-divide the labels into smaller hierarchies, essentially breaking down the table into a number of smaller, simpler tables. By employing a variety of algorithms to handle the different patterns, we can repeat the process recursively on these smaller tables until we reach the lowest level of headings. Once we understand the hierarchical structure of the labels (for both the row and column labels), we can use this knowledge to reconstruct the Table into a LongForm Dataframe. Of course this is the greatly simplified explanation, for the nitty-gritty details refer to the Literate Document[c].

[b] New Zealand Qualifications Authority (2012). URL http://www.nzqa.govt.nz/
[c] URL https://github.com/joh024/TableToLongForm

3 User Manual

Though the algorithms that go into TableToLongForm may be complicated, its usage by the end user is very simple - so simple that the User Manual can fit on a small corner of a poster! If the Table is something TableToLongForm can handle, a single call is all that is required. Additional arguments can be supplied for diagnostic output and more information on this can be found in the Literate Document.

```
## Step 1 - Read in original Table
> LabourForce = read.csv("StatsNZLabourForce.csv")

## Step 2 - Call TableToLongForm
> LabourForce.converted = TableToLongForm(LabourForce)

## Step 3 - All sorts of magic can now be performed
```

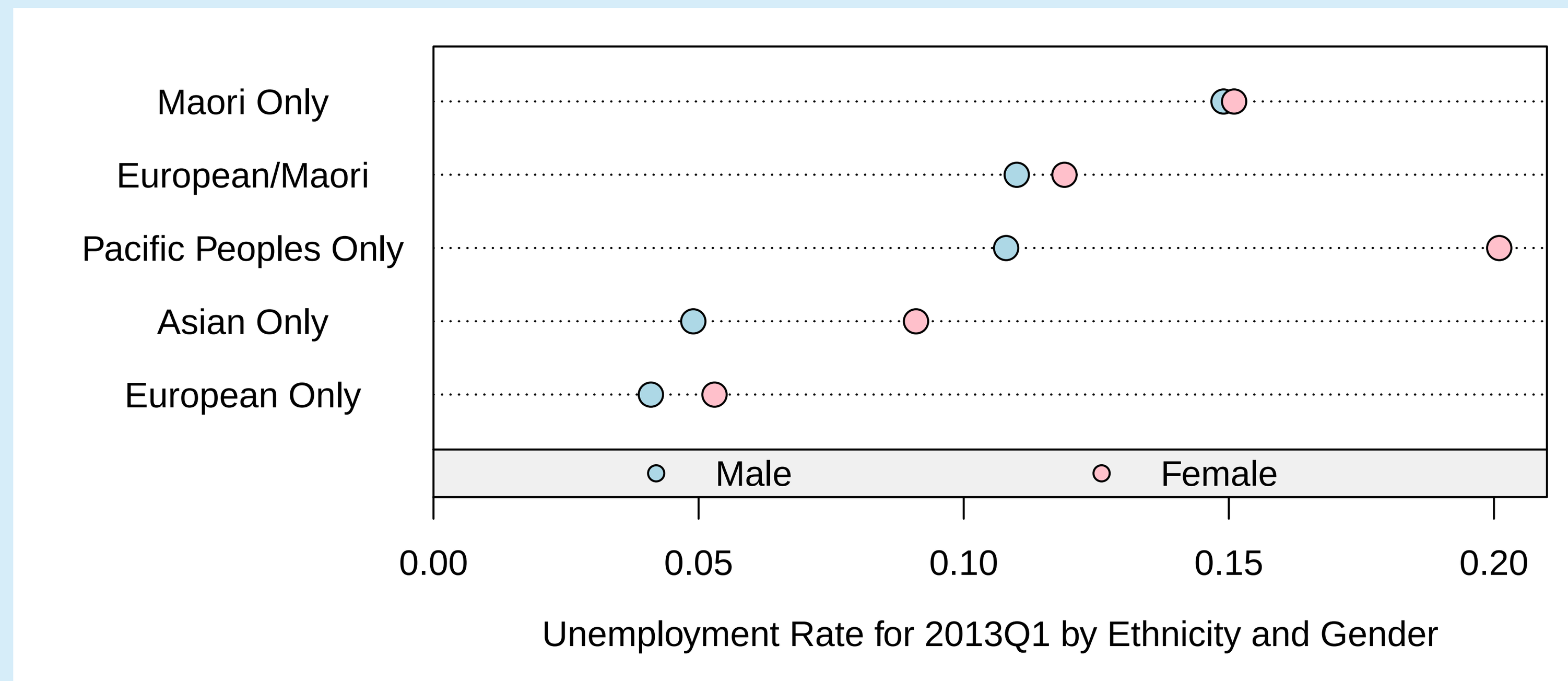

Figure 3: With the data in a convenient Dataframe, it's easy to draw all sorts of plots of the data. Here we see that for Asians and Pacific Peoples, the Unemployment Rate for Females is roughly twice that of the Males. For Europeans, Maori and European/Maori, the rates are roughly the same for both genders.

```
## Step 4 - Profit
```

4 Summary

We have introduced TableToLongForm, an **R** function that can automatically convert hierarchical Tables that would normally rely on the discerning powers of a human brain, to a simple LongForm Dataframe that any decent software package can easily manipulate and use, thereby unlocking the true value of the underlying data. Additional algorithms are planned for TableToLongForm, with the eventual aim of converting most conventional Tables automatically.