Williams College Graduates

Ben Czekanski, Middlebury '18

January 23, 2017

Introduction

The package **wcgrads** examines data on graduating seniors published in the Williams College Course Catalog. The Catalog is published as a PDF document, so this package converts catalog from .pdf to .txt and then cleans it into a tidy format for analysis. **wcgrads** includes analyses of different elements of this data. These analyses cover graduates initials' and name lengths, as well as the distribution of latin honors over time. While the included analyses are interesting and informative, **wcgrads** allows the user to perform their own examinations by including the cleaned version of these catalogs.

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. After being converted to .txt, the catalog is cut down to just the list of graduating seniors and then cleaned into a single line of data per graduate.Once in this single line format, all graduating classes are combined into a single data frame of all students that graduated from 2001-2016. This data frame is called "allyrs". "allyrs" provides all of the information from the catalogs, and adds gender. It is important to note that this is gender predicted by the R pacakge **gender** using the method "ssa" and is almost certainly not completely correct. However, providing predicted gender gives package users additional, interesting information to analyze.

Analyses

The first question we can answer using this data is "What is the longest name?". This question can be answered using the namelength() function.

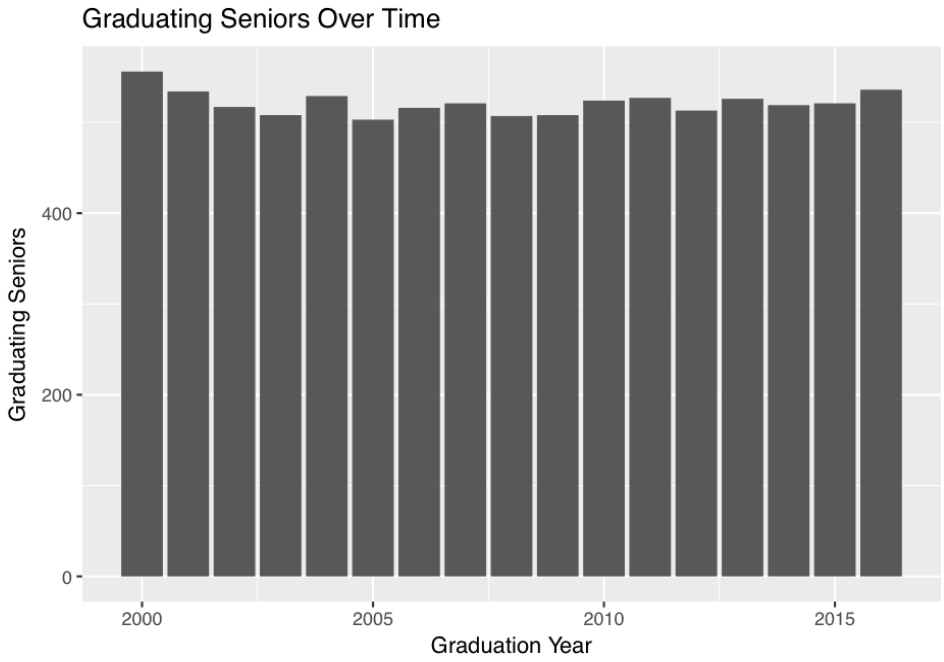
```
library(wcgrads)
names <- namelength()
head(names, 10)
```

```
## # A tibble: 10 × 3
##                                    entirename namelength grad.year
##                                         <chr>      <int>     <dbl>
## 1  Susannah Sarah Mahinaokapolanialoha Fyrberg         40      2000
## 2    William Oliver Sweetwater Parker Bobseine         37      2010
## 3      JenniferAnne Lorraine McLellan Morrison         36      2012
## 4       Oloruntosin Adepeju Ifedadepo Adeyanju         35      2008
## 5       Christophe Alexander Dorsey-Guillaumin         35      2010
## 6      Pierre-Alexandre Charles Meloty-Kapella         35      2010
```

```
## 7       Bhuvaneswari Ettinamane Narendra Reddy         35      2011
## 8     Chloë Iambe Naomi Illyria Feldman Emison         35      2012
## 9        Katharine Elizabeth Huntress Peterson         34      2008
## 10       Quinn Christianna Sanborn Brueggemann         34      2010
```

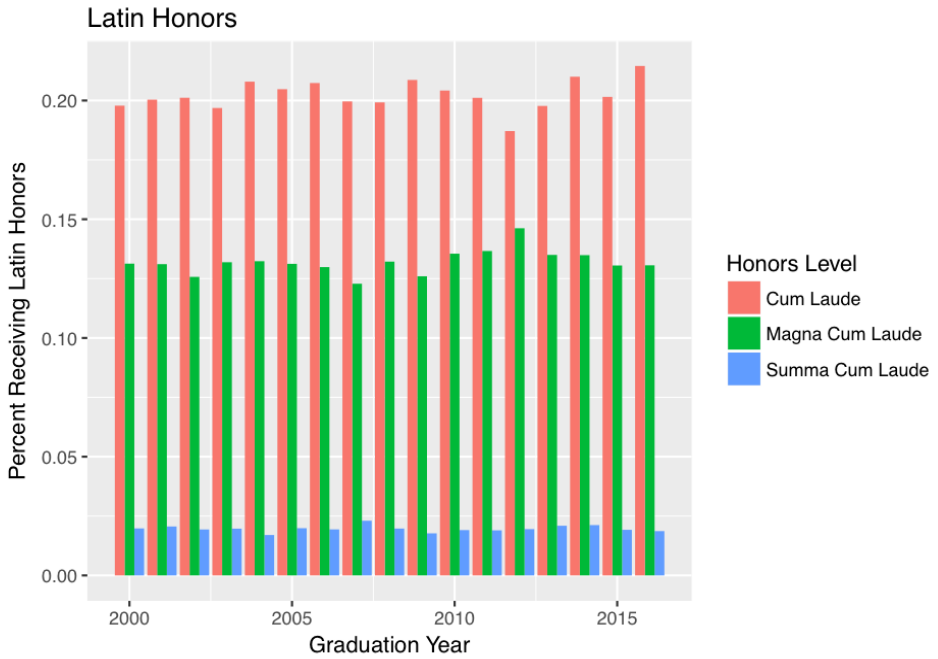
Next we can answer the question "How has the number of graduating seniors changed over time?". The gradcount() function can answer this question easily by provinding a graph showing the change over time.

```
library(wcgrads)
gradcount()
```



Yet another question that the allyrs data allows us to answer is "How has the distribution of latin honors changed over time?". **wcgrads** has a function for that too. latin_honors() shows the percentage of students that received each level of latin honors by year. There is no general trend, but there is a decent amount of variation. Perhaps the variation can be explained by equal GPA's and the methodology used in the case of such ties.

```
library(wcgrads)
latin_honors()
```

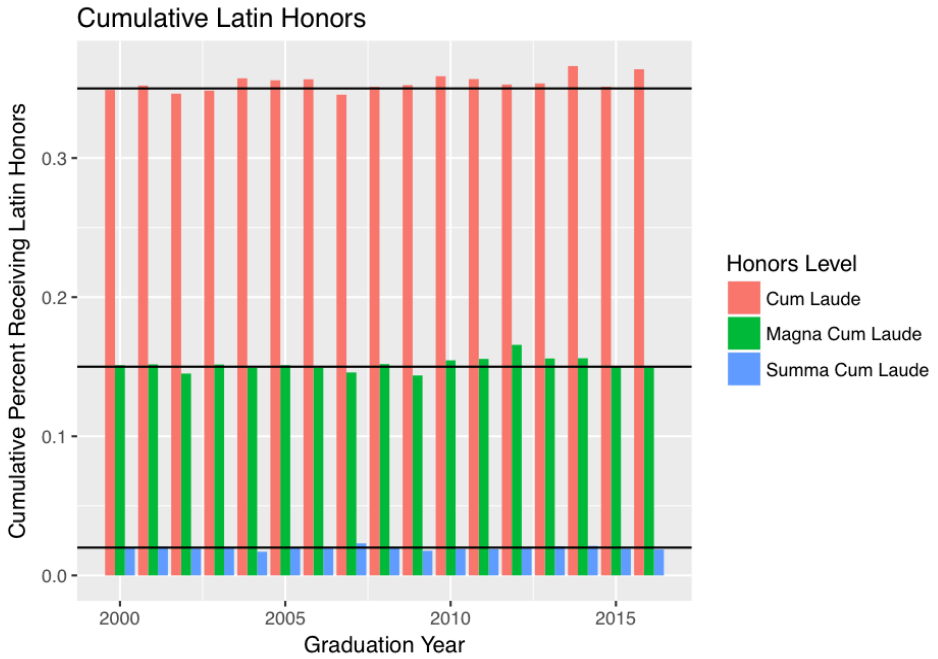


Delving further into latin honors, we find that the following in the Course Catalogs:

"The Faculty will recommend to the Trustees that the degree of Bachelor of Arts with distinction be conferred upon those members of the graduating class who have passed all Winter Study Projects and obtained a four year average in the top: 35% of the graduating class — Bachelor of Arts cum laude or higher 15% of the graduating class — Bachelor of Arts magna cum laude or higher 2% of the graduating class — Bachelor of Arts summa cum laude"

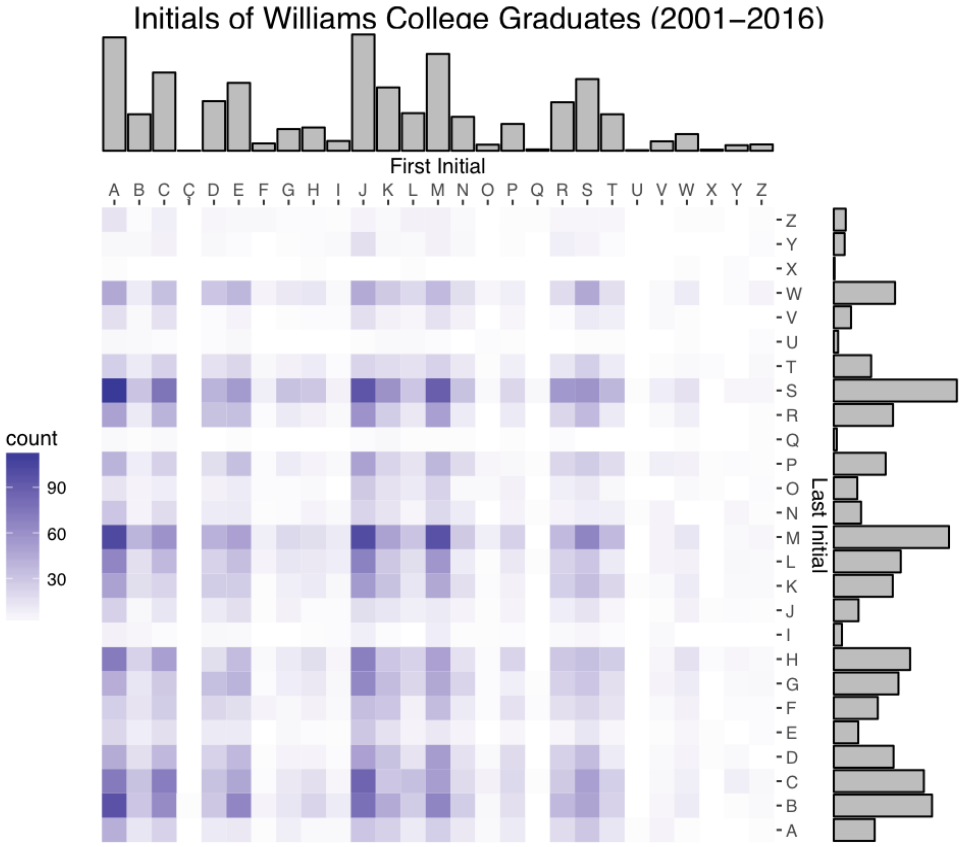
The total_latin_honors() function shows the percentage distributions of cumulative latin honors, and compares them to the guidelines(in black) found in the Course Catalogs. Again there is no real trend, but there is inconsistency. For all of the levels of honors there are years where the distribution exceeds the guidelines, and years where it fails to reach the guideline. This is confusing because one would expect that if a tie did not allow the exact percentage to earn a certain level, there would be consistent rounding up or rounding down, such that the cumulative percentages are consistently above or below that recommendation. This variation causes concern because these honors are distinctions that appear on resumes, and that people work hard to achieve. It therefore seems unfair that latin honors are given liberally in some years and conservatively in others, without a clear methodology.

```
library(wcgrads)
total_latin_honors()
```



A final, somewhat more trivial analysis of the data involves the distribution of initials. The initials() function performs this analysis and shows both the joint and marginal distribution of initials. 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 often 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.

```
library(wcgrads)
initials()
```



Conclusion

The **wcgrads** package uses the "allyrs" data to answer a number of questions, both important and trivial about Williams College graduates. The limited analyses included in the package also raise interesting questions ranging from how names are chosen to how the College handles ties in GPA when awarding latin honors. More important than any question it asks or answers, **wcgrads** makes information buried in the Williams College Course Catalog easily available to anyone who wants to analyze it.