Data Science Capstone

First-Gen @ Wellesley:

Comparisons of Post-Bac Outcomes Between First-Gen and non First-Gen Alums at Wellesley

Briana Vigil ’22

Objectives & Research Question:

First-generation students make a small yet integral part of the Wellesley community, and their experiences in higher education are varied, but research shows that first-generation students “often face psychological, academic, financial, and social challenges, and about one in three leave college within the first three years” (mghclaycenter.org) This brings me to my question of: Having persevered against the obstacles at Wellesley and once graduated, do first-gen Wellesley alums have different outcomes than other peers in terms of pursuing higher degrees?

Data Collection & Cleaning

In order to analyze this question, I used data scraped from the WellesleyHive, an online networking platform created exclusively for Wellesley alums, current students, and faculty. I filtered the data for ‘First-generation’ and ‘Alum’ to ensure that everyone I have in my first-generation data are as accurate as possible. For the non first-gen alums, I simply filtered for “Alum” and scraped the data from those profiles. I cleaned and organized the data using Python’s BeautifulSoup and Pandas libraries. Once cleaned, the data were converted into csv files and imported into R to begin with the analysis.

Variable Selection & Methodologies

To facilitate my analysis, I created several “dummy variables” that indicated whether an alum has obtained a Master’s or PhD, or if they are a current Graduate Students, as well as whether the alum’s major was in the STEM, Social Science, or Humanities field. I also extracted the year they graduated and approximated their age that would be accurate given the assumption that they graduated within 4 years and were a traditional student rather than a Davis Scholar. Although it may be inaccurate to assume so, this method facilitated my approach to come to my analysis. With these variables, I was able to begin my analysis for the pursuit of higher education for first-generation and non first-generation Wellesley alums.

Results:

Upon seeing the results of the boxplot, I made the decision to subset my original data to only look at the people who are 65 and younger, due to decreasing of data as well as overfitting that was skewing the results. The boxplot in Fig. 1 aided in my decision since it clearly depicts the sparseness of the data once the alum age is over 65. In Fig. 2, we can see that the scatterplot matrix depicts the strongest correlation between ApproxAge and Masters, which aided in my final decision of subsetting of the data. In addition to this, I also wanted to visualize the proportion of First Generation alums to the different predictors that are in my models. Figures 3-7 depict this, where we can see the proportion of alums with: Masters degrees, PhDs, in STEM, Humanities, and Social Science fields. In all of these visualizations, we see that First Generation alums are slightly lower in proportion to non First Generation counterparts.

To view the variables that were most important, I also conducted a series of logistic regressions for both Masters and PhD with all other predictors that were relevant to the research question. Figure 8 depicts the series of logistic regression that were run for the Masters outcome, where I ultimately concluded that the model with all of the predictors plus the interaction between Humanities and Age was the best one because of the lower AIC of 1707.65. Figure 9 depicts the series of logistic regression that were run for the PhD outcome, where I ultimately concluded that the model with all of the predictors plus the interaction between First Generation and Age was the best one because of the lower AIC of 617.5. In addition to this, what we can also see from this summary is that the most significant predictor is age, but there are also negative correlations with the outcome and being First Generation.

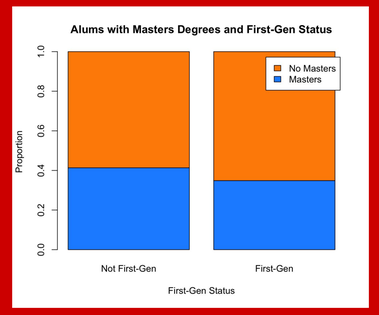

Fig. 3. Barchart that depicts the proportion of Wellesley alums (both First Generation and non) and the amount of Masters degrees that they received

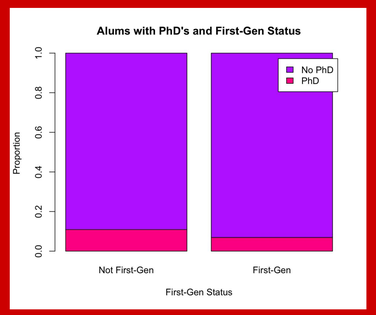

Fig. 4. Barchart that depicts the proportion of Wellesley alums (both First Generation and non) and the amount of PhD’s that they received

	Masters Degree - all predictors	Masters Degree - no disciplines	Masters Degree - interactions for first gen and age variables	Masters Degree - interactions between age and STEM	Masters Degree - interactions between age and Social Science	Masters Degree - interactions between age and Humanities
(Intercept)	−3.222***	−2.776***	−3.357***	−3.222***	−2.947***	−3.694***
	(0.268)	(0.229)	(0.295)	(0.285)	(0.350)	(0.342)
FirstGen	−0.167	−0.183	−0.040	−0.166	−0.176	−0.270
	(0.138)	(0.136)	(0.528)	(0.138)	(0.138)	(0.138)
STEM	0.131		0.132	−0.246	0.121	0.133
	(0.160)		(0.160)	(0.575)	(0.159)	(0.160)
SocialScience	0.636***		0.635***	0.636***	−0.048	0.651***
	(0.128)		(0.128)	(0.128)	(0.452)	(0.128)
Humanities	0.358**		0.358**	0.359**	0.360**	1.100*
	(0.120)		(0.120)	(0.120)	(0.120)	(0.458)
ApproxAge	0.042***	0.043***	0.063***	0.040***	0.052***	0.071***
	(0.006)	(0.006)	(0.007)	(0.006)	(0.008)	(0.008)
FirstGen × ApproxAge			−0.003			
			(0.013)			
STEM × ApproxAge				0.010		
				(0.015)		
SocialScience × ApproxAge					0.018	
					(0.012)	
Humanities × ApproxAge						−0.019*
						(0.011)
Num.Obs.	1392	1392	1392	1392	1392	1392
AIC	1758.4	1733.8	1718.3	1759.9	1707.9	1707.7
BIC	1739.8	1751.5	1747.0	1746.6	1744.6	1744.3
Log.Lik.	−946.191	−864.905	−848.160	−847.954	−846.941	−846.826
RMSE	1.11	1.12	1.11	1.11	1.11	1.11

*p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 8. Model Summary of all logistic regression models conducted for Masters Degrees. This shows the different kinds of models that were run to see which one was optimal for the purposes of this research question

	PhD Degree - all predictors	PhD Degree - no disciplines	PhD Degree - interactions for first gen and age variables	PhD Degree - interactions between age and STEM	PhD Degree - interactions between age and Social Science	PhD Degree - interactions between age and Humanities
(Intercept)	−13.054	−13.661***	−12.763	−12.781	−13.126	−11.784
	(613.711)	(1.346)	(618.960)	(604.623)	(613.632)	(2142.281)
FirstGen	−0.554*	−0.433+	−2.407*	−0.553*	−0.553*	−0.554*
	(0.267)	(0.236)	(1.096)	(0.270)	(0.267)	(0.267)
STEM	1.923***		1.548***	0.260	1.923***	1.923***
	(0.272)		(0.275)	(1.187)	(0.275)	(0.272)
SocialScience	0.288		0.274	0.277	0.419	0.288
	(0.310)		(0.324)	(0.324)	(0.823)	(0.310)
Humanities	19.588		19.608	19.795	19.591	18.318
	(613.710)		(618.959)	(604.623)	(613.632)	(2142.281)
ApproxAge	0.044***	0.040***	0.036***	0.038***	0.045***	0.010
	(0.010)	(0.008)	(0.011)	(0.010)	(0.014)	(56.497)
FirstGen × ApproxAge			0.045+			
			(0.025)			
STEM × ApproxAge				0.045		
				(0.031)		
SocialScience × ApproxAge					−0.003	
					(0.019)	
Humanities × ApproxAge						0.033
						(56.497)
Num.Obs.	1392	1392	1392	1392	1392	1392
AIC	618.8	874.6	617.5	618.5	620.7	620.8
BIC	650.2	892.3	654.2	655.2	657.4	657.4
Log.Lik.	−303.378	−435.313	−301.750	−302.275	−303.362	−303.378
RMSE	0.66	0.79	0.66	0.66	0.66	0.66

*p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 9. Model Summary of all logistic regression models conducted for PhDs. This shows the different kinds of models that were run to see which one was optimal for the purposes of this research question

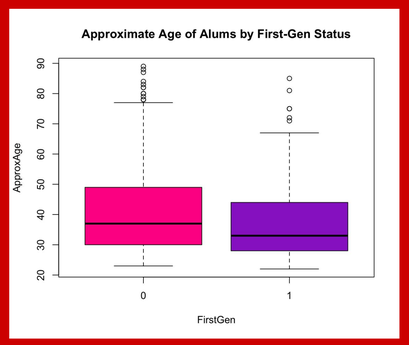

Fig. 1: Boxplot of Approximate Age and First Gen status of Wellesley Alums

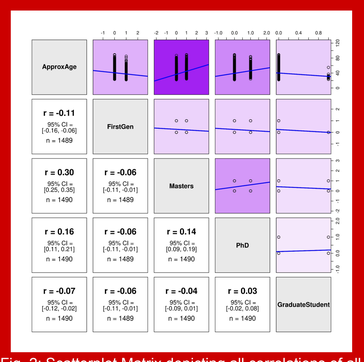

Fig. 2: Scatterplot Matrix depicting all correlations of all variables in dataset. Darker colors signify stronger relationships between variables

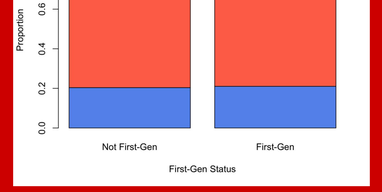

Fig. 5. Barchart that depicts the proportion of Wellesley alums (both First Generation and non) and the amount of alums who are primarily STEM oriented in their academic history

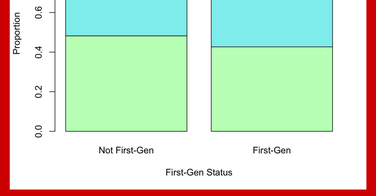

Fig. 6. Barchart that depicts the proportion of Wellesley alums (both First Generation and non) and the amount of alums who are primarily Humanities oriented in their academic history

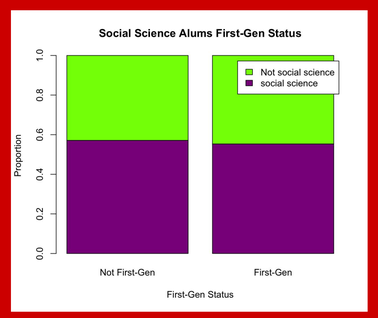

Fig. 7. Barchart that depicts the proportion of Wellesley alums (both first gen and non) and the amount of alums who are primarily Social Science oriented in their academic history

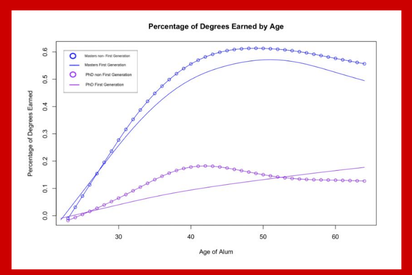

Fig. 10. Spline that depicts the percentage of Masters and PhD degrees received by both First Generation and non through the different ages of alums.

Discussion:

Although the results were not exactly what I had hoped (ideally there would have been no differences), these findings reveal that the gap between First Generation and non First Generation alums is relatively small and although not as significant as the age variable, an alum identifying as First Generation does negatively correlate with having a Masters and/or a PhD. It is also important to note that there may be missing data in terms of current graduate students since alums may have not updated their education history yet.

To close the gap, there is a need for further support and programming for First-Generation upperclassmen at Wellesley. There are already amazing structures in place such as WellesleyPlus and the FLI Network, and these are especially helpful for the transition to Wellesley. Wellesley also needs structures for the transition out of Wellesley, so that first-generation sibs are fully equipped with the resources and confidence needed for further pursue of education.

Further Research:

Prof. Jeremy Wilmer provided direction to the Standard Occupation Classification (SOC) by the CDC that I hope to include in further iterations of this project in order to observe any differences or lack thereof between occupations pursued by First-Gen or non First-Gen alums.

Special thanks to Rebecca Garcia, former dean at Wellesley College, for her unconditional support of myself and all other First-Generation students that had the honor to work with her.