Mosaic plots for data visualization

Using the "Adult" census income data set

Anton Antonov
MathematicaForPrediction project GitHub
MathematicaForPrediction blog at WordPress
March 2014

Introduction

This document gives a description and examples of using the function `MosaicPlot` of the *Mathematica* package MosaicPlot.m provided by the project MathematicaForPrediction at GitHub, see [1].

The function `MosaicPlot` summarizes the conditional probabilities of co-occurrence of the categorical values in a list of records of the same length. The list of records is assumed to be a full array and the columns to represent categorical values. (Note, that if a column is numerical but has a small number of different values then it can be seen as categorical.)

I have read the descriptions of mosaic plots in the book "R in Action" by Robert Kabakoff, [2], and one of the references provided in the book ("What is a mosaic plot?" by Steve Simon, [3]). I was impressed how informative mosaic plots are and I figured they can be relatively easily implemented using Prefix trees (also known as "Tries") [4,5]. I implemented `MosaicPlot` while working on a document analyzing the census income data from 1998, [6]. This is the reason that data set is used in this document. A good alternative set provided by *Mathematica* is `ExampleData[{"Statistics","USCars1993"}]`.

Data set

The data set can be found and taken from http://archive.ics.uci.edu/ml/datasets/Census+Income, [6].

The description of the data set is given in the file "adult.names" of the data folder. The data folder provides two sets with the same type of data "adult.data" and "adult.test"; the former is used for training, the latter for testing.

The total number of records in the file "adult.data" is 32 561; the total number of records in the file "adult.test" is 16 281.

Here is how the data looks like:

	age	workclass	fnlwgt	education	education−num	marital−status	occupation	relationship
1	37	Private	182 675	Some-college	10	Married-civ-spouse	Exec-managerial	Wife
2	24	Private	333 505	HS-grad	9	Married-spouse-absent	Transport-moving	Own-child
3	45	State-gov	36 032	HS-grad	9	Divorced	Protective-serv	Unmarried
4	30	Private	202 450	HS-grad	9	Married-civ-spouse	Transport-moving	Husband
5	20	Private	194 630	HS-grad	9	Married-civ-spouse	Adm-clerical	Husband
6	17	?	170 320	11th	7	Never-married	?	Own-child
7	39	Private	255 503	Bachelors	13	Never-married	Exec-managerial	Not-in-family
8	40	Private	240 124	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband
9	46	Private	321 327	Some-college	10	Married-civ-spouse	Transport-moving	Husband
10	18	Private	109 702	Some-college	10	Never-married	Sales	Own-child
11	28	Private	125 527	Some-college	10	Never-married	Sales	Not-in-family
12	20	Private	164 219	HS-grad	9	Never-married	Handlers-cleaners	Own-child
13	36	Private	32 334	Assoc-voc	11	Married-civ-spouse	Exec-managerial	Wife
14	63	?	257 659	Masters	14	Never-married	?	Not-in-family
15	19	Private	358 631	HS-grad	9	Never-married	Adm-clerical	Not-in-family
16	45	Private	329 603	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband
17	45	Private	101 320	HS-grad	9	Divorced	Adm-clerical	Unmarried
18	19	Private	307 496	Some-college	10	Never-married	Other-service	Own-child
19	25	Private	112 847	HS-grad	9	Married-civ-spouse	Transport-moving	Own-child
20	27	Local-gov	162 404	HS-grad	9	Never-married	Protective-serv	Not-in-family

Since I did not understand the meaning of the column “fnlwgt” I dropped it from the data.

Here is the summary table of the data:

1 age		2 workclass		3 education		4 education−num		5 marital−status	
Min	17.	Private	22 696	HS-grad	10 501	Min	1.	Married-civ-spouse	14 976
1st Qu	28.	Self-emp-not-inc	2541	Some-college	7291	1st Qu	9.	Never-married	10 683
Median	37.	Local-gov	2093	Bachelors	5355	Median	10.	Divorced	4443
Mean	38.5816	?	1836	Masters	1723	Mean	10.0807	Separated	1025
3rd Qu	48.	State-gov	1298	Assoc-voc	1382	3rd Qu	12.	Widowed	993
Max	90.	Self-emp-inc	1116	11th	1175	Max	16.	Married-spouse-absent	418
		(Other)	981	(Other)	5134			Married-AF-spouse	23

6 occupation		7 relationship		8 race		9 sex		10 capital−gain	
Prof-specialty	4140	Husband	13 193	White	27 816	Male	21 790	1st Qu	0.
Craft-repair	4099	Not-in-family	8305	Black	3124	Female	10 771	3rd Qu	0.
Exec-managerial	4066	Own-child	5068	Asian-Pac-Islander	1039			Median	0.
Adm-clerical	3770	Unmarried	3446	Amer-Indian-Eskimo	311			Min	0.
Sales	3650	Wife	1568	Other	271			Mean	1077.65
Other-service	3295	Other-relative	981					Max	99 999.
(Other)	9541								

11 capital−loss		12 hours−per−week		13 native−country		14 income	
1st Qu	0.	Min	1.	United-States	29 170	<=50K	24 720
3rd Qu	0.	1st Qu	40.	Mexico	643	>50K	7841
Median	0.	Median	40.	?	583		
Min	0.	Mean	40.4375	Philippines	198		
Mean	87.3038	3rd Qu	45.	Germany	137		
Max	4356.	Max	99.	Canada	121		
				(Other)	1709		

On the summary table the numerical variables are described with min, max, and quartiles. The category variables are described with the tallies of their values. The tallies of values are ordered in decreasing order. The tallies of truncated values are summed under the value “(Other)”.

Note that:
-- only 24% of the labels are “>50K”;
-- 2/3 of the records are for males;
-- “capital-gain” and “capital-loss” are very skewed.

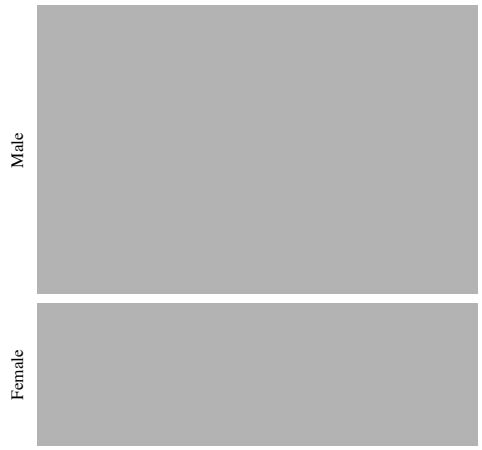
Load the package

```
Get["~/MathFiles/MathematicaForPrediction/MosaicPlot.m"]
```

Explanations

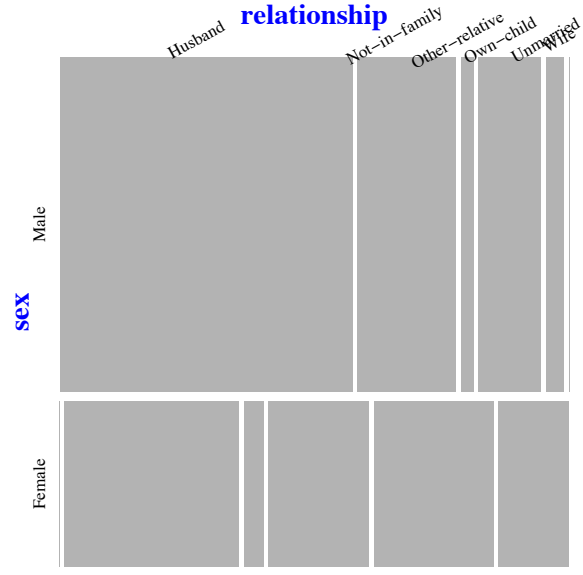
If we pick a categorical variable, say “sex”, we can visualize the frequencies of the appearance of the variable values with the following plot:

```
MosaicPlot[censusData[[All, {9}]],
 ColorRules → {_ → GrayLevel[0.7]}, ImageSize → 250]
```



The size of the rectangles depends on the frequencies of appearance of the values “Male” and “Female” in the data records. From the rectangle sizes we can see what we already knew from the data summary table: approximately $2/3$ of the records are about males.

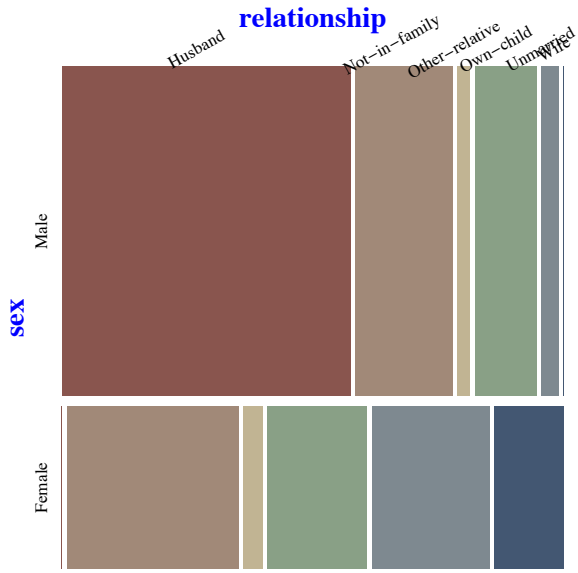
We can subdivide every rectangle *r* according to the frequencies of co-occurrence of *r*’s value with the values of a second categorical variable, say “relationship”:



The labels corresponding to the values of “relationship” are rotated for legibility. The "relationship" labels are placed according to the co-occurrence with the value "Male" of the variable "sex". The correspondent fractions of the pairs ("Female","Husband"), ("Female","Not-in-family"), etc., are deduced from the order of the "relationship" labels.

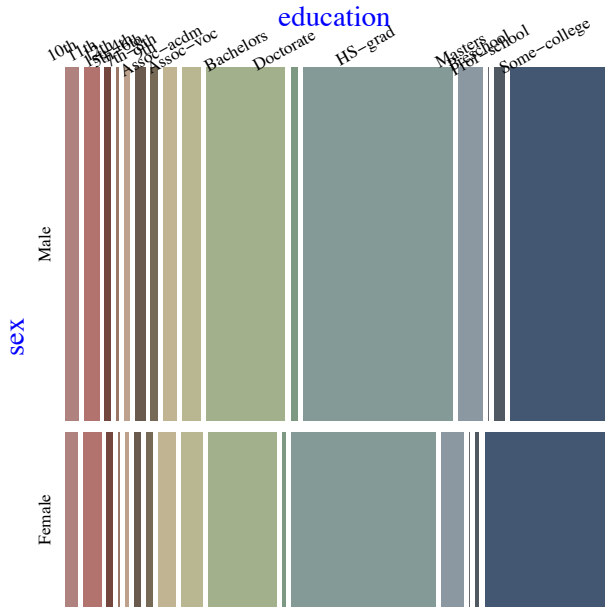
Using colored mosaic plots can help distinguishing which rectangles correspond to which values. Here

is the last plot with rectangles colored across the "relationship" data variable:



From the visual representations of the “sex vs. relationship” mosaic plot we can see that large fraction of the males are husbands, none (or a very small fraction) of them are wives. We can also see that none (or a very small fraction) of the females are husbands, the largest fraction of them are “Not-in-family”, and the “Not-in-family” females are approximately three times more than the females that are wives.

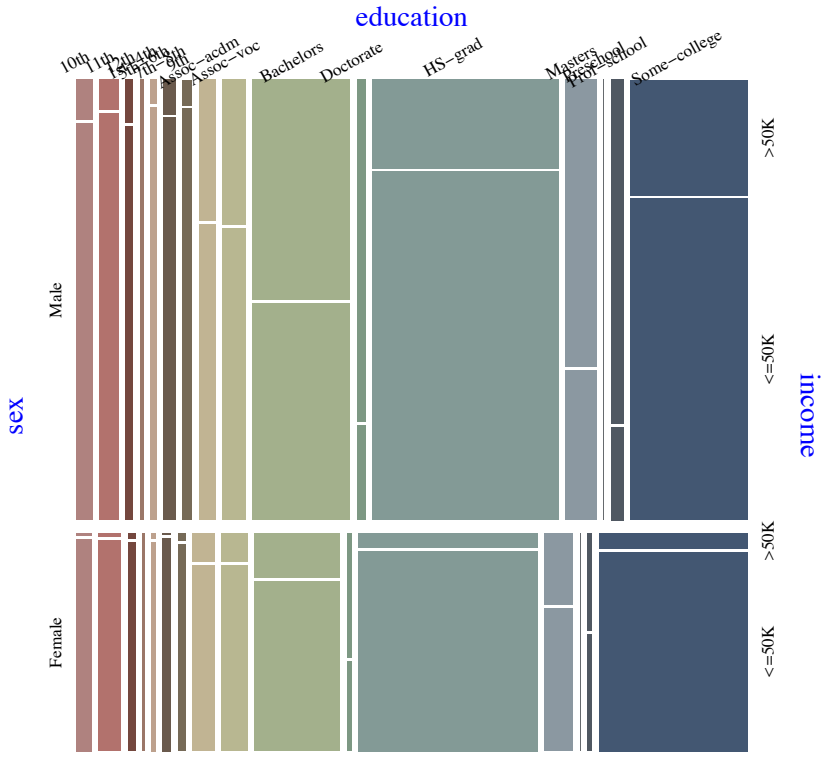
Let us make another mosaic plot for a different kind of relationship, “sex vs. education”:



By comparing the sizes of the rectangles corresponding to the values “Bachelors”, “Doctorate”, “Masters”, and “Some-college” on the “sex vs. education” mosaic plot we can see that the fraction of men that have finished college is larger than the fraction of women that have finished college.

We can further subdivide the rectangles according the co-occurrence frequencies with a third categorical variable. We are going to choose that third variable to be “income”, the values of which can be seen

as outcomes or consequents of the values of the first two variables of the mosaic plot.

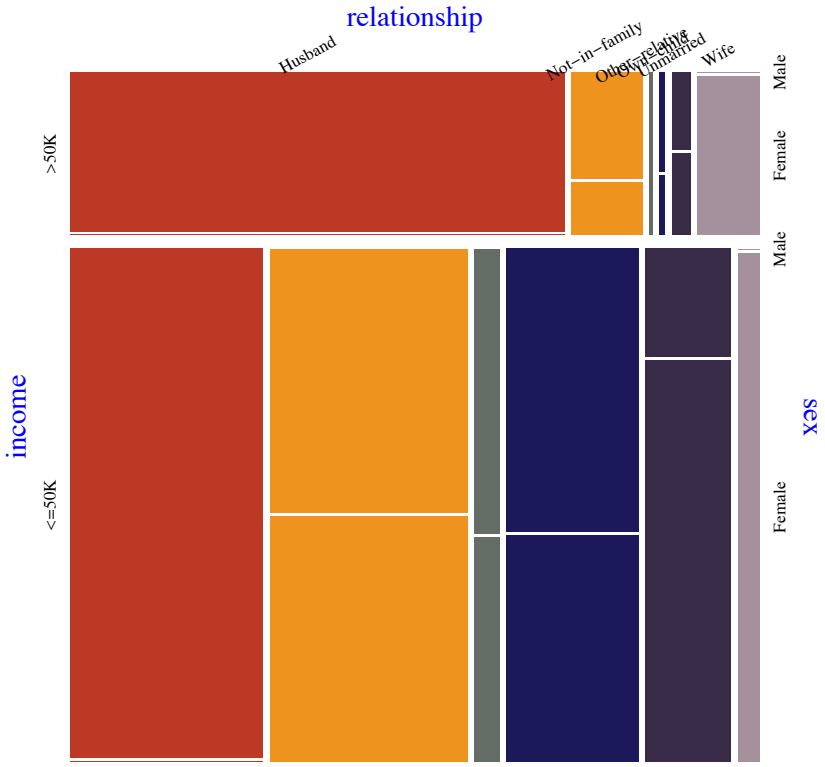


From the mosaic plot “sex vs. education vs. income” we can make the following observations.

1. Approximately 75% of the males with doctorate degrees or with a professional school degree earn more than $50 000 per year.
2. Approximately 60% of the females with a doctorate degree earn more than $50 000 per year.
3. Approximately 45% of the females with a professional school degree earn more than $50 000.
4. Across all education type females are (much) less likely to earn more than $50 000 per year.

(The exact numbers of these observations can be seen tooltip table shown when hovering with the mouse over the rectangles.)

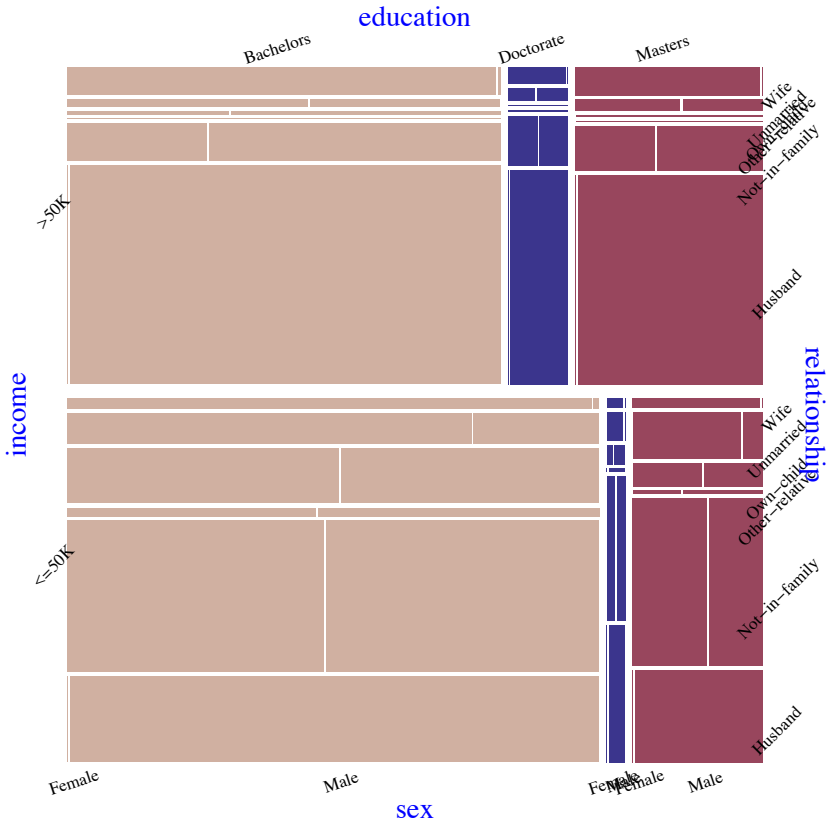
Instead of having the consequent (or outcome) variable to be the last variable in the mosaic plot, it is also useful to start with the consequent variable to get a perspective of how the attributes breakdown for it. Here is an example of a mosaic plot for “income vs. relationship vs. sex” (using a different color scheme):



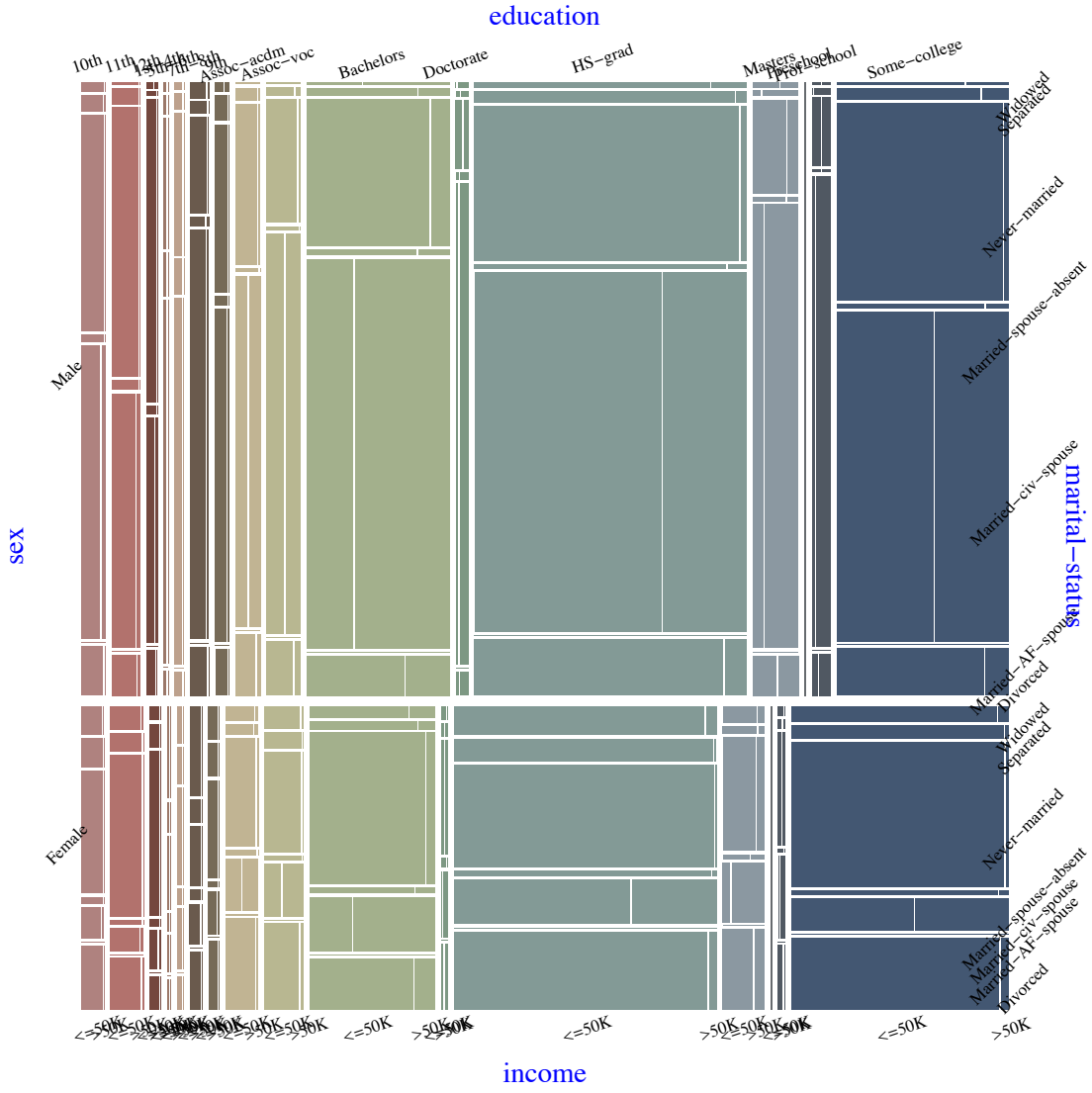
From the mosaic plot “income vs. relationship vs. sex” we can see that 75% of the people with income higher than $50 000 are male and husbands. We can also see that large fraction, 30%, of the people with income less than $50 000 are “Not-in-family” and they are equally likely to be male or female. People who are not in a family are only 10% of the people with income higher than $50 000.

It might be useful to make a mosaic plot for a subset of the records. Here is an example of a mosaic plot with splitting across four columns made only for people who have bachelor, master, or doctorate degrees:

```
Block[{censusData =
    Cases[censusData, {___, "Bachelors" | "Masters" | "Doctorate", ___}]},
 MosaicPlot[censusData[[All, {14, 3, 7, 9}]], "Gap" → 0.02,
  "GapFactor" → 0.56, "ColumnNamesOffset" → 0.07, "ColumnNames" →
   Map[Style[#, Blue, FontSize → 15] &, columnNames[[{14, 3, 7, 9}]]],
  "LabelRotation" → {{3, 1}, {1, 1}},
  ColorRules → {2 → ColorData[13, "ColorList"]}, ImageSize → 430]]
```



Similar to the previous mosaic plot is this plot of “sex vs. education vs. marital-status vs. income”:



Options

MosaicPlot takes the following options:

```
ColumnNames → None
ColumnNamesOffset → 0.05
ExpandLastColumn → False
FirstAxis → y
Gap → 0.02
GapFactor → 0.5
LabelRotation → {{1, 0}, {0, 1}}
LabelStyle → {}
Tooltips → True
ZeroProbability → 0.001
ColorRules → Automatic
```

In addition, MosaicPlot takes all the options of Graphics. (Because MosaicPlot is implemented with Graphics.)

The options are explained in the sub-sections below.

Visualizing categorical columns + a numerical column ("ExpandLastColumn")

If the last data column is numerical then MosaicPlot can use it as pre-computed contingency statistics. This functionality is specified with the option "ExpandLastColumn"→True.

In order to explain the functionality we are going to use following interpretation. If the last of column of the data is numerical then we can treat the data as a contracted version of a longer list of records made only of the categorical columns. For example, consider the following table with observations of people's hair and eyes color:

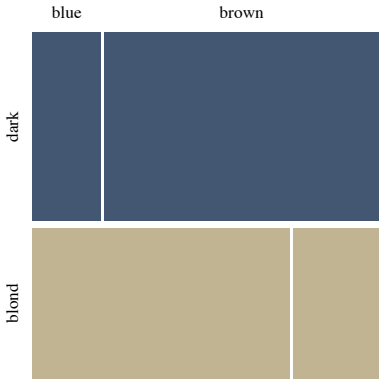
hair color	eyes color	number of observations
blond	blue	3
blond	brown	1
dark	blue	1
dark	brown	4

The table above can be considered as a contracted version of this table:

hair color	eyes color
blond	blue
blond	blue
blond	blue
blond	brown
dark	blue
dark	brown
dark	brown
dark	brown
dark	brown

Setting the option "ExpandLastColumn" to True gives a mosaic plot corresponding to that latter, observations-expanded table:

```
MosaicPlot[sData, "ExpandLastColumn" → True, ImageSize → 200]
```

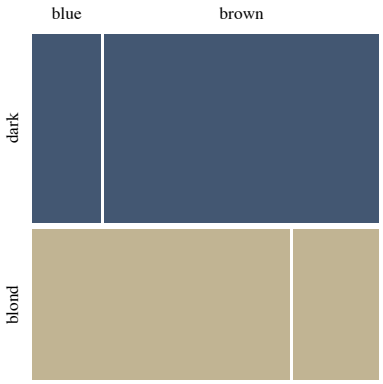


The last data column (which is numerical) does not need to be made of integers:

```
sData[[All, 3]] = sData[[All, 3]] / 20.;
sData
```

```
{{blond, blue, 0.15}, {blond, brown, 0.05},
 {dark, blue, 0.05}, {dark, brown, 0.2}}
```

```
MosaicPlot[sData, "ExpandLastColumn" → True, ImageSize → 200]
```



Controlling the size of the gap between the rectangles ("Gap" and "GapFactor")

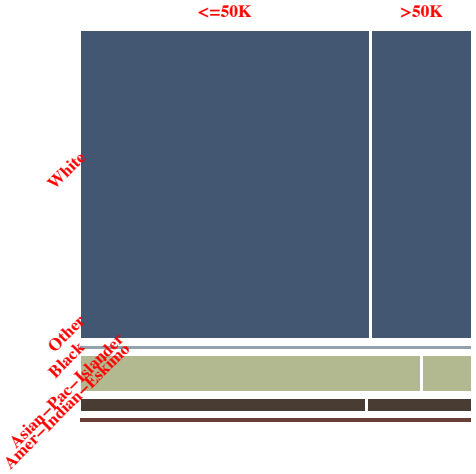
The size of the gaps between the rectangles is controlled with the options "Gap" and "GapFactor". The value "Gap" specifies the size of the gap between the rectangles derived from the first column. `MosaicPlot` splits the data into rectangles recursively. In order to derive the gaps for the subsequent data column the values of "Gap" and "GapFactor" are multiplied. In other words, if `MosaicPlot` is given the options {"Gap"→*g*,"GapFactor"→*f*} then the gap between the rectangles corresponding to the *i*-th column have the size is $g\, f^{(i-1)}$.

Contingency values labels ("LabelRotation" and "LabelStyle")

The labels derived from the distinct values (levels) of each column of the data can be rotated and given style options.

The option "LabelRotation" takes directional specification for `Text` (the fourth argument of `Text`). The option "LabelStyle" takes options and arguments for the function `Style`.

```
MosaicPlot[censusData[All, {8, 14}], "LabelRotation" → {{1, 0}, {1, 1}},
 "LabelStyle" → {Bold, Red, FontFamily → "Times"}, ImageSize → 250]
```



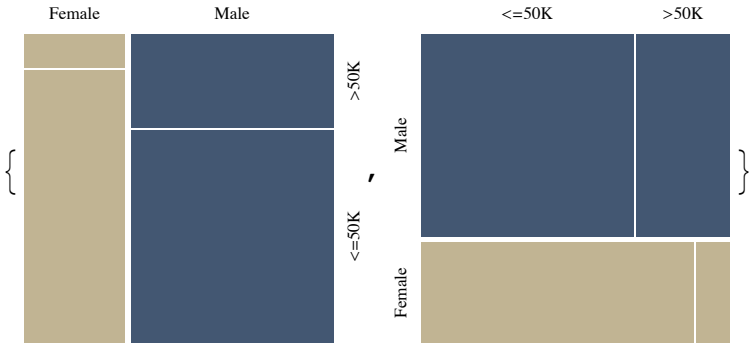
Labels for categorical variables (“ColumnNames” and “ColumnNamesOffset”)

The names of the data columns (data’s variables) are specified with the option “ColumnNames”. (The list of names given to “ColumnNames” can be formatted with `Style`.) The distance of the column names from the rectangles is specified with the option “ColumnNamesOffset”.

Start of the rectangle splitting (“FirstAxis”)

The starting axis of the data splitting is specified by “FirstAxis”.

```
MosaicPlot[censusData[All, {9, 14}], "FirstAxis" → #] & /@ {"x", "y"}
```



Tooltips with exact contingency statistics (“Tooltips”)

`MosaicPlot` has an interactive feature using `Tooltip` that gives a table with the exact co-occurrence (contingency) values when hovering with the mouse over the rectangles. The option “Tooltips” takes the values `True` or `False`.

Visualizing non-existing contingencies (“ZeroProbability”)

The non-existing contingencies have to be represented in the mosaic plot. `MosaicPlot` uses very thin rectangles for them and the size of these rectangles is controlled with the option “ZeroProbability”.

Coloring of the rectangles (ColorRules)

The rectangles can be colored using the option `ColorRules` which specifies how the colors of the rectangles are determined from the indices of the data columns.

More precisely, the values of the option `ColorRules` should be a list of rules, $\{i_1 \to c_1, i_2 \to c_2, \ ...\}$, matching the form

```
{(_Integer → (_RGBColor | _GrayLevel)) ..}.
```

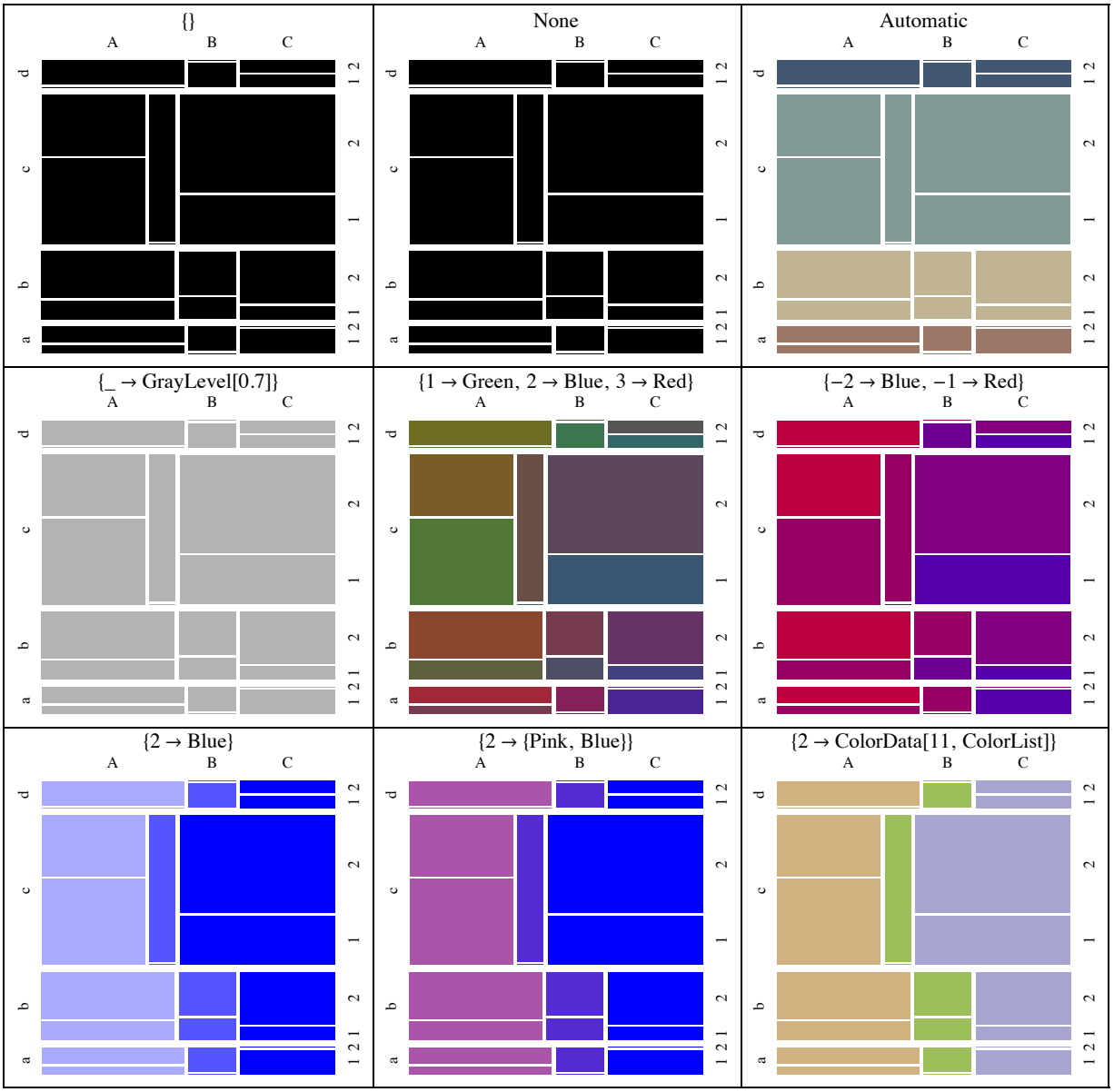
The column indices i_k can be negative (-1 meaning the last column).

If coloring for only one column index is specified the value of `ColorRules` can be of the form

```
{_Integer → {(_RGBColor | _GrayLevel) ..}}.
```

The colors are used with `Blend` in order to color the rectangles according to the order of the unique values of the specified data columns.

The grid of plots below shows mosaic plots of the same data with different values for the option `ColorRules` (given as plot labels).



The default value for `ColorRules` is `Automatic`. When `Automatic` is given to `ColorRules`, `MosaicPlot` finds the data column with the largest number of unique values and colors them according to their order using `ColorData[7,"ColorList"]`.

References

[1] Anton Antonov, Mosaic plot for data visualization implementation in *Mathematica*, source code at GitHub, https://github.com/antononcube/MathematicaForPrediction, package MosaicPlot.m, (2013).

[2] Robert Kabacoff, R in Action, Manning Publications, 1 edition, 2011, URL: http://www.amazon.-com/R-Action-Robert-Kabacoff/dp/1935182390 .

[3] Steve Simon, What is a mosaic plot?, URL: http://www.pmean.com/definitions/mosaic.htm .

[4] Anton Antonov, Tries with frequencies *Mathematica* package, source code at GitHub, https://github.-

com/antononcube/MathematicaForPrediction, package TriesWithFrequencies.m, (2013).

[5] Wikipedia, Trie, http://en.wikipedia.org/wiki/Trie .

[6] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Census Income Data Set, URL: http://archive.ics.uci.edu/ml/datasets/Census+Income .