Mosaic Plots in the ggplot2 Framework ggmosaic

by Haley Jeppson, Heike Hofmann

Abstract The options of graphical methods for categorical variables are not well developed in comparison to what is available for numeric variables. One method of visualizing multidimensional data is through a mosaic plot. Mosaic plots can be an easy and powerful option for identifying relationships between multiple variables. However, while mosaic plots have been implemented in a variety of packages, the ordinary grammar of graphics does not support mosaic plots. With the R package ggmosaic, a custom ggplot2 geom designed for mosaic plots is implemented. Equipped with the functionality and flexibility of ggplot2, ggmosaic creates plots that can be converted into interactive plotly graphs. This paper provides an overview of the implementation and examples that highlight the versatility and ease of use of ggmosaic while demonstrating the practicality of mosaic plots.

Introduction

Is there a social contract we should all be following when flying? Who gets to use that extra arm rest? Is it okay to recline your seat or to bring your baby or child on board? Are people upset if the child is unruly and, like Donald Trump, wish to "get that baby out of here?" (Killough, 2016) How often are you allowed to leave your seat before being perceived as rude? Does the occupant of the window seat reserve all rights to the window shade? And is it okay to switch seats for family, friends, or to an unsold seat?

To learn a little about people's opinions on what is considered to be rude behavior while on an airplane, FiveThirtyEight ran a SurveyMonkey Audience poll for two days in August of 2014. (Hickey, 2014) The survey had 1,040 respondents (874 of whom had flown) aged 18-60+ from across the county and asked twenty-six questions that ranged from background information regarding the respondent to feelings regarding potentially aggravating behavior one might encounter on an airplane. The data set is available from FiveThirtyEight's data git hub repository.

To explore the data a bit more, I used the `ggmosaic` package. The original analysis done by FiveThirtyEight focused primarily on the perceived rudeness of a behavior, one behavior at a time, or comparing the response "Very Rude" across all behaviors. By using a mosaic plot, multiple categorical counts can be viewed simultaneously which will provide a clearer view of the underlying data and allow for more comparisons to be made. We can perhaps gain more insight into how the seat recliner, the chatterbox, or the unruly child is perceived and how those perceptions may be related.

The `ggmosaic` package

The `ggmosaic` package was designed to create visualizations of categorical data, and has the capability to produce bar charts, stacked bar charts, mosaic plots, and double decker plots. The main focus of this paper, however, will be on mosaic plots. A mosaic plot is a convenient graphical summary of the conditional distributions in a contingency table, and in a mosaic plot, the area of each graphical element is proportional to the underlying probability of that category. This allows us to easily visualize how the joint distribution is composed of the product of the conditional and marginal distributions – which, in turn, allows us to see any association that may be occurring between the variables. Because the plot is constructed hierarchically, the ordering of the variables is very important. There are many features that can be customized in `ggmosaic`, including the type of partitioning, the ordering of variables, conditioning or faceting, and the spacing between the categories.

While mosaic plots have been implemented in a variety of packages in R (R Core Team, 2016), the ordinary grammar of graphics does not support mosaic plots. However, with version 2.0.0 of `ggplot2` (Wickham, 2009), a way for other R packages to implement custom geoms was introduced. With the R package ggmosaic, a custom ggplot2 geom designed for mosaic plots is implemented. The `ggmosaic` package can be installed from `https://github.com/haleyjeppson/ggmosaic`.

`ggmosaic` was created primarily using `ggproto` and the `productplots` package which was created by Wickham and Hofmann (2016); **?**. They refer to their framework as product plots, alluding to the computation of area as a product of height and width, and the statistical concept of generating a joint distribution from the product of conditional and marginal distributions.

To begin, `ggmosaic` began as a geom extension of the `rect` geom with the data handling provided in the `productplots` package which calculates `xmin`, `xmax`, `ymin`, and `ymax` for the `rect` geom to plot.

Having a geom designed for mosaic plots does more than simply allow us to utilize the ggplot2

customization options such as faceting and layering, it allows for a `ggplotly()` hook so we can create interactive mosaic plots. Although the `ggplotly()` function translates most of the geoms bundled with the `ggplot2` package, it has no way of knowing about the rendering rules for custom geoms. The `plotly` package does, however, contain the infrastructure to provide translations of custom geoms to plotly. In `ggplot2`, many geoms are special cases of other geoms. For example,`geom_line()` is equivalent to `geom_path()` once the data is sorted by the `x` variable. Sievert (2016) Because GeomMosaic can be reduced to the lower-level geom GeomRect, we were able to write a method for the `to_basic()` generic function in `plotly`.

`ggmosaic` does not come without its own set of limitations and the main hurdle `ggmosaic` faces is that `ggplot2` is not capable of handling a variable number of variables. The current solution is to read in the variables `x1` and `x2` as `x = product(x1, x2)`. The `product` function creates a list of the variables specified which allows for it to pass `check_aesthetics`, and then splits the variables back into a dataframe for the calculations.

The aesthetics set up the formula that determines the how the joint distribution will be broken down. The following aesthetics can be set:

- weight: select a weighting variable
- `x`: select variables to add to formula
  - declared as `x = product(x1, x2, ...)`
- `fill`: select a variable to be filled
  - if the variable is not also called in `x`, it will be added to the formula in the first position
- `alpha`: select a variable to receive a transparency-scale
  - if the variable is not also called in `x`, it will be added to the formula in the first position
- `conds` : select a variable to condition on

These values are then sent through `productplots` functions to create the formula for the desired distribution: `weight` $\sim$ `alpha` $+$ `fill` $+$ `x` $|$ `conds`

The other parameters that can be set include the `offset` and the `divider`. When there is a variable with many categories, it may be of interest to decrease the size of the spacing between the spines. This can be achieved by declaring `offset =`. The default setting is `offset = 0.01`. There are two main ways to partition the area - into bars or into spines. When the area is partitioned into bars, the height is proportional to value and the width equally divides the space. Bars can be arranged horizontally ("hbar") or vertically ("vbar"). Alternatively, the space can be partitioned into spines, where the width is proportional to value, height occupies full range. Spines are space filling and can be arranged horizontally ("hspine") or vertically ("vspine"). In `ggmosaic`, the type of partitioning desired can be specified by setting `divider = " "`. The default divider for one variable is "hspine". When more than one variable is to be considered, a type of partition needs to be selected for each variable. By selecting `divider = mosaic()`, the default, or `divider = ddecker()`, the correct number of partitions will be selected. For example, if three variables were to be plotted, the default, `divider = mosaic()`, would partition the plot with spines in alternating directions, beginning with a horizontal spine, i.e. `divider = c("hspine", "vspine", "hspine")`. It is also an option to manually select the type of partition that will be used for each variable, i.e. `divider = c("hbar", "vspine", "hspine")`. It should be noted that the first partition in the vector will be the last partition made in the plot. As mentioned above, when no divider is declared, the default `divider = mosaic()` will begin with a horizontal spine and alternate directions with each subsequent variable.

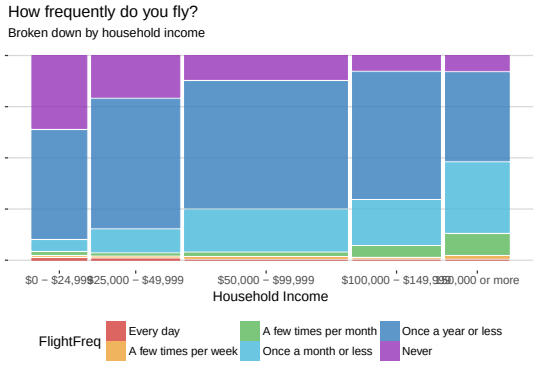
Using mosaic plots to visualize this data will allow for the hierarchical structure of the counts and proportions, which is important for understanding the multivariate discrete distributions, to be seen.

Visualizing the Data

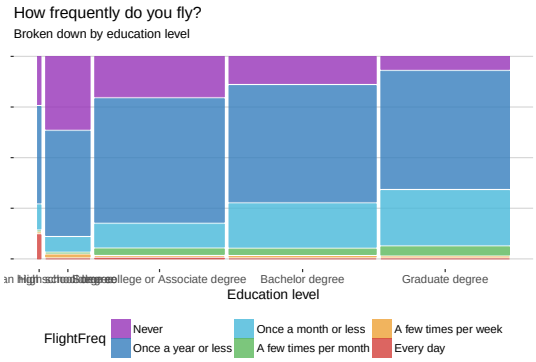
I began by taking a look at who is flying and how frequently they fly. In addition to answering questions about common aggravating behaviors, the respondents answered questions regarding their demographics. Some of the demographic types of questions asked included age, gender, household income bracket, education level and region.

Figure ~**??** was created by first dividing the space into horizontal spines each representing the proportion of respondents within that household income bracket. We can view these horizontal spines in the final product and they will answer questions such as, "what proportion of the respondents have a household income of under $100,000?" The next step was to split each horizontal spine into vertical spines, which were subsequently filled with different colors, representing how frequently one flew.

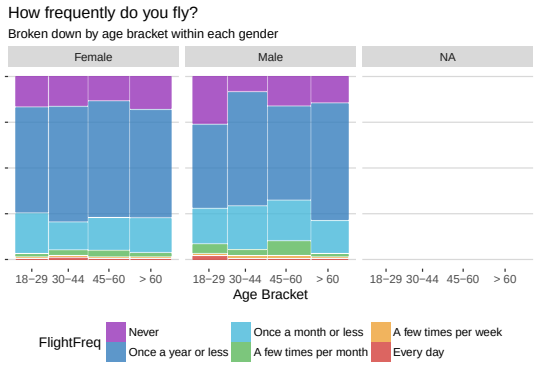
This plot can be used to answer questions such as "what proportion of those that have a household income of under $24,999 flies a few times per month?"



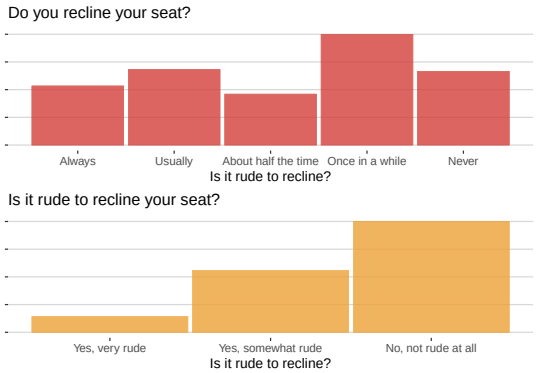
From Figure ~**??** we can see that the largest majority of the survey participants fall within the household income bracket of $50,000-99,000, and that as household income increases, the more likely one is to frequently fly. Even more interesting, the few participants that responded that they flew every day were of the two lower income brackets or chose to not respond to the question.



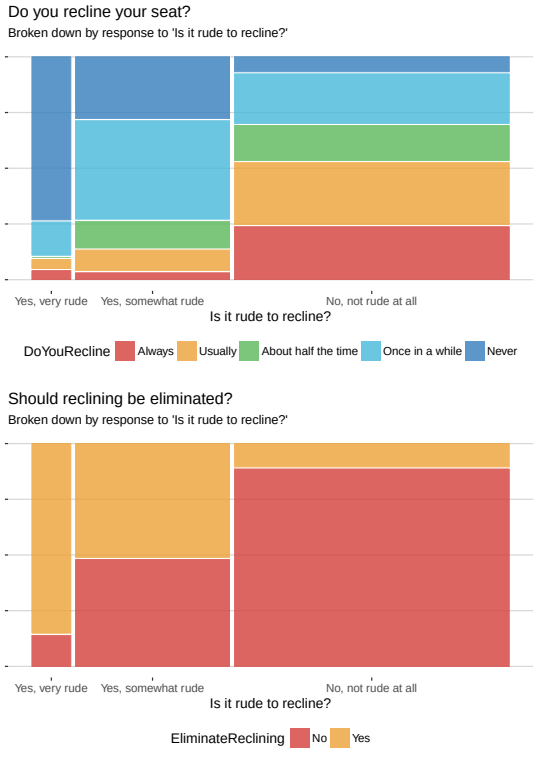
Continuing the investigation into who is flying and how frequently they fly, Figure ~**??** presents how frequently the respondents flew by education level. From the mosaic plot, we can see that the majority of respondents had some type of college education and there is a trend people flying more frequently as education level increases.



To add to this, Figure ~**??** shows that gender and age don't play too large of a role in how often one flies, though interestingly the largest group of those that never fly was made up of males in the age bracket 18-29. Figure ~**??** exemplifies how having mosaic plots implemented as a geom allows for the ggplot2 customization features to easily be accessed. Here, rather than partitioning into spines for the gender categories, I have faceted on gender. Statistically, this relates to conditioning on the variable on gender. The distribution that is being displayed by Figure ~**??** is f(FlightFreq, Age | Gender)



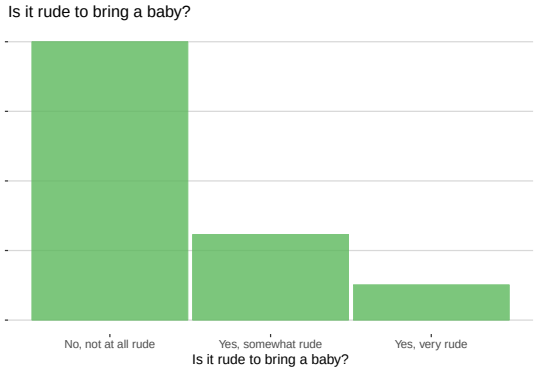
Next I dove into the perceived level of rudeness of some of the more common behaviors. First up, how do the participants view "The seat recliner" and are the participants seat recliners themselves? As airlines look to shrink costs, the airline seats appear to be shrinking as well. In the past thirty years, the average cheap seat has shrank from 18 inches wide to about 16 and a half inches wide. Halsey III (2016) Has this smaller space brought strong opinions on a flight passenger's right to recline their seat? We can look at how often the participants reclined their seat and also at how rude participants believe the behavior to be. However, perhaps a more interesting question is "How often does one recline given they feel the behavior is very rude?" Figure ~**??** is a breakdown of participant responses to the two questions "Do you recline?" and "Is it rude to recline?" that can allow for such types of questions to be answered. While some of the results were as expected - the more rude someone considers the act reclining, the less likely they were to recline their own chair- there were three respondents who felt it was very rude to recline your seat on an airplane, yet always reclined their own seat. The plot also lets us know that the majority of the respondents do not find it at all rude to recline your seat and are fairly evenly divided on how often they themselves recline their own seat.



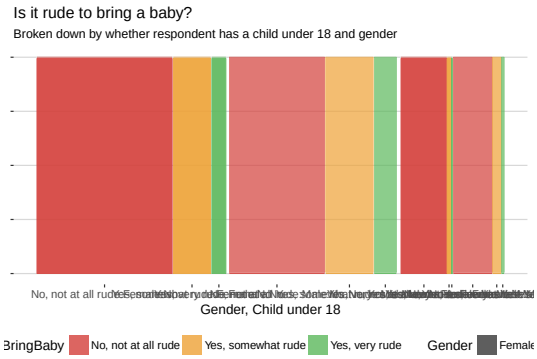
Not a particularly surprising revelation, but Figure ~**??** displays how the small propotion of participants that view the act of reclining your seat on an airplane as rude would also like to eliminate the option of reclining. In contrast the larger proportion that do not view the act as rude do not see a reason to eliminate the option.

Next, how did the participants view babies and unruly children on flights? Are there certain demographics that are more likely to be offended by their presence? Are parents less likely or more likely to be offended by other children? The firstplot adressing this topic is Figure ~**??**. Here we can get an idea of how the particpiants viewed the act of bringing a baby on an aircraft. Fortunately, we see that most were not bothered, but there is a slight proportion that finds it to be very rude.

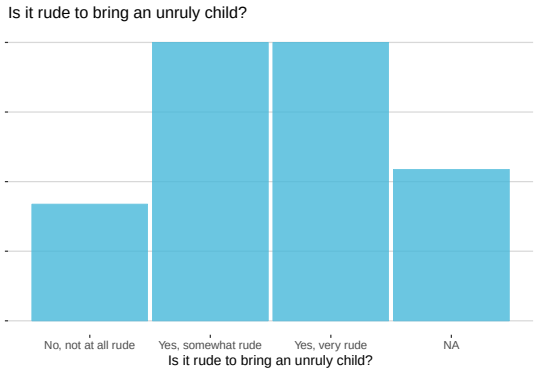
To see where those respondents might be coming from, the next plot looks at how the participants judged the level of rudeness of bringing a baby on board broken down by whether the participant had a child under the age of 18 and then by gender. **??** is an example of one of the other types of plots `ggmosaic` is capable of creating, a double decker plot. A modification of a mosaic plot, a double decker plot, is composed of n-1 hspines and ends with a vspine rather than alternating hspines and vspines. In ~**??**, the plot is first split horizontally by response to "Do you have a child under the age of 18?". From there each hspine is split into gender resulting in a plot where each combination of gender and child under the age of 18 is represented by a vertical bar where the width represents the proportion of that category. Lastly, each vertical bar is split vertically by the responses to "Is it rude to bring a baby on a plane?" The finished product is a plot that allows for the three variables to be presented concisely and displays how the proportions differ.



In ~**??**, we see that those without a child under the age of 18 make up the largest proportion of the respondents and they are more likely to consider bringing a baby on board as rude. Additionally, men or more likely than women to consider it rude to bring a baby on a flight.

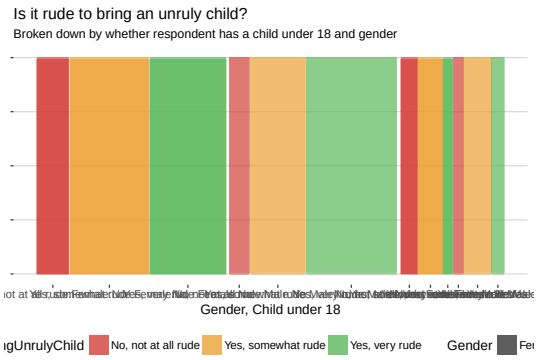


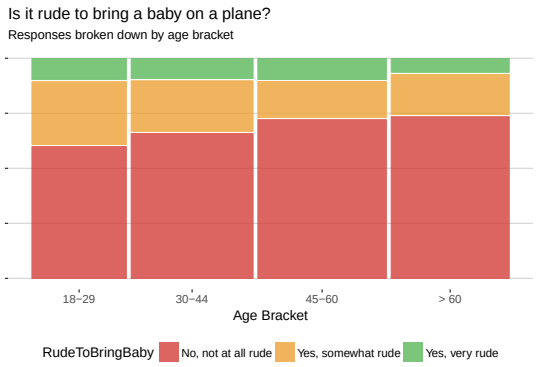
The next figure, Figure ~**??**, looks into how the participants viewed sharing the plane ride with an unruly child. In Figure ~**??** we can see that the participants were likely to be bothered by an unruly child. Perhaps this has to do with the wording of the question in the SurveyMonkey, but I continued on with the analyis anyway.

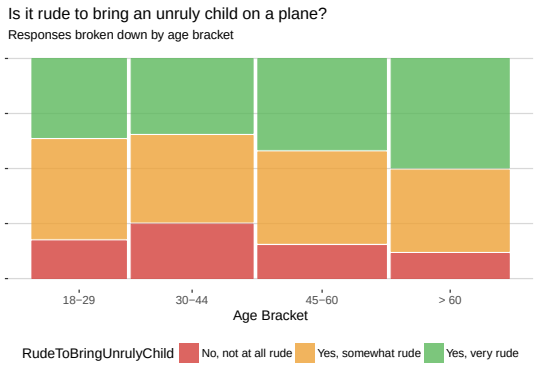


To see where those opinions stemmed from, or to see if those opinions varied by gender and whether or not the participant has a child under the age of 18, Figure ~**??** was created. In the double

decker plot, it is clear that females are more forgiving of unruly children than males and that parents of children under the age of 18 are also more forgiving. It does, however, seem fairly clear that most find the presence of an unruly child as rude.







To continue the investigation of how often faircraft passengers find an unruly child or baby on board as rude and how this opinion varies person to person, Figures ~**??** and ~**??** were created. An interesting revelation brought about by these two plots is that while those older seem to be less likely to be bothered by a baby being on a flight, they are more likely to be upset by an unruly child.

Conclusion

After exploring the data a bit more, it is clear that there does not seem to be a general consensus on what passengers perceive to be acceptable behavior on an airplane. The use of mosaic plots aided in the exploration of how certain opinions or behaviors are related.

Summary

This file is only a basic article template. For full details of *The R Journal* style and information on how to prepare your article for submission, see the Instructions for Authors.

Bibliography

A. Halsey III. Trying to squeeze into that airline seat? congress is feeling the pinch, too., Mar 2016. URL https://www.washingtonpost.com/local/trafficandcommuting/trying-to-squeeze-into-that-airline-seat-congress-is-feeling-the-pinch-too/2016/03/01/a0b51c72-df22-11e5-846c-10191d1fc4ec_story.html?utm_term=.73ebc6d3a2dc. [p4]

W. Hickey. 41 percent of fliers think you're rude if you recline your seat, Sep 2014. URL http://fivethirtyeight.com/datalab/airplane-etiquette-recline-seat/. [p1]

A. Killough. Trump: 'you can get the baby out of here', Aug 2016. URL http://www.cnn.com/2016/08/02/politics/donald-trump-ashburn-virginia-crying-baby/. [p1]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. [p1]

C. Sievert. plotly for r, 2016. URL https://cpsievert.github.io/plotly_book/. [p2]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org. [p1]

H. Wickham and H. Hofmann. *productplots: Product Plots for R*, 2016. URL https://CRAN.R-project.org/package=productplots. R package version 0.1.1. [p1]

Haley Jeppson
Iowa State University
line 1
line 2
author1@work

Heike Hofmann
Iowa State University
line 1
line 2
author2@work