Project Proposal

Haley Jeppson

March 16, 2016

Project info

Project title:

Mosaicplots in the ggplot2 framework: ggmosaic

Project short title: ggmosaic

URL of project idea page: ggmosaic

Bio of Student Provide a brief (text) biography, and why you think your background qualifies you for this project.

CONTACT INFORMATION

Student name: Haley Jeppson

Melange Link_id:

Student postal address: 1124 North 2nd Street, Ames IA 50010

Telephone(s): (319) 327-0459

Email(s): hjeppson@iastate.edu

Other communications channels: Skype/Google+, etc. :

Student affiliation | |

Institution: Iowa State Univeristy

Program: Statistics

Stage of completion: 1.5 years completed

Contact to verify:

Schedule Conflicts: I am not currently aware of anything that would interfere with me treating the ggmosaic project as a full time job this summer.

MENTORS

Mentor names: Heike Hoffman and Diane Cook

Mentor emails: hofmann@iastate.edu

Mentor link_ids: hofmann@iastate.edu

Have you been in touch with the mentors? When and how?

I have been in touch with Heike via email several times.

CODING PLAN & METHODS

Describe in detail your plan for completing the work. What functions will be written, how and when will you do design, how will you verify the results of your coding? The sub-section headings below are examples only. Each project is different, please make your application appropriate to the work you propose.

Outcomes:

1. R package for generalized version of mosaic plots implemented as a `geom` for the `ggplot2` package based on `ggproto`.
2. A set of examples documenting the use and flexibility of `geom_mosaic`.
3. A shiny app highlighting the mosaicplot functionality interactively, to allow users to specify parameters and see the impact immediately to allow them to familiarize to the more abstract concepts of mosaicplots.

Describe perceived obstacles and challenges, and how you plan to overcome them.

TIMELINE

Date	Event
23 May	Begin coding
20 June	Midterm evaluations
15 Aug	Final week. Tidy code, write tests, improve documentation and submit code sample.

Provide a detailed timeline of how you plan to spend your summer. Don't leave testing and documentation for last, as that's almost a guarantee of a failed project.

What is your contingency plan for things not going to schedule?

MANAGEMENT OF CODING PROJECT

How do you propose to ensure code is submitted / tested?

How often do you plan to commit? What changes in commit behavior would indicate a problem?

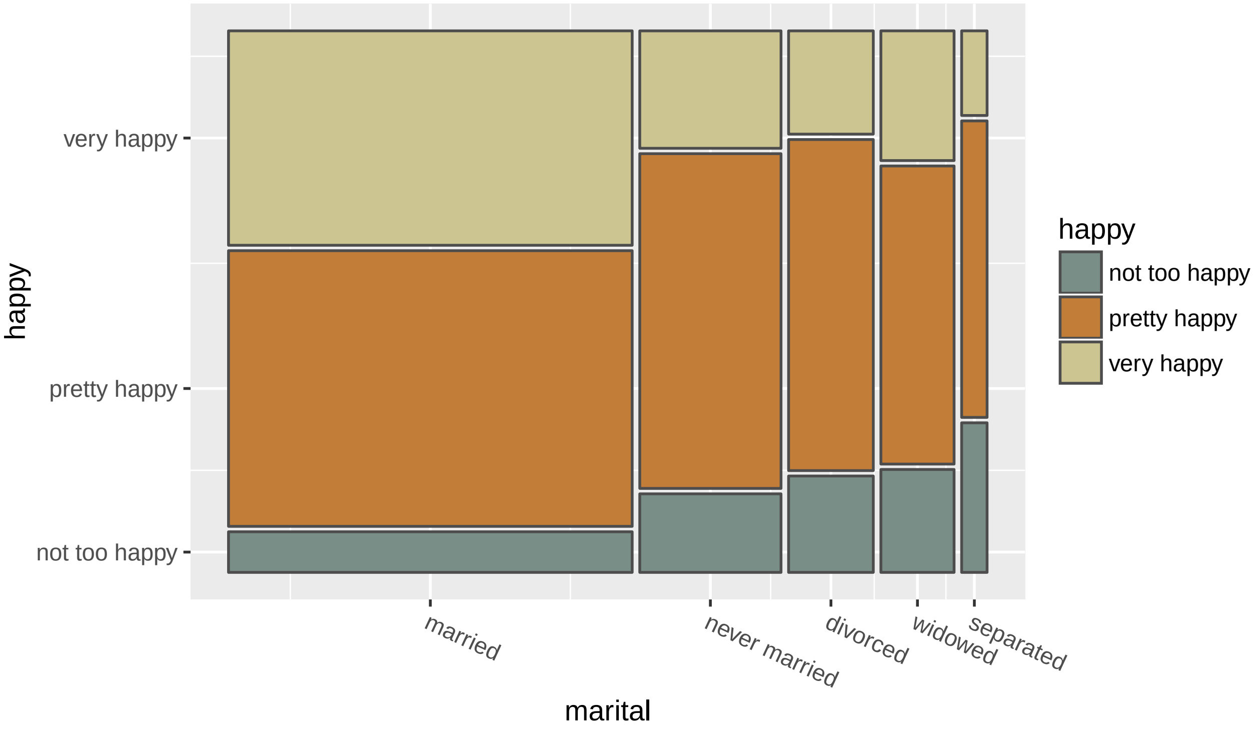
TEST

Describe the qualification test that you have submitted to you project mentors. If feasible, include code, details, output, and example of similar coding problems that you have solved.

1. Install the productplots package from github (you might have to install the devtools package first). Run one of the examples, put the chart in a knitr/Rmarkdown document and write a paragraph to explain the chart.

A mosaic plot is a convenient graphical summary of the conditional distributions in the contingency table. The area of the graphical element is propotional to the underlying probability, so we are easily able to visualize how the joint distribuiton is composed of the product of the conditional and marginal distributions which allows us to see any association that may be occuring between the variables. Because the plot is built hierarchically, the ordering of the variables is very important.

The `productplots` package is used below to produce an example of mosaic plot for two categorical variables. In addition, below the plot is an interpretation of the plot. The data set used in the `happy` data set and the 2 variables to be considered in this example are happy (with 3 levels not too happy, pretty happy, very happy) and marital (with 5 levels married, never married, divorced, widowed, separated). For this example, all `NA`s have been removed.



In this first example of a mosaic plot we are viewing the joint distribuiton of the variables `happy` and `marital` as the product of the marginal distribtion of `marital` and the conditional distribution of `happy` conditioned on the variable `marital`, i.e. $f(happy, marital) = f(happy|marital)f(marital)$. We can consider the rows as the categories of the response variable and the columns are the categories of the explanatory variable. When the segments in the mosaic plot do not line up, there is an indication of an assocation between the variables. For example, from the plot, it appears the marital group `separated` is the least likely to have responded with a `very happy` level of happiness.

2. Write a shiny app that shows a mosaicplot (usingp`prodplot`) of a few variables and allows to interactively change at least one aspect of the mosaic.

3. Based on Hadley Wickham's introduction to extending `ggplot2` write a function that implements a geom of your choice. Document the function using Roxygen, and include it into an R package.

Anything Else