names

Jian Lu '19

2/12/2017

Abstract The **names** package takes data for each graduating class from 2003 to 2015 published in the Williams College Course Catalogs. It reformats it into a data frame for easy manipulation. The second component of this package displays the proportions of Jewish people and the level of merit received upon graduation using line graphs. This project was completed as part of the application process for Hutchin Hill at the request of Dr. David Kane.

Introduction

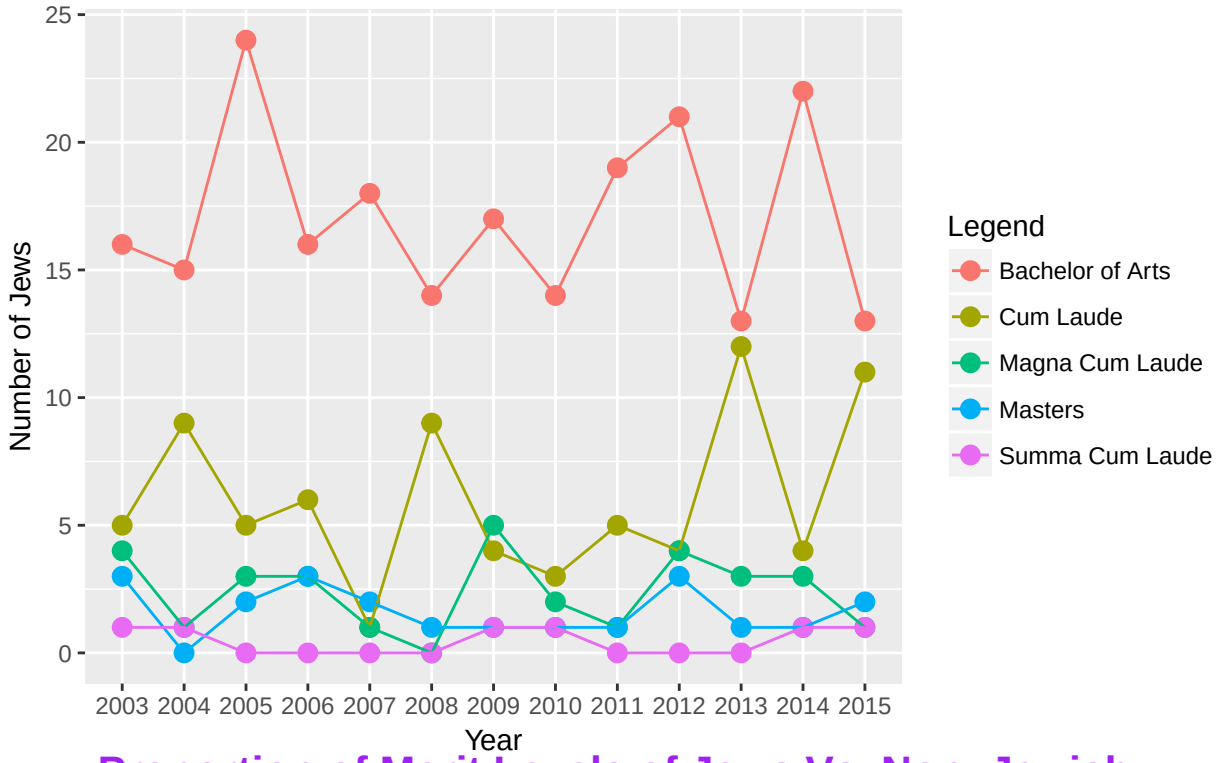
The data on graduating students of Williams College shows the merit each individual student has received at Williams. The data frame constructed from this data contains information from all students from the Class of 2003 to the Class of 2015. However, majors are only listed for students who graduated with honors, not the entire graduating class. With the tools available in this package, the success rate of Jewish people at Williams College can be measured. To do so, the proportion of Jewish students that graduated at certain merit levels are graphed against the proportion of non-Jewish students. The data set includes both undergraduate students and economic policy graduate school students. The variables in the data set include graduation year, major (if available), type of honors (if available), Phi Beta Kappa, Sigma XI, and ethnicity (specifically if the student of Jewish descent).

Data

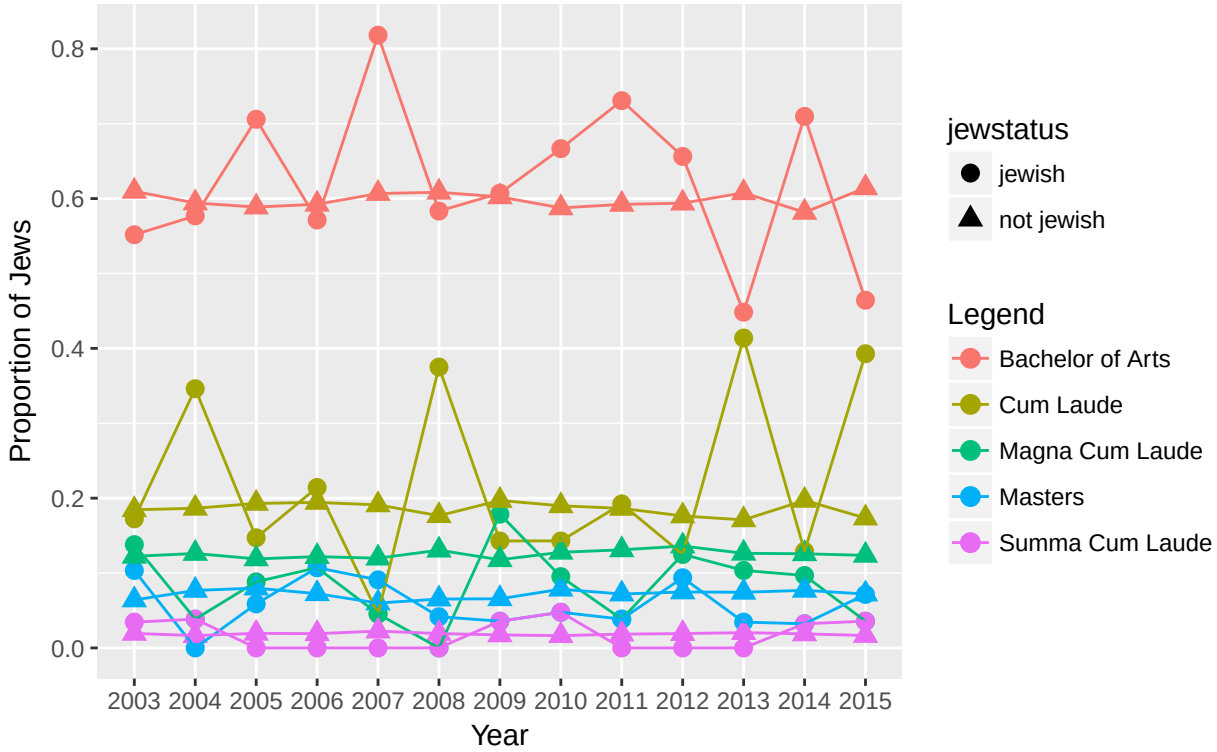
The data used to construct the overall data frame in this package was taken from the Williams College Course Catalogs found on the website of the Office of The Registrar of Williams College for the years 2003 through 2015. Since there are some slight variation in the method used to present each year's information, some modifications were made to the text files. The data files were changed so that every line contained the information of only one student, as well any page numbers, and irrelevant information was removed. All of the data files were organized in the same homogeneous fashion, which allowed for the function **readnames** to easily organized the information into a data frame. Each row includes the full name of each person and is followed by personal information in the same row. The last variable of ethnicity is not provided by the college. Thus, a list of around 3000 Jewish last names was used to evaluate whether or not the student was of Jewish descent.

The process of creating the data frame begins with iterating through every single line of text within each individual text file of each years data. All these text files were copy pasted from the Williams College Course Catalogs files that are posted each year. Then lists of all students whom graduated with Summa Cum Laude, Magna Cum Laude, Cum Laude, and Master's degrees were created so that they could be identified later. This allowed for the deletion of any extraneous strings within the text file such as category separations, and page numbers. 9 variables were then constructed and filled out depending on the data of each line of text. An example of the process is shown below.

Merit Levels of Graduating Jews



Proportion of Merit Levels of Jews Vs. Non−Jewish



Conclusion

The data implies that each year there is an average of 27.53 students of Jewish descent attending Williams, copared to an average of 530.61 students that are not of Jewish descent. The range of Jewish students in each class year was 13 with the minimum of 21 students in the 2010, and a maximum of 34 students in 2005. It can be seen from the timeline graphic that the count of Jewish people at Williams College has stayed relatively the same over the course of 13 years. We can see that there is not much deviation from the mean of 27 Jewish students attending Williams throughout the years. This may point toward the fact that Williams College is fulfilling quotas of certain ethnic groups. Significant conclusions, however, can only be drawn after receiving data and testing not just Jewish people but African Americans, Asians, and Caucasians. The comparison graphic show that Jewish people at Williams perform, on average, the same as the non-Jewish population. In order to compare with statistical significance, I used a difference of means test to check if being Jewish or not really affects academic success. In the following five hypothesis tests, we set the H_0: Proportion of Jewish - Proportion of Non-Jewish = 0 and the H_A: Proportion of Jewish - Proportion of Non-Jewish $\neq$ 0. We are 95% confident that the true difference in means of proportions of Jewish people and non-Jewish who receive a **Bachelor of Arts** degree lies between (-.04, .08), however the p-value = 0.4335, so we cannot reject the null. We are 95% confident that the true difference in means of proportions of Jewish people and non-Jewish who receive **Cum Laude** lies between (-.04, .11), however the p-value = 0.3526, so we cannot reject the null. We are 95% confident that the true difference in means of proportions of Jewish people and non-Jewish who receive **Magna Cum Laude** lies between (-.07, -.01), however the p-value = 0.01133, so we can reject the null that the H_0: Proportion of Jewish - Proportion of Non-Jewish = 0. We are 95% confident that the true difference in means of proportions of Jewish people and non-Jewish who receive **Summa Cum Laude** lies between (-.01, .01), however the p-value = 0.7939, so we cannot reject the null. We are 95% confident that the true difference in means of proportions of Jewish people and non-Jewish who receive a **Masters Degree** lies between (-.033, .006), however the p-value = 0.1654, so we cannot reject the null. If success is akin to receiving a Master's degree or achieving any sort of cum laude, then the Jewish population typically performs on par with the rest of the student population as we can see that the proportions for both groups are relatively the same. From the five hypothesis tests above, we see that we do not have a large enough sample size to draw any statistically significant conclusion, except for the test on Magna Cum Laude. However, if we look at the confidence interval for Magna Cum Laude, we can see that it is not very far from zero. Although there are some dips in performance in the classes of 2007 and 2015, and some out performance by the Jewish population in the classes of 2008, 2013 and 2015. From this data and the hypothesis tests, we can infer that the Jewish ethnicity is independent of success at Williams, as the proportions of merit levels of the Jewish population and the non-Jewish population are relatively similar.

The **names** package includes functions that easily create graphics of the proportion of Jewish students at Williams College and the level of merit they have received upon graduation. The data set created in the package is organized so that it can be easily subsetted and tested for other correlations, or interests of the user. For example, one can view all the majors at Williams college, and subset the majors with ethnicity to view which majors are most popular with the Jewish population. However there are many improvements that could be made to this package. Currently, it only tests for students of Jewish descent based on last name. So one potential development for the package could be adding ethnicity tests for other ethnic groups to yield better results. With the current method of labeling students of Jewish descent, there may be some Type I and Type II errors, so a huge development for the package would be to get the ethnic data on each student from Williams and implementing it into the data frame. As more years go by, I hope to gather more data and increase the sample size resulting in smaller p-values for the hypothesis tests. It would be ambitious to conclude that ethnicity is completely independent of success at Williams, but this poses another question and with further manipulation of the data on students, it would be possible to look into.