Williams College Initials

Ben Czekanski

January 9, 2017

Introduction

The package **wcgrads** examines the distribution of initials of Williams College Graduates. It uses information published by the Registrar of the College in PDF and analyzes it to find which combinations of initials are most popular, and visualizes this information.

Data

The data for this package comes from Course Catalogs posted online by the Williams College Registrar. These files are first downloaded and then converted to pdf using the pdftotext executable. They are then cut down to just the list of graduating seniors and further broken into first and last names. Once in first name last name format, the multiple graduating classes are combined into one data frame of all students that graduated from 2001-2016.

Functions

There are four functions within the **wcgrads** package:

download_pdfs

download_pdfs() simply downloads the Course Catalogs in pdf form

get_data

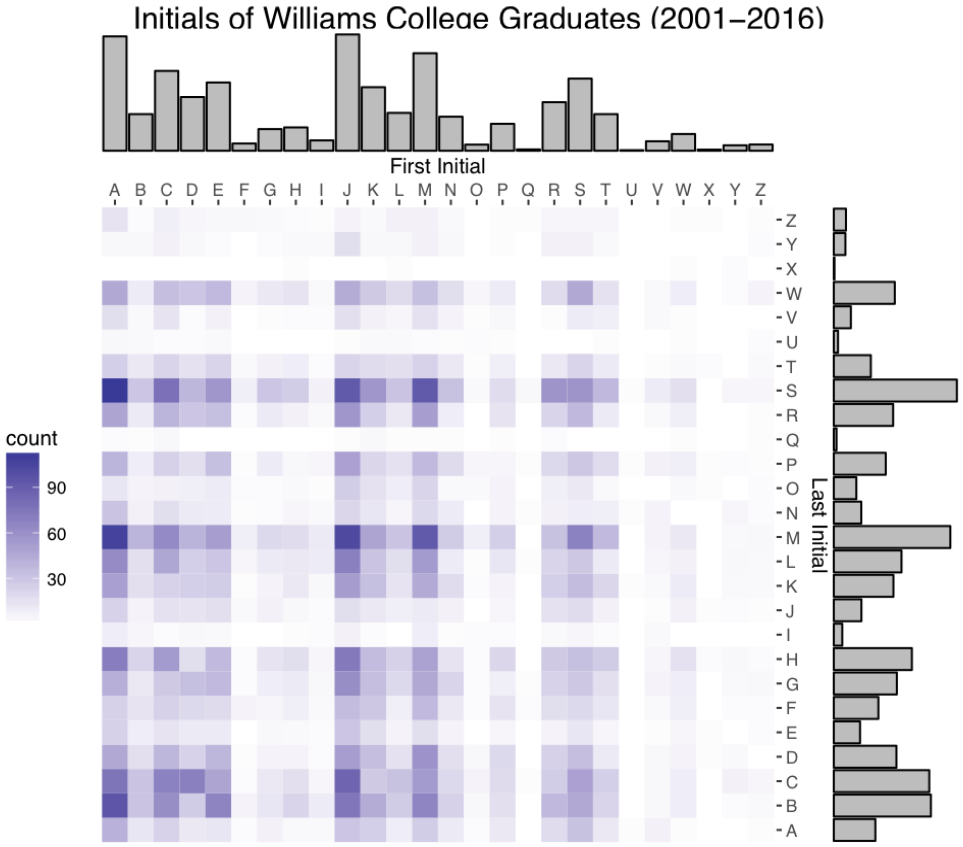
get_data(file) converts these pdfs to txt to a data frame and then cuts out the parts that aren't names. It also breaks the strings of names and honors into just first and last initials.

combine_data

combine_data() uses all *get_data* to make all of the pdfs into txt documents and then one large, cleaned data frame of all the initials and graduation years

visual

visual(x) creates a visualization of initials showing both the joint and marginal distributions of first and last initials for all graduates from 2001-2016, as shown below.



Conclusion

Looking at the plot generated, we see several main trends. First, the distribution of last initials appears to be much more even than the distribution of first initials. This suggests that there may be some letters that are more popular than others. When given the choice of first initial, letters A-E, J-N, and R-T are given more than other letters. This could also simply be because there are more options for first names that start with these initials. The difference between first and last initals appears to be greatest in F-H and W, which all occur as last initials much more than they do as first initials. Another interesting aspect of the graphic is observing which first initials are more or less popular given a last initial. For example, L and K occur as last initials at a similar rate. However, the pairing CL occurs much more frequently than the pairing CK, possibly due to avoidance of alliteration. Inversely, while E and K occur about as frequently as first names, EC occurs much more than KC. The plot offers the opportunity to make many other observations, and I invite the reader to explore them. Further advancements on this subject could include showing change over time, further breakdown by name, or breaking this down by sound rather than letter to further examine how sound affects name choice. Another possible use of the data is to compare Williams graduates to other NESCAC schools or the population as a whole.