# Geographic Distribution of Williams College Students

*by Jackson Moss '19*

**Abstract** The **wgd** package uses publically available data from the Williams College Registrar to load and determine statistics about the geographic distribution of the student body at Williams College between 2001 and 2016. The analysis is completed for the 50 states, District of Columbia, Puerto Rico, and Guam. The package creates various statistics and graphics illustrating the changes in the geographic distribution of the student body over the past 15 years.

## Introduction

The geographic distribution of the study body, or more simply, where students are from, is a dynamic that heavily influences the social, ideological, and academic environment of the college. An understanding of how this structure changes over time is important to the diversity of the college and should be watched closely. The package wgd uses string manipulation on tables extracted from PDFs and organizes the information in the form of a usable data frame.

The package **wgd** also creates a variety of statistics and graphics that illustrate changes in the geographic distribution on a country, region, and state level. Users can create bar plots showing how the number of students from a certain state or region changes over time. These introductory graphics show geographic distribution trends a clear and concise fashion. The data can also be used to perform more advanced analysis. For example, external data sets could be used to determine what variables are associated with changes in the Williams College geographic distribution.

## Data

The raw data used for this analysis is in the form of PDFs posted to the website of the Office of the Registrar of Williams College. The data of interest are stored in tables inside the PDFs, and are available for the years between the years 2001 and 2016. In 2011, the Office of the Registrar began recording the geographic distribution data in a different way in previous years. Thus, two different methods were employed to read the data into a usable format, one for the years 2001 to 2010, and one for the years 2011 to 2016.

For 2001 to 2010, the data was converted into text files using PDFelement, an online PDF to text editor. The result is a character vector containing two distinct sections. The first x lines reference the names of the states and the next x lines reference the number of students from each state. There are many unneeded spaces and periods in the raw text file. To clean this, spaces and non-alphanumeric characters are removed.

```
rm_spaces <- gsub(" ", "", frame)
rm_stuff <- gsub("[[:punct:]]", "", rm_spaces)
```

Two new character vectors are created, one referring to the name of the state and one referring to the number of students from that state. Binding them together creates the format we want.

```
names <- rm_one[nchar(rm_one) > 3]
numbers <- rm_one[nchar(rm_one) <= 3]
full <- cbind(names, numbers)
```

For 2011 to 2016, the data can simply be copy and pasted from the PDF into a text file, and then manually edited to contain a new line for each state reference. Unneeded spaces are removed, and vectors of the names of the states and the number of students from each state are created and bound together.

```
rm_spaces <- gsub(" ", "", data)
states <- gsub("[0-9]", "", rm_spaces)
numbers <- gsub("[a-zA-Z]", "", rm_spaces)
total <- cbind(states, numbers)
```

These operations yield a two column matrix, one column representing the names of the states and one column representing the number of students from each state. However, this 2 column matrix has a variable number of rows depending on the text file that is being manipulated. In the instance that

a state has zero students from it, the raw data does not contain that state in the table. Since certain states sometimes have zero students at Williams, matrices will often contain a different number of rows and have missing states. In order to report data on all states between the years of 2001 and 2016, the relevant data must be added in. A solution is to create a 1 x 2 matrix with the relevant information and bind it to the data. Several if statements check for this condition and create the missing data. Do not mind the matrix created for the else condition, as it is a placeholder and will be deleted.

```
if(identical(total[total[, 1] %in% c("Nebraska", "Nebraska\t"), 1], character(0)))

{ a1 <- matrix(c("Nebraska", "0"), 1, 2) } else { a1 <- matrix(1:2, 1, 2) }
```

Next, the original data and the new matrices are bound together. The placeholder matrices are removed and the matrix is ordered by alphabetically by state name. The matrix is coerced into a data frame that has 53 rows: The 50 states, District of Columbia, Guam, and Puerto Rico.

```
readEphData("williams2012.txt")[1:4, ]

#>      state number
#> 1  Alabama      7
#> 2   Alaska     11
#> 3  Arizona      9
#> 4 Arkansas      7
```

**Use readEphData**

The information in the locally stored text files can be loaded and parsed using the readEphData function. The function creates a dataframe containing the geographic distribution of students for the year of the user coded text file. Subsequent functions combine the data and format it into a data frame. The function has one argument and can be used as follows.

```
readEphData("dataset")
```

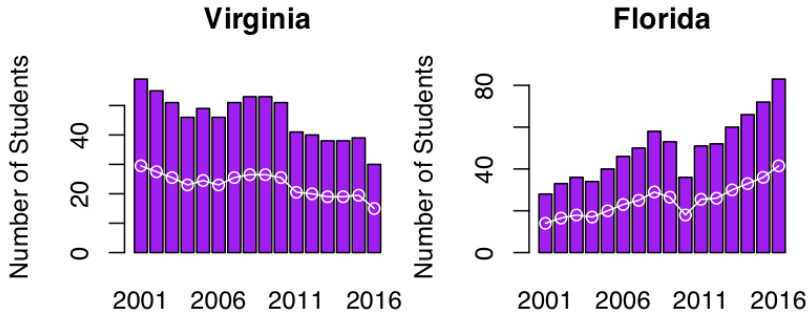
**Use plot_state**

The function **plot_state** is used to generate graphics that show the geographic distribution over time for a user-coded state. The function subsets the data for a particular state and then displays the information in a bar plot. The function has one argument and is used as follows.

```
plot_state("State")
```

The function can display graphics for any of the 50 states, District of Columbia, Guam, and Puerto Rico. The user should code the state without spaces and with capital letters where normal. For example:

```
plot_state("Virginia")
plot_state("Florida")
```




**Use plot_region**

The function **plot_region** is used to generate graphics that show the geographic distribution over time for a user-coded region. The function subsets the data for a particular region and then displays the information in a plot. The function has one argument and is used as follows.

```
plot_region("Region")
```

# Geographic Distribution of Williams College Students

*by Jackson Moss '19*

**Abstract** The **wgd** package uses publically available data from the Williams College Registrar to load and determine statistics about the geographic distribution of the student body at Williams College between 2001 and 2016. The analysis is completed for the 50 states, District of Columbia, Puerto Rico, and Guam. The package creates various statistics and graphics illustrating the changes in the geographic distribution of the student body over the past 15 years.

## Introduction

The geographic distribution of the study body, or more simply, where students are from, is a dynamic that heavily influences the social, ideological, and academic environment of the college. An understanding of how this structure changes over time is important to the diversity of the college and should be watched closely. The package wgd uses string manipulation on tables extracted from PDFs and organizes the information in the form of a usable data frame.

The package **wgd** also creates a variety of statistics and graphics that illustrate changes in the geographic distribution on a country, region, and state level. Users can create bar plots showing how the number of students from a certain state or region changes over time. These introductory graphics show geographic distribution trends a clear and concise fashion. The data can also be used to perform more advanced analysis. For example, external data sets could be used to determine what variables are associated with changes in the Williams College geographic distribution.

## Data

The raw data used for this analysis is in the form of PDFs posted to the website of the Office of the Registrar of Williams College. The data of interest are stored in tables inside the PDFs, and are available for the years between the years 2001 and 2016. In 2011, the Office of the Registrar began recording the geographic distribution data in a different way in previous years. Thus, two different methods were employed to read the data into a usable format, one for the years 2001 to 2010, and one for the years 2011 to 2016.

For 2001 to 2010, the data was converted into text files using PDFelement, an online PDF to text editor. The result is a character vector containing two distinct sections. The first x lines reference the names of the states and the next x lines reference the number of students from each state. There are many unneeded spaces and periods in the raw text file. To clean this, spaces and non-alphanumeric characters are removed.

```
rm_spaces <- gsub(" ", "", frame)
rm_stuff <- gsub("[[:punct:]]", "", rm_spaces)
```

Two new character vectors are created, one referring to the name of the state and one referring to the number of students from that state. Binding them together creates the format we want.

```
names <- rm_one[nchar(rm_one) > 3]
numbers <- rm_one[nchar(rm_one) <= 3]
full <- cbind(names, numbers)
```

For 2011 to 2016, the data can simply be copy and pasted from the PDF into a text file, and then manually edited to contain a new line for each state reference. Unneeded spaces are removed, and vectors of the names of the states and the number of students from each state are created and bound together.

```
rm_spaces <- gsub(" ", "", data)
states <- gsub("[0-9]", "", rm_spaces)
numbers <- gsub("[a-zA-Z]", "", rm_spaces)
total <- cbind(states, numbers)
```

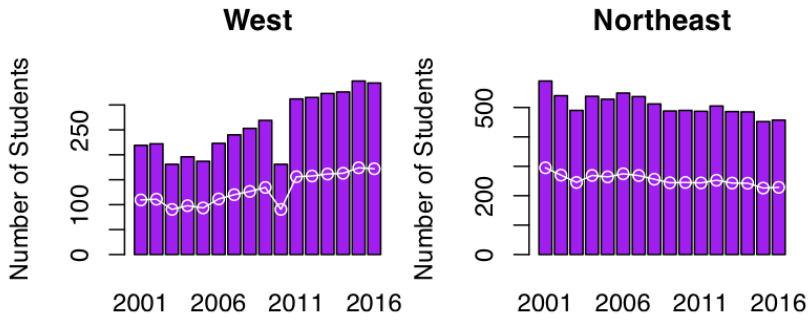
These operations yield a two column matrix, one column representing the names of the states and one column representing the number of students from each state. However, this 2 column matrix has a variable number of rows depending on the text file that is being manipulated. In the instance that

The function can display graphics for any of the six regions that this package defines. The regions and states they contain are as follows:

- West: CA, OR, WA, ID, NV, AZ
- Mountain West: NM, CO, WY, MT, ND, SD, UT
- South: OK, TX, AR, MS, FL, GA, SC, NC, TN, LA, AL
- Midwest: NE, KA, IA, MO, OH, KY, IN, IL, WI, MI, MN
- Northeast: NY, MA, VM, NH, ME, RI, CT
- Mid Atlantic: VA, DE, DC, NJ, WV, PA

The user should code the state without spaces and with capital letters where normal. For example:

```
plot_region("West")
plot_region("Northeast")
```




**Use total_dist**

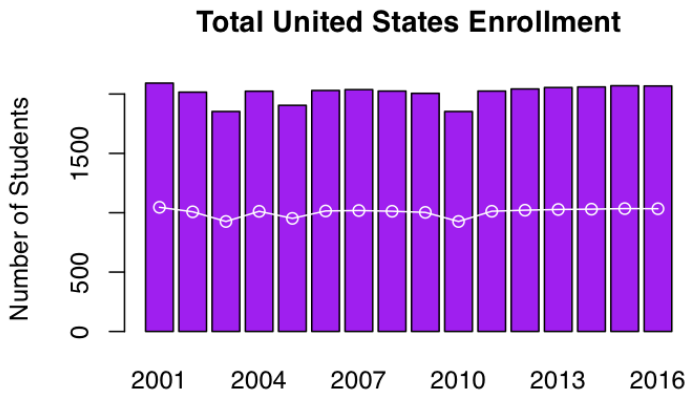
The function **total_dist** uses the geographic distribution data from the readEphData function to generate graphics and summary statistics for user coded options. The function has the argument type which dictates the type of output generated. The possible options are Summary, US, and SD. Summary generates a data frame containing the geographic distribution for the years 2001 to 2016. US generates a bar plot displaying how total United States enrollment has changed at Williams from 2001 to 2016. SD generates a data frame containing the standard deviation in total United States enrollment for the year blocks 2001 to 2010 and 2011 to 2016. The function can be used as follows

```
total_dist("Summary")[1:4, ]
```

### Discussion

Before delving deeper into the data by state or region, let's look at the total number of students from the 50 states, District of Columbia, Guam, and Puerto Rico between 2001 and 2016.

```
total_dist("US")
```



A quick glance at the bar plot shows that the total number has not changed much in the past 15 years. What is surprising is how flat the distribution has been over the past six years versus the ten years before that. Quick calculations of the standard deviation for each time period yields an interesting result.

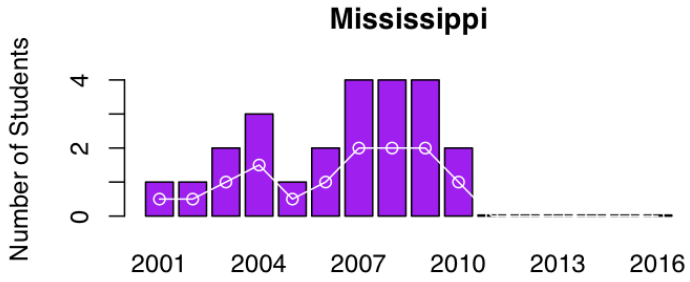
```
total_dist("SD")

#>    2001-2010 2011-2016
#> SD  82.95976  17.02645
```

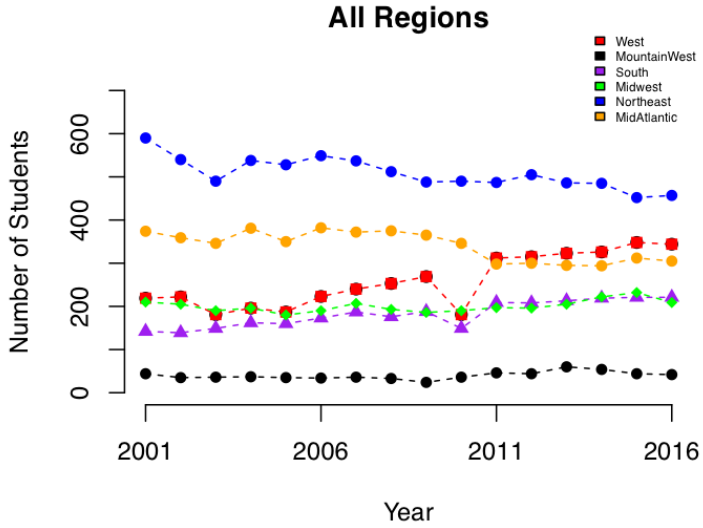
The standard deviation of total United States enrollment for the past six years is almost five times smaller than it was for the preceding ten years. Did admissions become "smarter" about yield calculations sometime near 2011? What other factors could potentially explain this?

Output from the **plot_state** function shows marked upward and downward trends for different states. California, with only 156 students in 2001, has become a Williams powerhouse with 267 students in 2016. Florida has also seen huge gains, rising from 28 in 2001 to 83 in 2016. Some states have opposite trends. District of Columbia had 26 students at Williams in 2001 and 2002, but only 12 in 2015 and 2016. Massachusetts has also seen decreases, from 341 to 280 from 2001 to 2016.

One especially interesting state is Mississippi. The southern state had an average of 2.4 students at Williams between 2001 and 2010. Not a single student from Mississippi has matriculated since then.




Output from the **plot_region** function shows marked upward and downward trends for different regions.




A quick glance shows decreases in the Northeast and Mid Atlantic Regions, and increases in the West and South Regions. The most notable aspect of the graphic is that the Northeast and West regions have been converging over the past 15 years.

## Conclusion

The student geographic distribution data, on both a state and region level, show marked downward and upward trends. This is an important result that has no doubt influenced the social, ideological, and academic environment of the college. A rough statement summarizing the data would be as follows: Between the years of 2001 and 2016, Williams has become increasingly represented by students from the West and South and decreasingly represented by students from the Northeast and Mid Atlantic.

What is more important than this result are the reasons why these trends have occurred. Why has the number of Students from Florida increased almost by a factor of three. Why hasn't a single student from Mississippi matriculated to Williams since 2010 after an average of 2.4 matriculated the previous ten years? It is unclear whether these trends are due to internal bias (i.e. changing admissions goals/quotas for students from certain states or areas of the country) or the quality of the students from different areas of the country.

As follows is a quick conceptual model outlining potential reasons why the geographic distribution might change. This model may guide further research in this topic.

**Internal Bias** - Williams desires to admit students from all 50 states for diversity purposes - Williams desires more/less students from x state or region for y reason

**Quality of Students** - Williams has become more well known in areas like the West and South over the past 15 years, leading to more and better qualified applications. Thus, more students have matriculated from those areas - The educational quality in state x has changed over time, resulting in more or less students at Williams from state x

**Other Reason** - A Williams College admissions counselor in charge of state or region x is replaced by a counselor who is better/worse at convincing quality students to apply to Williams, and thus more/less students from state or region x matriculate to Williams

The actual underpinnings for the geographic distribution of the student body are almost certainly a combination of the three described above, and many others. Isolating and compiling data for many of these reasons would be a difficult task.

An interesting data set to compare with the results from the **wgd** package would be data on the admission rates by state. If the admission rate for a certain state or region remains constant while the number of students from that state increases or decreases, the change is likely attributed to the number of students applying from that area. If the admission rate for a certain state or region trends up or down, then the change is likely due either the quality of students or internal bias in Williams admissions. Unfortunately, it is unlikely that Williams would release this data.

Another interesting data set for comparison would be geographic distribution data from a school similar to Williams. Under the premise that students at Williams from the West and South have increased over the past 15 years due to increased awareness or recognition of Williams, it would be interesting to see if a similar trend has occurred at schools with recognition and awareness that has presumably not changed, or at least not a much. Harvard, for example, has always been well known and attractive to students all across the country.

Regardless, the package **wgd** includes functions that allow easy reading and analysis of geographic distribution data published by Williams College. Depending on the format of the data, this method can be generalized to other institutions that publish data in a similar format. **wgd** provides a good starting point for analyzing the geographic distribution of the Williams College student body. As mentioned above, outside data sets would need to be pulled in order to analyze the possible underpinnings for the results.

*Jackson Moss '19*
*Williams College*
*Williamstown, MA, USA*
jm21@williams.edu