Balloon Plot

Graphical tool for displaying tabular data

by Nitin Jain and Gregory R. Warnes

Introduction

Numeric data is often summarized using rectangular tables. While these tables allow presentation of all of the relevant data, they do not lend themselves to rapid discovery of important patterns. The primary difficulty is that the visual impact of numeric values is not proportional to the scale of the numbers represented.

We have developed a new graphical tool, the `balloonplot`, which augments the numeric values in tables with colored circles with area proportional to the size of the corresponding table entry. This visually highlights the prominent features of data, while preserving the details conveyed by the numeric values themselves.

In this article, we describe the balloonplot, as implemented by the `balloonplot` function in the `gplots` package, and describe the features of our implemenation.

Function description

The `balloonplot` function accepts a table (to be displayed as found) or a vector or list of vectors for x (column category), y (row category) and z (data value) of vectors from which a table will be constructed.

The `balloonplot` function plots a graphical table, where each cell displays the appropriate numeric value plus a colored circle whose size reflects the relative magnitude of the corresponding component. The *area* of each circle is proportional to the frequency of data. (The circles are scaled so that the circle for largest value fills the available space in the cell.)

As a consequence, the largest values in the table are "spotlighted" by the biggest cicles, while the smaller values are displayed with very small circles. Of course, circles can only have positive radius, so the radius of circles for cells with negative values are set to zero. (A warning is issued when this "truncation" occurs.)

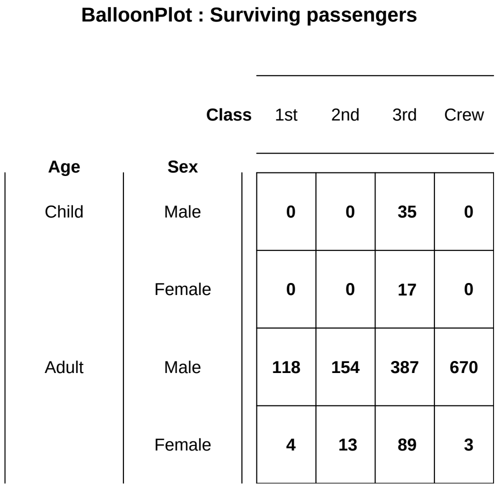
Of course, when labels are present on the table or provided to the function, the graphical table is appropriately labeled. In addition, options are provided to allow control of various visual features of the plot:

- rotation of the row and column headers
- balloon color and shape (globally or individually)
- number of displayed digits
- display of entries with zero values
- display of marginal totals
- display cumulative histograms
- x- and y-axes group sorting
- formatting of row and column labels
- as well as the traditional graphics parameters (title, background, etc.)

Example using the `Titanic` data set

For illustration purposes, we use the `Titanic` data set from the `datasets` package. `Titanic` provides survival status for passengers on the tragic maiden voyage of the ocean liner "Titanic", summarized according to economic status (class), sex, and age.

Typically, the number of surviving passengers are shown in a tabular form, such as shown in Figure **??**. (This was created by calling balloonplot with the balloon color set to match the background color.) Note that one must actively focus on the individual cell values in order to see any pattern in the data.



BalloonPlot : Surviving passengers

	Class	1st	2nd	3rd	Crew
Age	**Sex**				
Child	Male	**0**	**0**	**35**	**0**
	Female	**0**	**0**	**17**	**0**
Adult	Male	**118**	**154**	**387**	**670**
	Female	**4**	**13**	**89**	**3**

Figure 1: Tabular representation of survived population by gender and age

Now, we redraw the table with light-blue circles ('balloons') superimposed over the numerical values (figure 2). This is accomplished using the code:

```
library(gplots)
library(datasets)

data(Titanic)

dframe <- as.data.frame(Titanic)
 ## convert to 1 entry per row format

survived <- dframe[dframe$Survived=="No",]
attach(survived)

balloonplot(x=Class, y=list(Age, Sex), z=Freq,
            sort=TRUE, show.zeros=TRUE,
            cum.margins=FALSE, main=
            "BalloonPlot : Surviving passengers")
mtext("by class, gender, and age")

detach(survived)
```

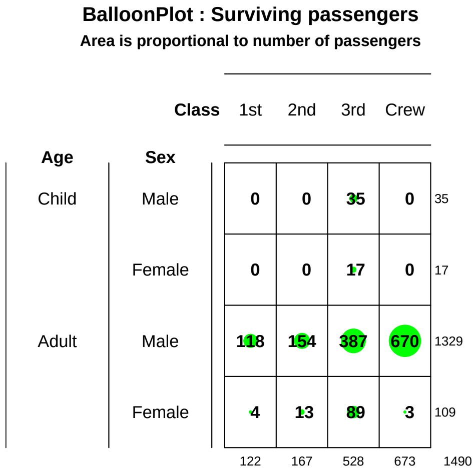


Figure 2: Balloon plot of surviving individuals by class, gender and age

With the addition of the blue “spotlights”, whose area is proportional to the magnitude of the data value, it is easy to see that only adult females and adult male crew members survived in large numbers. Note that we also added row and column marginal totals.

Of course, the number of surviving passengers is only half of the story. We can create a similar plot showing the number of passengers who did not survive. Alternatively, we can simply add surivival status as another variable to the display:

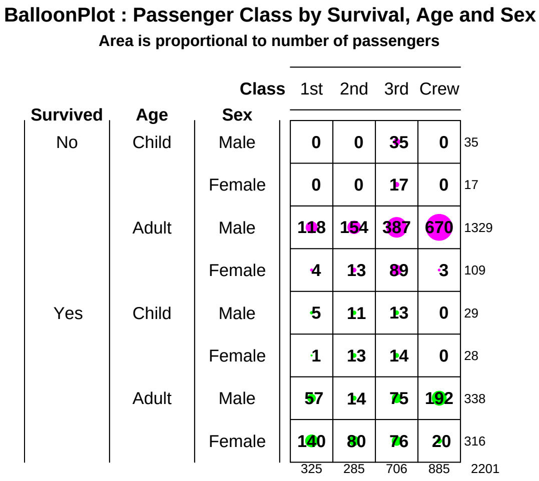


Figure 3: Balloon plot of Titanic passengers by gender, age and class. Green circles represent passengers who survived and magenta circles represent the passengers who did not survive.

Figure 3 conveys much more information than figures 1 and 2 without loss of clarity. Bigger circles of magenta color make it clear that the number of passengers who did not survive is significantly bigger than that of passengers who survived.

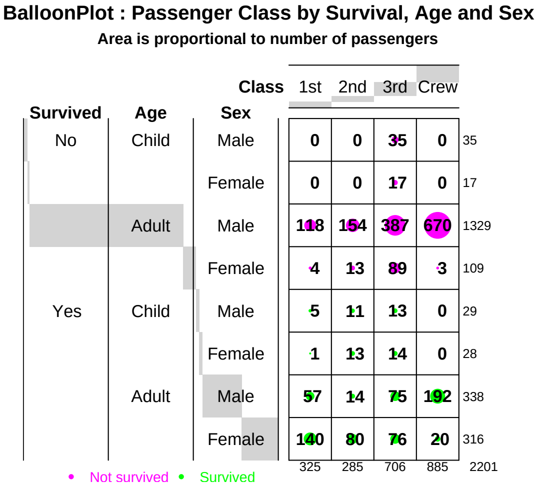
Still, to make further improvements in the display, we redraw the same data and add the cummulative sums accross each row and column, histograms across each section, and legend, as shown in figure 4. This is accomplished using the code:

```
library(gplots)
data(Titanic)
dframe <- as.data.frame(Titanic)
attach(dframe)
colors <- ifelse( Survived=="Yes", "green", "magenta")

balloonplot(x=Class, y=list(Survived, Age, Sex),
            z=Freq, sort=TRUE, dotcol=colors,
            show.zeros=TRUE, main="BalloonPlot :
            Passenger Class by Survival, Age and Sex")

title(main=list("Area is proportional to number of
            passengers", cex=0.9), line=0.5)

legend(3,0.5, legend=c("Not survived","Survived"),
       col=c("magenta","green"), pch=20, cex=.8,
       pt.cex=0.8, text.col=c("magenta", "green"),
       horiz=TRUE, xjust=0.5, bty="n")
```



		Class	1st	2nd	3rd	Crew	
Survived	Age	Sex					
No	Child	Male	0	0	35	0	35
		Female	0	0	17	0	17
	Adult	Male	118	154	387	670	1329
		Female	4	13	89	3	109
Yes	Child	Male	5	11	13	0	29
		Female	1	13	14	0	28
	Adult	Male	57	14	75	192	338
		Female	140	80	76	20	316
			325	285	706	885	2201

Figure 4: Balloon plot of all the passengers of Titanic, stratified by survival, age, sex and class

It is now easy to see that adult females and adult male crew members survived in large numbers. Figure 4 displays many features of the Titanic passengers. Passengers who survived are shown in green circles and passeengers who did not survive are shown in magenta circles. Sums across rows and columns are represented in the numbers at the right and bottom respectively. The shaded rectangular blocks show the density of data. From the figure, it can be clearly seen that the largest number of individuals who did not survive are adult male crew (670). 1st class adult females were among the group of people who survived the most. Since there were no child-crews and all the children survived, some fileds are left empty in the figure.

It turns out that there is a well known reason for this difference. Passengers in 1st and 2nd class, as well as crew members, had an easier time reaching the lifeboats. Since there were too few lifeboats for the number of passengers and crew, most women and children among the first and second class passengers as well as female crew found space in a lifeboat, while many of the later arriving 3rd class women and children were too late: the lifeboats had already been filled and had moved away from the quickly sinking ship.

Conclusion

Using the well worn Titanic data, we have shown how balloonplots help to convey important aspects of tabular data, without obscuring the exact numeric values. We hope that this new approach to visualizing tabular data will assist other statiticians in more effectively understanding and presenting tabular data.

We wish to thank *Ramon Alonso-Allende* `allende@cnb.uam.es` for the discussion on R-help which lead to the development of `balloonplot`. Ramon also added the code to display the row- and column sums.

Gregory R. Warnes, Pfizer Inc., USA
`gregory.r.warnes@pfizer.com`
Nitin Jain, Pfizer Inc., USA
`nitin.jain@pfizer.com`