Williams Graduates

Yuxin Wu ’19

2017-2-10

Abstract

The `Williams Graduates` package uses the data published online by the Williams College Registrar to compute statistics about the gender distribution of the graduated classes. This is done for sixteen years from 2000 to 2015 by scanning data directly from the text versions of the published documents. Data of 8354 undergraduate students were collected and analyzed to produce a series of statistical summaries in terms of timeplots to understand the change in gender distribution of the students as time progresses.

Introduction

The gender distribution of the students is an important factor of the college. The package `williamsgraduates` implements string manipulation on the data about each student in the graduated class from yearly catalogs and organizes the information of each student to an accessible dataframe.

The dataframe is further manipulated, and the data is presented with different timeplots of the gender distribution. These summary functions offer some preliminary analysis on the data, however the data can be used to perform more complex analyses, potentially with information combined from external data sources.

Data

The information used to construct the data in this package was taken from the website of the Office of The Registrar of Williams College in the form of pdf documents. The documents for each year from 2000 to 2015 were collected and manually converted into text files and pared down to only the pages that contained information on graduated students.

The information on students of the graduated class is first read line by line into the function.

```
> readLines(system.file("extdata", str_c(2000,".txt"), package = "williamsgraduates"), warn
= FALSE)
```

```
##  [1] "Bachelor of Arts, Summa Cum Laude"
##  [2] "*DoHyun Tony Chung, with honors in Political"
##  [3] "Economy"
##  [4] "*Rebecca Tamar Cover, with highest honors in"
##  [5] "Astrophysics"
##  [6] "*Amanda Bouvier Edmonds"
##  [7] "*Douglas Bertrand Marshall III, with highest"
##  [8] "honors in Philosophy"
```

```
##  [9] "*Michelle Pacholec, with honors in Chemistry"
## [10] "*Grace Martha Pritchard, with honors in English"
```

A series of operations is then conducted to produce the output dataframe. Lines that separated by line breaks are first joined back together into a single line. The "*" and "+" signs in front of the students' names are then removed. Suffixes of the name like "Jr." or Roman numerals are also removed. The first name, last name, department honor and latin honor are then extracted and put into a dataframe. The gender of the student is estimated from the student's first name using `gender` and `genderdata` packages.

```
> readgraduates(2000)
```

```
##     lastName firstName             major departmentHonors     latinHonors gender
## 1      Chung    DoHyun Political Economy           Honors Summa Cum Laude   <NA>
## 2      Cover   Rebecca      Astrophysics   Highest Honors Summa Cum Laude female
## 3    Edmonds    Amanda              <NA>             <NA> Summa Cum Laude female
## 4   Marshall   Douglas        Philosophy   Highest Honors Summa Cum Laude   male
## 5   Pacholec  Michelle         Chemistry           Honors Summa Cum Laude female
## 6  Pritchard     Grace           English           Honors Summa Cum Laude female
## 7    Ramberg   Michael              <NA>             <NA> Summa Cum Laude   male
## 8  Schildgen    Taylor       Geosciences   Highest Honors Summa Cum Laude female
## 9        Sun     Qiang              <NA>             <NA> Summa Cum Laude   <NA>
## 10     Trice     Laura              <NA>             <NA> Summa Cum Laude female
```

However, since there are 8354 names to be estimated in total and the `gender` package takes a relatively long time to estimate the gender of each name, the data used for `showstats` is a pre-loaded dataframe using `data()`.

Use `readgraduates`

The information within the locally stored text files can be accessed by using the `readgraduates` function which uses the method described above. This function outputs a dataframe containing the first name, last name, major(if available), department honors, latin honors and estimated gender for all graduated students in any given year from 2000 to 2015. However, as the `gender` package that estimates the gender of students takes a approximately 3-5 minutes to process over 500 names for each year, using, for example, `data(graduates2000)` to obtain the pre-loaded dataframe that contains the information on graduated students of 2000 would be a faster way. The function is used in the following way:

```
> readgraduates(2000)
```

The only argument is:

`year`: Selects the year of graduated students to be scanned.

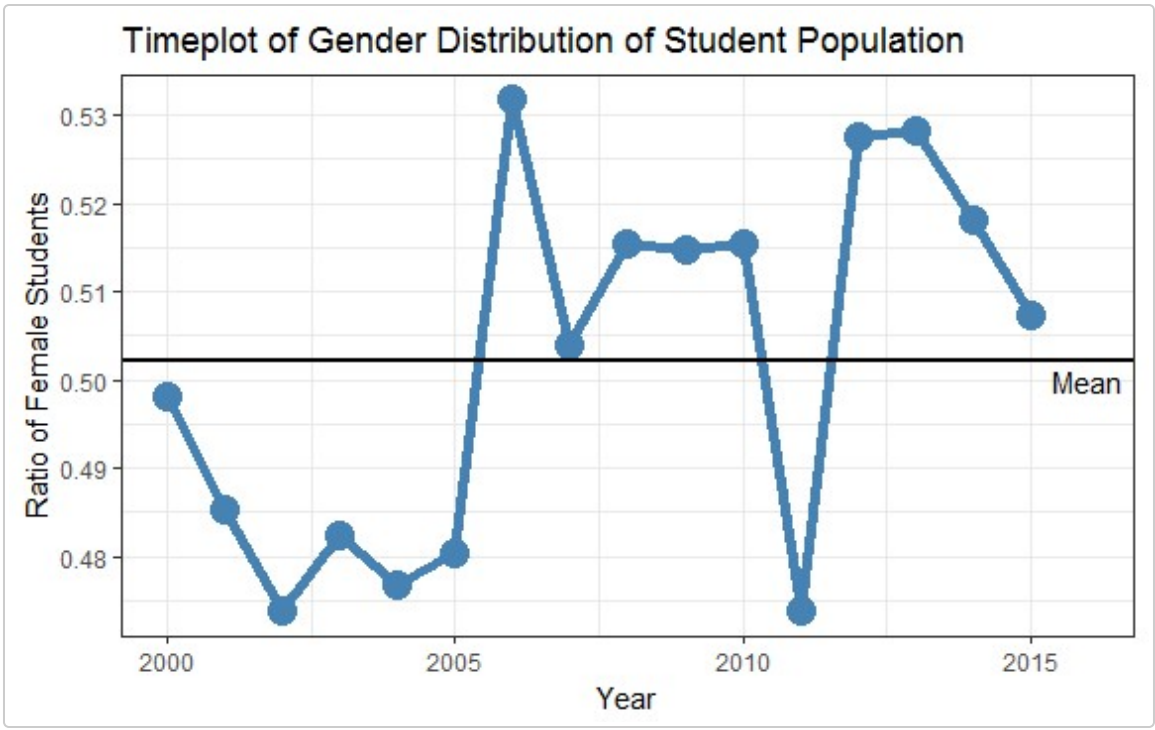
Use `showstats`

The function `showstats` uses the data from the `readgraduates` function to perform appropriate analysis and generates timeplots of the gender distribution of the student population. This function has only one argument and is used in the following way:

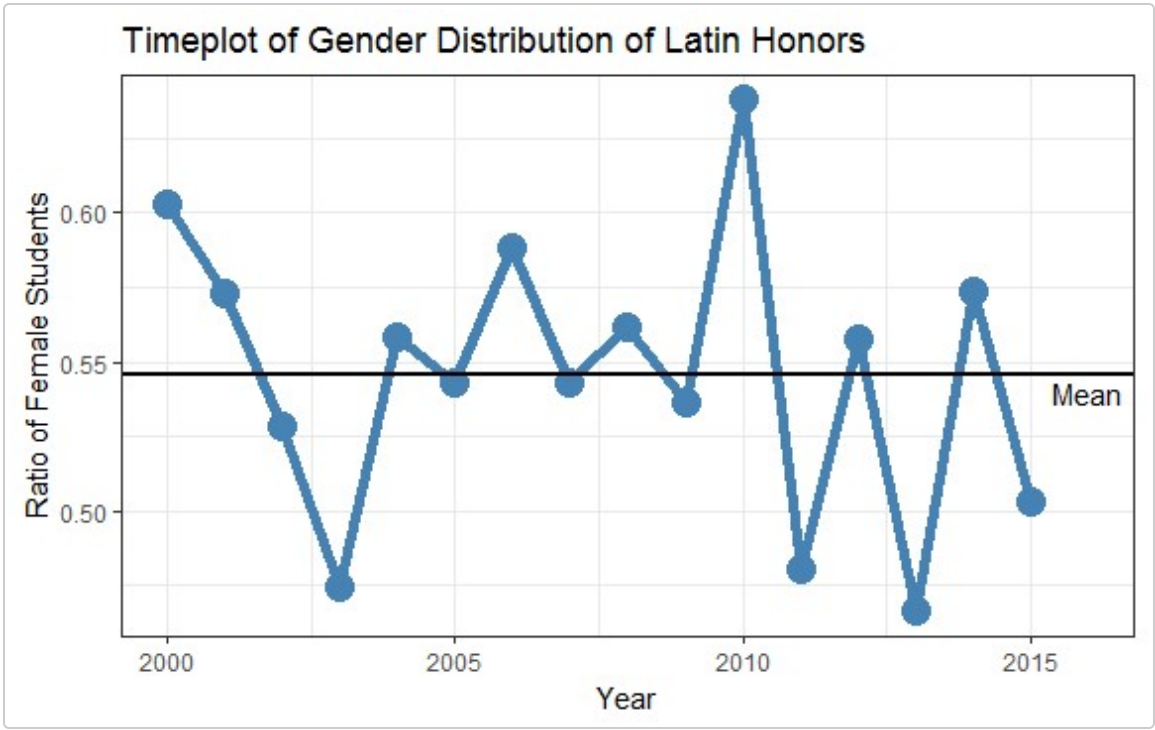
```
> showstats(type)
```

The argument `type` has six values (`graduates, latin honors, department honors, latin honors detail, department honors detail`), their results are as follows:

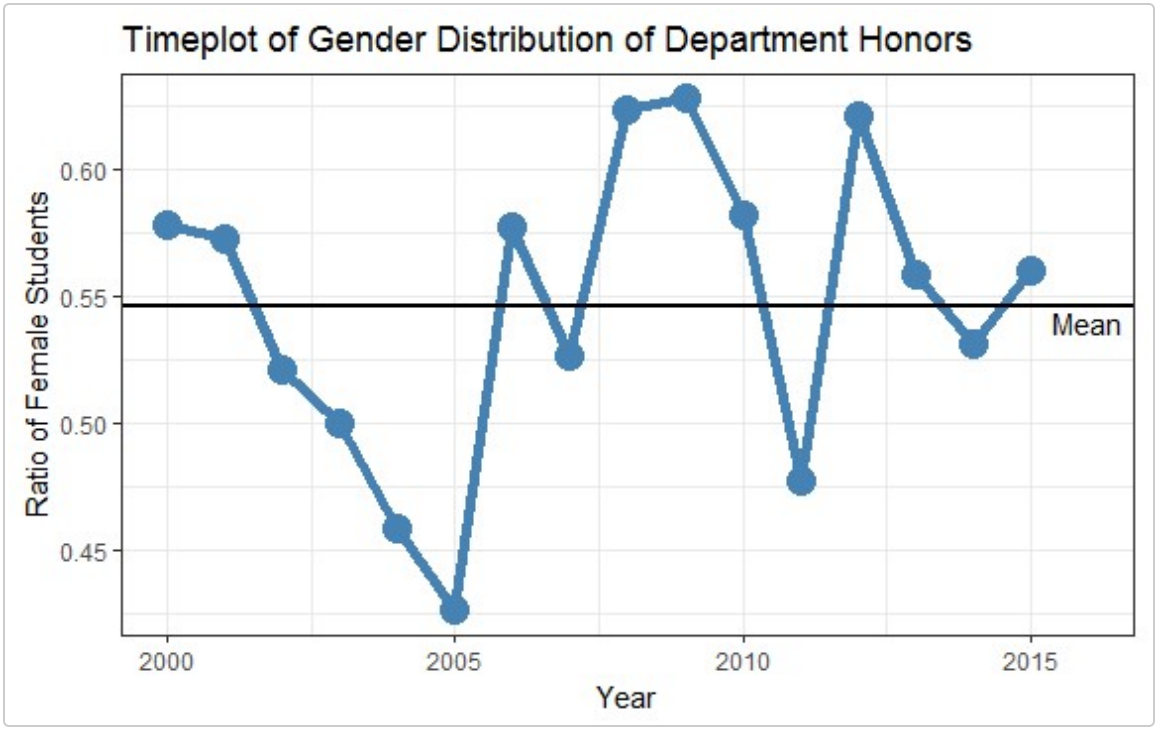
```
> showstats("graduates")
```



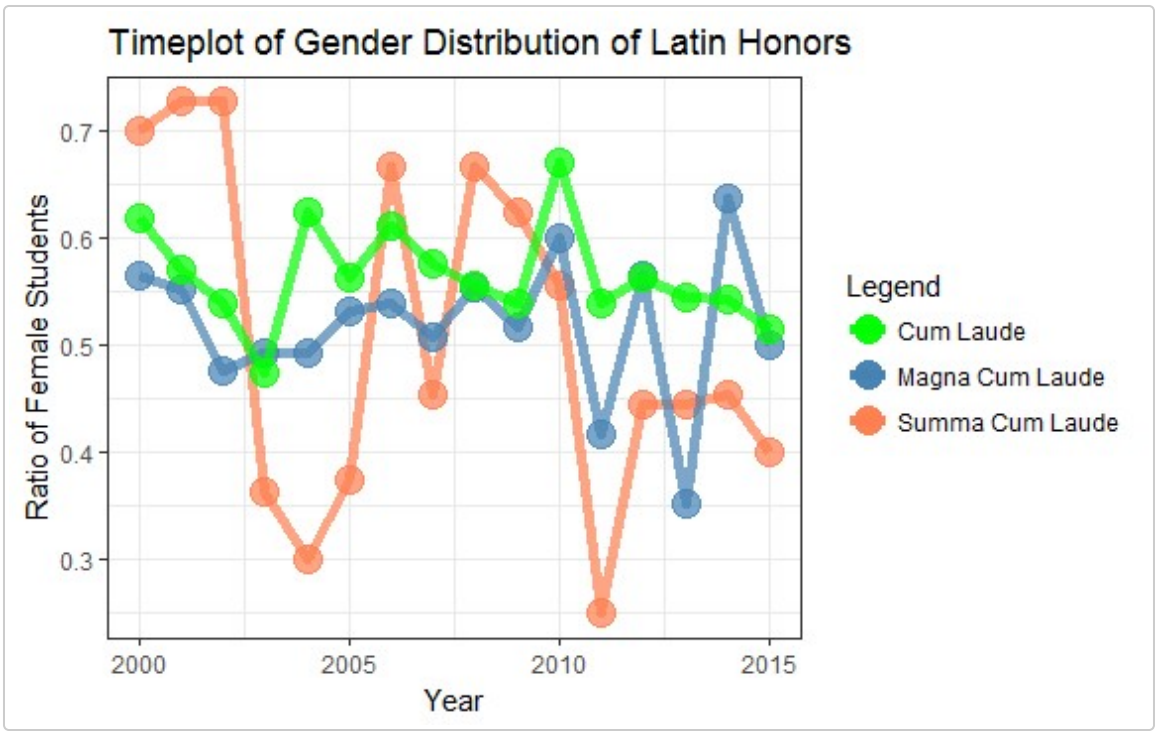

```
> showstats("latin honors")
```



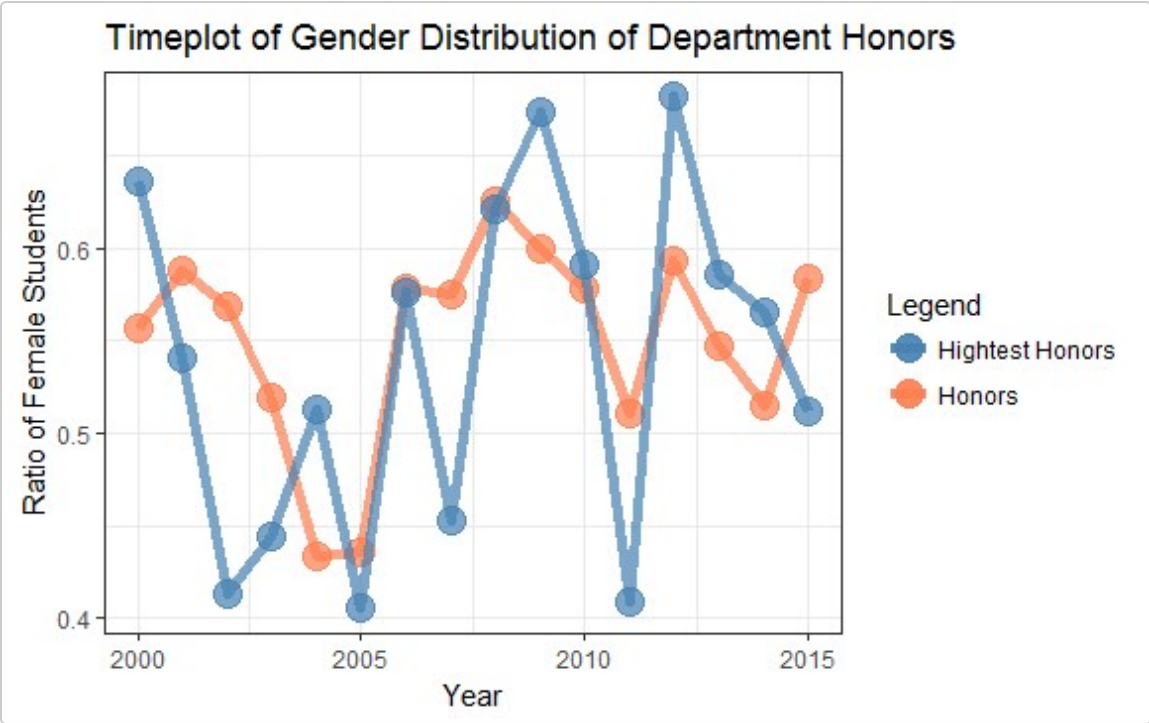

```
> showstats("department honors")
```




```
> showstats("latin honors detail")
```




```
> showstats("department honors detail")
```



Conclusion

The summary statistics indicate that the estimated gender distribution of graduated students in 2015 was around 50.7% female, and the mean gender distribution of graduated students from 2000-2015 was around 50.2% female, a very even distribution. The graduated students had the most percentage of female students in 2006 at around 53.1%. The graduated students had the least percentage of female students in 2002 and 2011, both at around 47.4%. The graphics produced above yield some interesting results that might be worth noting. While the graduated students had on average an equal amount of male and female students, the mean gender distribution of students that graduated with latin honors was 54.56% female and 54.65% female for students that graduated with department honors. This means that on average, there were proportionately more female students who has attained latin and department honors than male students from 2000-2015. However, it is also worth noting that the `gender` package could not produce a gender estimation to less common names or non-English names, which might have caused the data to be biased.

Bibliography

Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr

Lincoln Mullen (2015). gender: Predict Gender from Names Using Historical Data. R package version 0.5.1.

Lincoln Mullen (2017). genderdata: Historical Datasets for Predicting Gender from Names. R package version 0.5.0. https://github.com/ropensci/genderdata

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Office of The Registrar of Williams College
http://web.williams.edu/admin/registrar//catalog/archive.html