Simple missing functionalities

Anton Antonov
MathematicaForPrediction at GitHub
MathematicaVsR project at GitHub
November, 2016

Introduction

This notebook belongs to the Mathematica-part of MathematicaVsR at GitHub project DataWrangling.

This notebook illustrates commands that are (in my opinion) are missing from Mathematica but are present and used often in R. See these corresponding R-part HTML file or RMarkdown file.

In this notebook functionalities of those missing commands are obtained by the functions `RecordsSummary`, `VariableDependenceGrid`, `CrossTabulate`, and `MosaicPlot`.

Load packages

The following commands load the packages used in this notebook.

```
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master/
   MathematicaForPredictionUtilities.m"]
```

```
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master/
   MosaicPlot.m"]
```

Data load and rudimentary analysis

Titanic data

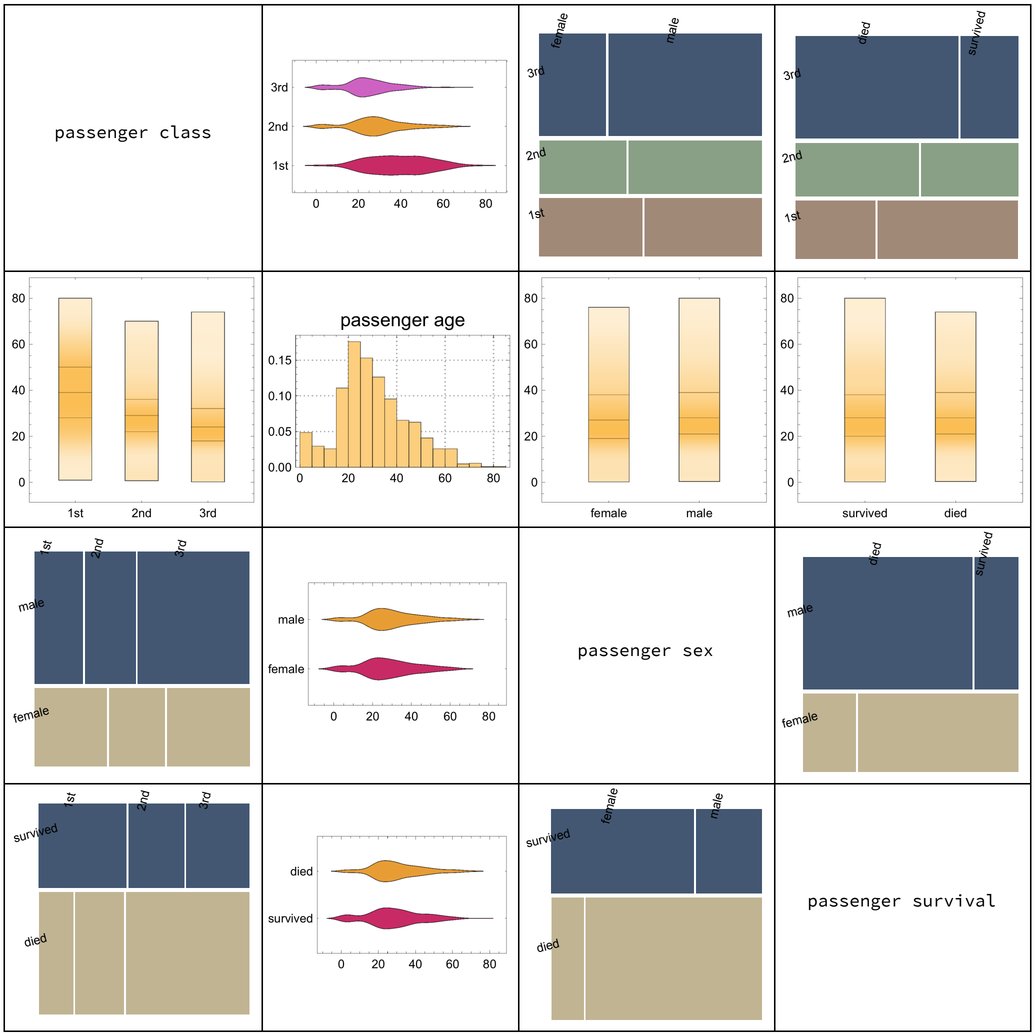
Here is the summary of the Titanic data used below:

```
titanicData =
  (Flatten@*List) @@@ ExampleData[{"MachineLearning", "Titanic"}, "Data"];
columnNames = (Flatten@*List) @@
   ExampleData[{"MachineLearning", "Titanic"}, "VariableDescriptions"];
titanicData = DeleteCases[titanicData, {___, _Missing, ___}];
RecordsSummary[titanicData, columnNames]
```

```
                     2 passenger age
                     Min     0.1667
 1 passenger class   1st Qu  21.       3 passenger sex   4 passenger survival
{3rd  501          , Median  28.     , male    658     , died      619         }
 1st  284            Mean    29.8811   female  388       survived  427
 2nd  261            3rd Qu  39.
                     Max     80.
```

This variable dependence grid shows the relationships between the variables.

```
Magnify[#, 0.7] &@VariableDependenceGrid[titanicData, columnNames]
```



Employee attitude dataset

Here is the summary of the Titanic data used below:

```
eaData = ExampleData[{"Statistics", "EmployeeAttitude"}];
eaColumnNames = (Flatten@*List) @@
   ExampleData[{"Statistics", "EmployeeAttitude"}, "ColumnHeadings"];
Multicolumn[RecordsSummary[N@eaData, eaColumnNames], 3, Dividers → All]
```

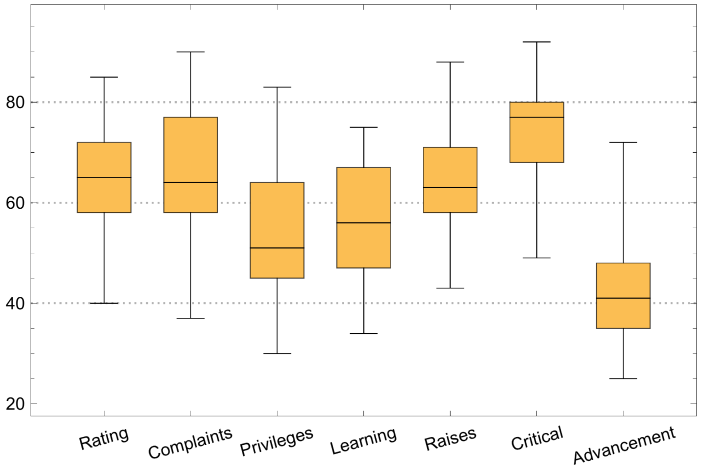
1 Rating Min 40. 1st Qu 58. Mean 64.6333 Median 65.5 3rd Qu 72. Max 85.	4 Learning Min 34. 1st Qu 47. Mean 56.3667 Median 56.5 3rd Qu 67. Max 75.	7 Advancement Min 25. 1st Qu 35. Median 41. Mean 42.9333 3rd Qu 48. Max 72.
2 Complaints Min 37. 1st Qu 58. Median 65. Mean 66.6 3rd Qu 77. Max 90.	5 Raises Min 43. 1st Qu 58. Median 63.5 Mean 64.6333 3rd Qu 71. Max 88.	
3 Privileges Min 30. 1st Qu 45. Median 51.5 Mean 53.1333 3rd Qu 64. Max 83.	6 Critical Min 49. 1st Qu 68. Mean 74.7667 Median 77.5 3rd Qu 80. Max 92.	

It is a good idea to get an impression of the numerical variables distributions in a given dataset.

There are several approaches for doing this (in Mathematica and in general.)

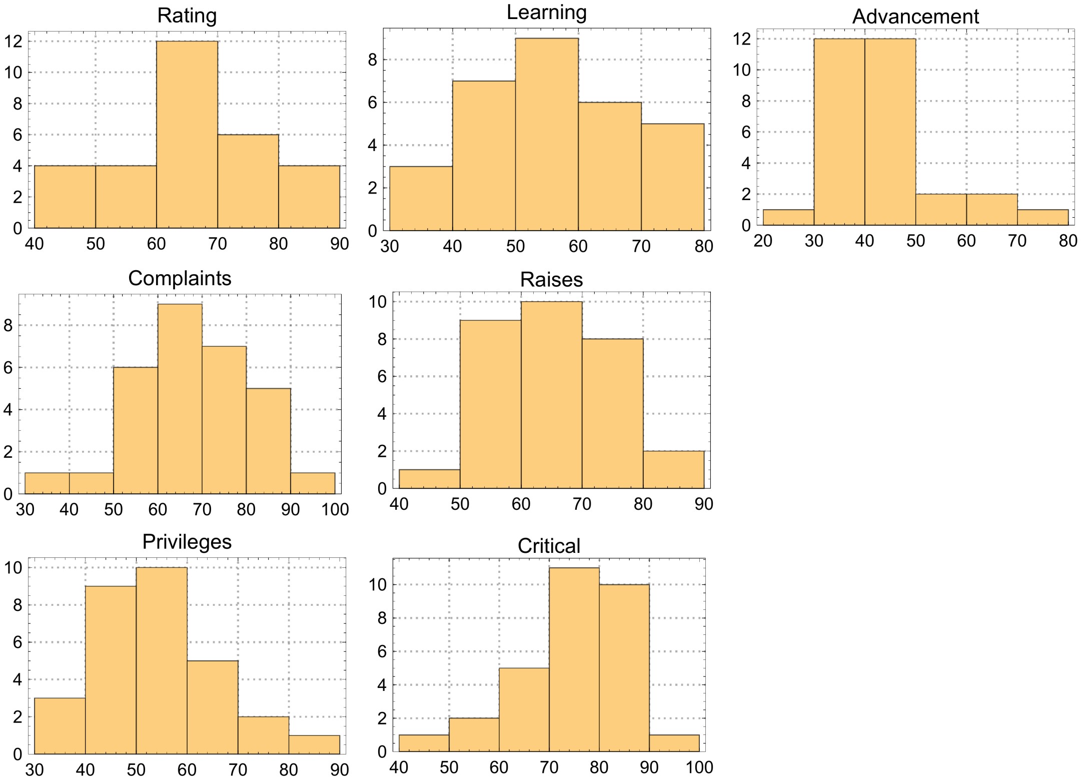
Box-and-whisker diagrams

```
DistributionChart[Transpose[eaData], ChartElementFunction → "BoxWhisker",
 ChartLabels → Map[Rotate[#, π / 12] &, eaColumnNames], PlotTheme -> "Detailed"]
```



Panel of histograms

```
Multicolumn[
 MapThread[Histogram[#1, PlotLabel → #2, PlotRange → All, PlotTheme -> "Detailed"] &,
  {Transpose[eaData], eaColumnNames}], 3]
```



Cross tabulation and mosaic plots

Cross tabulation

In statistics contingency tables are matrices used to show the co-occurrence of variable values of multi-dimensional data. They are fundamental in many types of research. Below are some examples of cross-tabulation. For a detailed discussion see the Markdown file “Contingency-tables-creation-examples.md” of this project or the corresponding PDF file.

```
CrossTabulate[titanicData〚All, {1, 3}〛] // MatrixForm
```

```
    | female  male
----+-------------
1st |  133    151
2nd |  103    158
3rd |  152    349
```

A generalization of `CrossTabulate` is the function `CrossTensorate` implemented in Mathematica-ForPredictionUtilities.m that takes a “formula” argument similar to R’s xtabs.

```
CrossTensorate[Count == "passenger class" + "passenger sex" + "passenger survival",
  titanicData, columnNames] // MatrixForm
```

```
    |       female                  male
----+------------------------------------------
1st | (   died   |  5  )   (   died   | 98 )
    | ( survived | 128 )   ( survived | 53 )
2nd | (   died   | 11 )    (   died   | 135 )
    | ( survived | 92 )    ( survived |  23 )
3rd | (   died   | 80 )    (   died   | 290 )
    | ( survived | 72 )    ( survived |  59 )
```

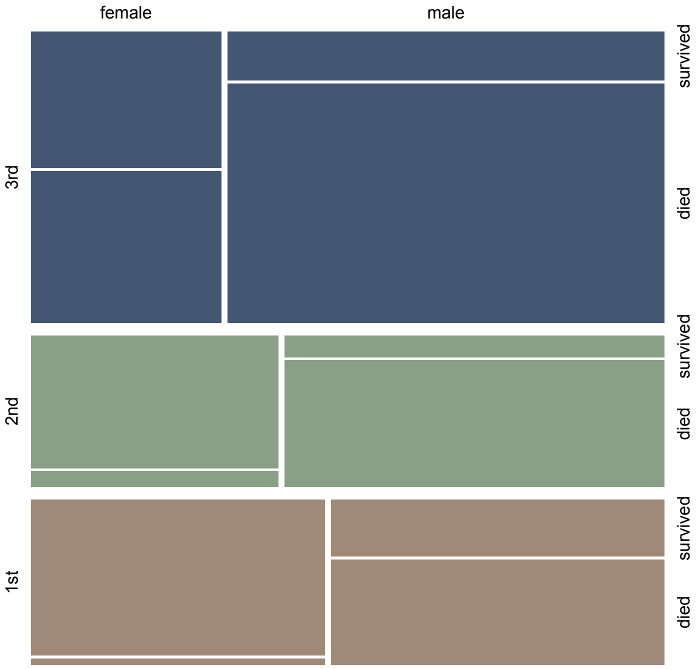
```
CrossTensorate["passenger age" == "passenger class" + "passenger sex",
  titanicData, columnNames] // MatrixForm
```

```
    | female    male
----+------------------
1st |  4926.   6195.42
2nd | 2832.42  4868.83
3rd | 3372.17  9060.83
```

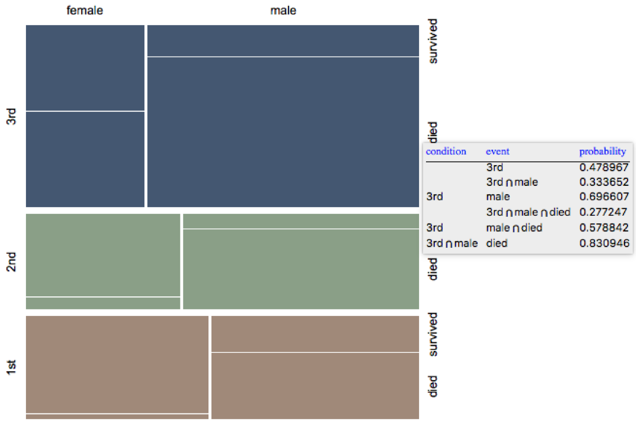
Mosaic plots

Mosaic plots can illustrate fairly well the (conditional) dependencies between the values of the categorical variables in a dataset.

```
MosaicPlot[titanicData⟦All, {1, 3, 4}⟧]
```



In contrast with R (and RStudio) in Mathematica's FrontEnd we can have tooltips showing the exact conditional values. (Hover with the mouse pointer over the rectangles in the plot above.)



condition	event	probability
	3rd	0.478967
	3rd ∩ male	0.333652
3rd	male	0.696607
	3rd ∩ male ∩ died	0.277247
3rd	male ∩ died	0.578842
3rd ∩ male	died	0.830946