Peter Kirgis
ADA
11/25/20

Acquire and Analyze: Emmaus, 2019-2020

Part 1: Executive Summary

This project looks at data from Emmaus, a homeless shelter and affordable housing developer in Haverhill, MA. To fulfill obligations to HUD and the Massachusetts DCHD, Emmaus must record extensive data for each client using their programs on both entry and exit, including demographics, employment, education, substance abuse, and domestic violence. They also track the outcomes for each client: Did they go back on the street? Or did they progress to temporary or permanent housing outside of the shelter?

All of this tracking takes place in Efforts to Outcomes (ETO), an online database for many kinds of nonprofits. While ETO has decent functionality, it was not amenable to many of the statistical techniques I wanted to pursue. So, I downloaded all of the data for a specific program at Emmaus, an individual homeless shelter called Mitch's Place, for both 2019 and 2020, to a set of CSV files. I then used Python to import and then merge these CSV files to a single working table in SQL for each year with all of my variables of interest, with one row for entry and one row for exit for each distinct enrollment in the program. I then brought the data into R to analyze the population served at Emmaus and the factors underlying differing outcomes for clients who entered the program.

As I had no knowledge of the program beforehand, my analysis uncovered many interesting facts and trends. The results did include one stunning finding, which will be the main topic of the *Future Directions* section of the paper. In conducting a logistic regression on a binary variable I created to distinguish successful from unsuccessful exit, I found that the variable *Domestic Violence Victim* to be the best indicator of whether someone would fail to move into housing. When keeping factors like age, gender, substance abuse, and history of

homelessness constant, a client with a history of domestic abuse reduces the probability of successful exit by almost 85%. As I will discuss more in the *Results* and *Future Directions*, I was surprised by this finding, and it suggests a potential hole in the support system that Emmaus provides.

Part 2: Introduction

On November 9, I began a year of service at Emmaus, Inc. as an *Americorps* member. My job title is "Grants and Data Specialist." Emmaus supports the homeless population of a small city of around 65,000 residents, and they run both an individual and family shelter, own over a dozen apartment buildings that they rent to clients at fixed rates (30% of monthly income), and provide additional assistance to the community in the form of rent subsidies, financial planning, and childcare services.

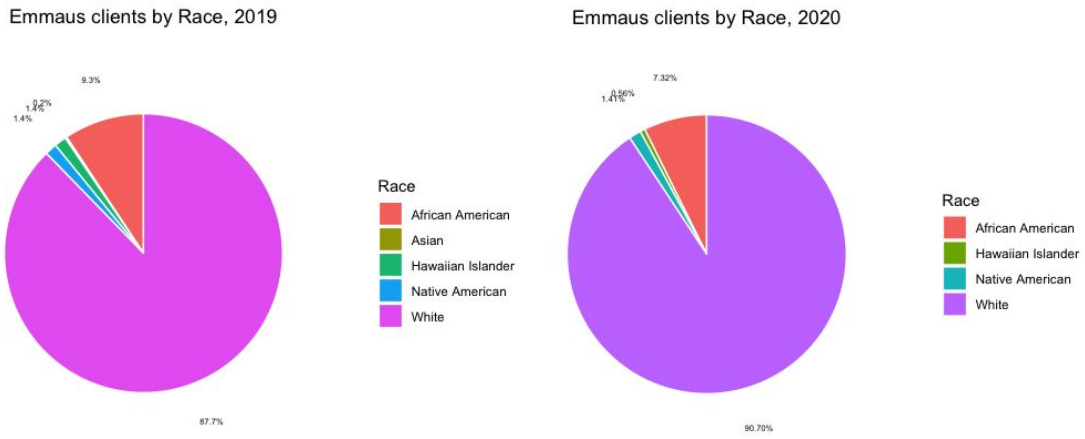
When I took the job, I was tasked by the Executive Director with the project of determining racial and ethnic descriptive stats and outcome differences among these populations in preparation for an upcoming report to the Massachusetts Department for Community Development (DCHD). She was concerned from preliminary tables she had done in Excel that African American clients in particular were not exiting the homeless shelter into housing at the same rates in White clients. Thus, this was the impetus and the guiding lens for my project. I knew, from the beginning of the project, that one of the main goals would be a logistic regression with a binary outcome variable and a categorical race variable with a variety of controls.

Part 3: Data

Unfortunately, the data extracted from the online ETO database was not in a format conducive to such an analysis. For starters, the variables were not all calculated on the same timelines. One of the main challenges for the program is that individuals come in and out of the

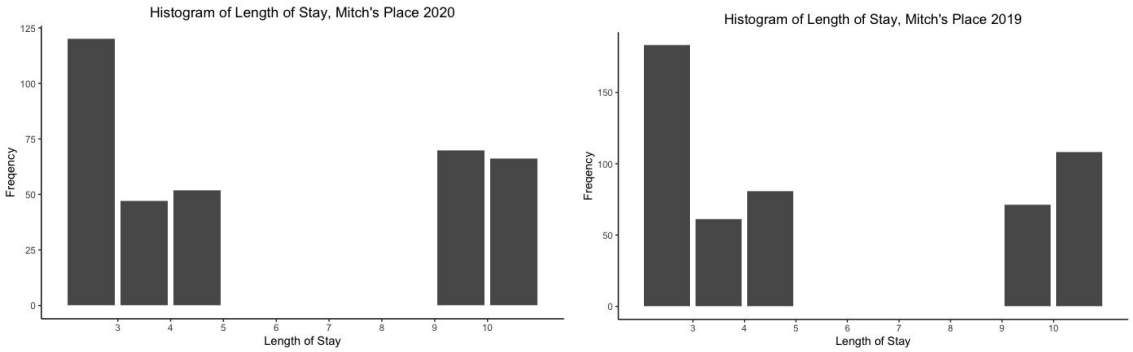
shelter multiple times in a single year or couple of years. Should these individuals be double counted for statistical analysis? Thankfully, each entry did have an *Enrollment ID* that I could use to join entry and exit data. The other blessing was that the data had a column titled *Data Collection Stage* which took the value "1" for data collected on entry and the value "3" for data collected on exit. Using these two variables, I was able to merge most of the data to my *Client* table, which contained *Enrollment ID* and *Personal ID*, along with key demographic characteristics for each client.[1] I now had two rows for each client entrance to the program for both 2019 and 2020, generating a population of 859 entrances and 820 exits for a total of 549 distinct clients with 76 variables for each.

The data suggests that Emmaus serves a predominantly White population that is consistent from year to year, although contains a decent minority of African American clients, around 10%. Here are side-by-side pie charts showing the racial demographics of the population at Mitch's Place for 2019 and 2020:



[1] Out of all of the CSVs, only one couldn't use the simple merge along *Enrollment ID and Data Collection Stage*. *Disabilities* was a file that contained an individual row for each response on entry for a client to questions about mental health, physical disability, and substance abuse, with one column for the type of question, *Disability Type*, and second for the type of response, *Disability Response*. To match these to my complete table, I needed to create a new table for *Client_Disability_Score*, which took the sum of all "Yes" responses to *Disability Response* for each client and then matched them using the *Enrollment ID*. As I will discuss in *Future Directions*, this new variable brough limitations to my model, because it assumes that all disabilities, mental and physical, count equally, an assumption that is clearly false (think of the differences between schizophrenia, alcoholism, and needing a wheelchair for living on one's own).

The data also show that the distribution of stays among Emmaus clients remained relatively constant from 2019 to 2020. It also shows that the distributions are bimodal, where the majority of clients are either staying for a short period of time--a matter of days or weeks--or are staying much longer, with many staying the entire year at the shelter.



Part 4: Methods

After exploring some of these descriptive statistics, I began to work to build a model of successful exit. To do this, I generated a binary variable out of my categorical variable, *Exit to Housing*, which took the value "1" if a client moved into a rental property or transitional housing, and the value "0" otherwise.[2] I then created a table that consisted only of exit data for distinct *Clients* for both 2019 and 2020, a sample of 549 exits. I then converted my categorical *Gender*, *Race*, and *Ethnicity* into factor variables so they could be used in the regression. From this, I constructed two logit models: (1) *Exit to Housing* on *Race*; and (2) *Exit to Housing* on *Age*, *Months Homeless in Past Three Years*, *Times Homeless in Past Three Years, LengthOfStay*, *Gender, DomesticViolenceVictim,* and *Disability Score*.

Part 5: Results

The results of the first regression (Appendix A) showed no significant effect for any racial group on the likelihood of a successful exit from the program. In part, this is likely the

[2] The R file shows a list of all of the destinations that clients exit the program into.

product of a vastly uneven sample with relatively few observations for most of the minority populations. Still, we can tentatively push back on the assumption that Emmaus is failing to treat their minority populations differently from their White populations, a reassuring finding.

The results of the second regression (Appendix B) were much more illuminating with respect to the probability of successful exit for a given client. *Age* was positively correlated, with a coefficient of 0.03, interpretable as each additional year of age increases a client's likelihood of successful exit by 3%. This is most likely a product of older clients having more assets than clients in their 20s. *Times Homeless in Past Three Years* was nearing significance with negative sign. This variable is definitely something to pursue in further regressions, because it has been documented that each case of homelessness recidivism has drastic consequences on an individual's ability to gain sustainable housing and employment. But, as I said in the *Introduction*, the most insightful variable for successful exit was *Domestic Violence Victim*. Victims of domestic violence successfully exited only 3% of the time, compared with a 14% success rate for non-victims. And when controlling for other variables, the odds ratio shows us that being a victim reduces your chances of successful exit by 85%, with a p-value of 0.008.

My best hypothesis for this result is that domestic violence victims have two traits that are preventing them from moving into housing. First, 75% of domestic violence victims are either females or transgender females. Likely, many of these individuals have chosen to leave a household with little or no resources at all to escape domestic violence. They will have a very hard time paying for housing in the next year after arriving at the shelter. Second, the effects of domestic violence may be causing significant trauma that hinders their ability to function on their own.

Part 6: Future Directions

The preliminary research I conducted leaves many possibilities for future directions. In the coming weeks, I will be actively pursuing this link between domestic violence and successful exit from programs at Emmaus. If, after more analysis, the link seems robust, there are multiple policy avenues that Emmaus could go down, from treatment and therapy to a prioritization in the aid decisions and a greater weight for history of domestic violence.

I will continue to work with this data by bringing records on all distinct clients for the period 2010-2020 to generate a larger sample for analysis. Another step will be to separate out the *Disability Score* into its component parts and observe whether this is any more predictive of divergent outcomes in placement. I also did not use any of the data on healthcare, assets, and employment history. These will take a good chunk of time, because they need to be aggregated in such a way that they are interpretable as estimates in the model, but these are certainly important inputs to the model that should be included. Together, each of these variables can be used as more precise inputs for a similar logistic regression.

Appendix A:

MODEL FIT:

$\chi^2(4) = 8.47$, $p = 0.08$

Pseudo-R^2 (Cragg-Uhler) = 0.03

Pseudo-R^2 (McFadden) = 0.02

AIC = 404.89, *BIC* = 426.44

Standard errors: MLE

	Est.	S.E.	z val.	p
(Intercept)	-16.57	2399.54	-0.01	0.99
race_fctNative American	-0.00	2545.10	-0.00	1.00
race_fctWhite	14.51	2399.54	0.01	1.00
race_fctAfrican American	14.93	2399.54	0.01	1.00
race_fctHawaiian Islander	16.57	2399.54	0.01	0.99

Appendix B:

MODEL FIT:

$\chi^2(8) = 31.88$, $p = 0.00$

Pseudo-R^2 (Cragg-Uhler) = 0.12

Pseudo-R^2 (McFadden) = 0.09

AIC = 343.01, *BIC* = 380.79

Standard errors: MLE

	Est.	S.E.	z val.	p
(Intercept)	-11.97	1013.27	-0.01	0.99
age	0.03	0.01	2.61	0.01
MonthsHomelessPastThreeYears	-0.01	0.03	-0.40	0.69
TimesHomelessPastThreeYears	-0.30	0.17	-1.76	0.08
LengthOfStay	0.03	0.04	0.77	0.44
Gender_fctMale	11.00	1013.26	0.01	0.99
Gender_fctFemale	11.50	1013.26	0.01	0.99
DomesticViolenceVictim	-1.85	0.64	-2.89	0.00
client_disability_score	-0.22	0.19	-1.15	0.25