Project Report

R against the Machine

4/27/2019

Problem Statement and Background

Since 1975, the United States government has welcomed over 3 million refugees for resettlement from all over the world (UNHCR, 2018). Recent cuts to refugee resettlement quotas have sparked debate over the program's national security implications and the ability of refugees to integrate into their host communities. We found that this conversation is often driven by emotional and ideological claims rather than evidence. Thus, we ask: what does the most recent available data tell us about refugee integration outcomes in the United States?

We seek to answer this question by looking at the Annual Survey of Refugees 2016 (ASR), the most recent, nationally representative survey of refugees who were resettled in the US between 2011 and 2015. The survey was carried out by the Office of Refugee Resettlement at the U.S. Department of Health and Human Services (HHS) and offers a window into respondents' first five years in the US and their progress towards learning English, participating in the workforce, and establishing permanent residence. The dataset includes information on 1,500 households and more than 4,000 individuals, and is available as a STATA database to researchers from accredited universities.

We recognize that survey data can come with problems in terms of its reliability and external validity. However, for this particularly policy area, there is a huge gap in capturing refugee data with very few agencies collecting information (other than ASR, the only other data sources are the Census Bureau data: American Community Survey (ACS) and the New Immigrant Survey).

Data science tools can equip governments with tools to optimize their institutional procedures. Scientists at Stanford University have developed a data-driven algorithm that uses supervised machine learning to discover synergies between refugees and resettlement sites, leading to 40 to 70% improvements, on average, in refugees' employment outcomes relative to current assignment practices.[1]

[1]Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. Science, 359(6373), 325-329.

Methods

Exploratory Data Analysis (EDA) allowed us to open-mindedly explore and gain significant insight into the survey results and answer our research question. We were also drawn by the potential that data visualization has to communicate our findings. We had considered creating a linear probability or machine learning model to predict the likelihood of a refugee household being successful at resettling. This approach was discarded because of data availability, the complexity of defining resettlement in strictly quantitative terms, and seeking to maximize our project's reach beyond specialized audiences.

We summarize the results of our EDA with a dashboard that enables users to navigate the various dimensions of refugee integration. Dashboards have gained significant popularity as a tool to communicate information on news outlets and research institutions.

As a step prior to EDA, we did basic data cleaning and wrangling tasks to ensure the data could be rendered into visualizations. We then coded the user interface.

Tools

We relied on the tidyverse and haven packages for data manipulation, particularly for importing and cleaning STATA data into R. We found difficulties understanding the nature of haven-labelled items that resulted from importing .dta, and discovered differences in coding results when we used the "readstata13" package. We ended up using haven for consistency due to an error of incompatible types (haven_labelled/charcter) when joining the data for the map.

We employed ggplot and ggplot2 for our visualizations and relied heavily on shiny, shinydashboard, and shinythemes to be able to code the actual dashboard. To host our product online, we created a free account on Shinyapps. Building a Shiny app added an additional dimension to our data exploration, since we had to use design thinking and not only data science skills. This project also made us realize that while R can be a powerful tool for map visualization and dashboard building, this is perhaps not the language's competitive advantage. We wonder whether other (although more expensive) resources such as Tableau would be better suited for these tasks.

Through the coding process we used Git and the Github platform for version control. While it overall served its purpose, we often found that it presented complications when edits were being made in real time.

Results

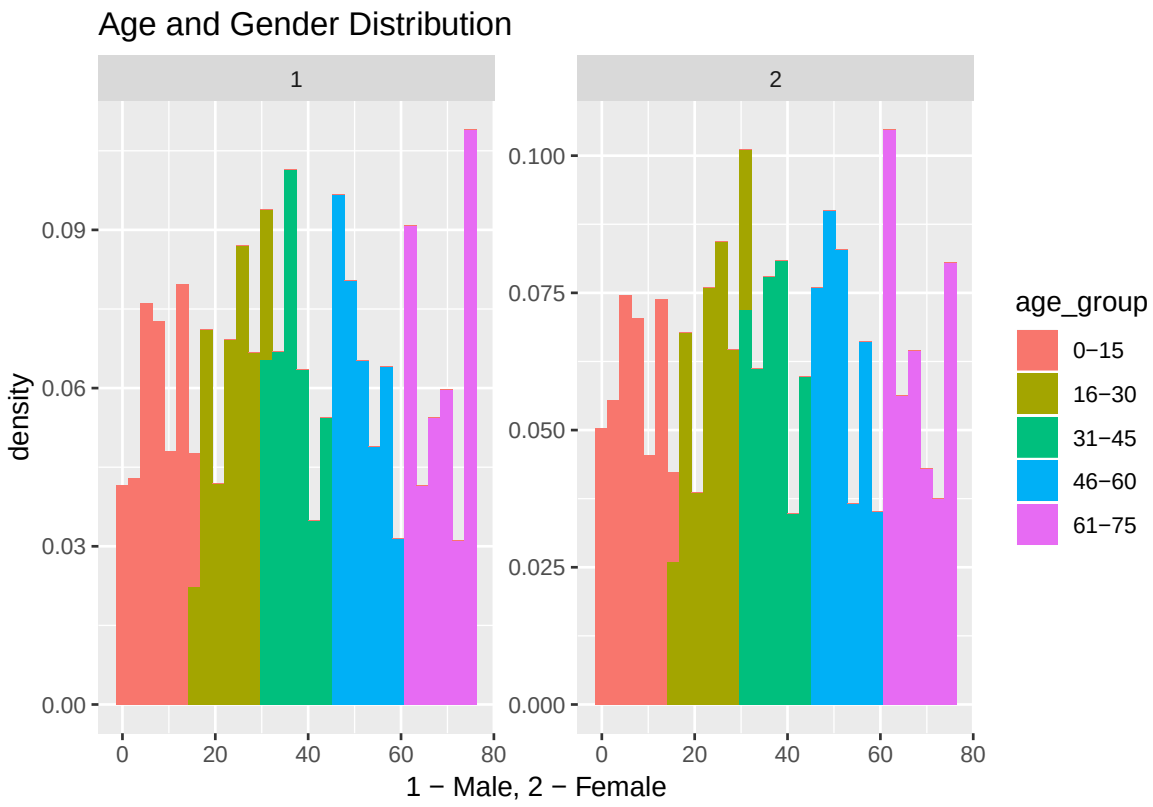
We have compiled the results of our EDA analysis into a dashboard deployed and accessible via ShinyApps. The target audience of the dashboard is the general public interested in discovering the evidence behind refugee outcomes.

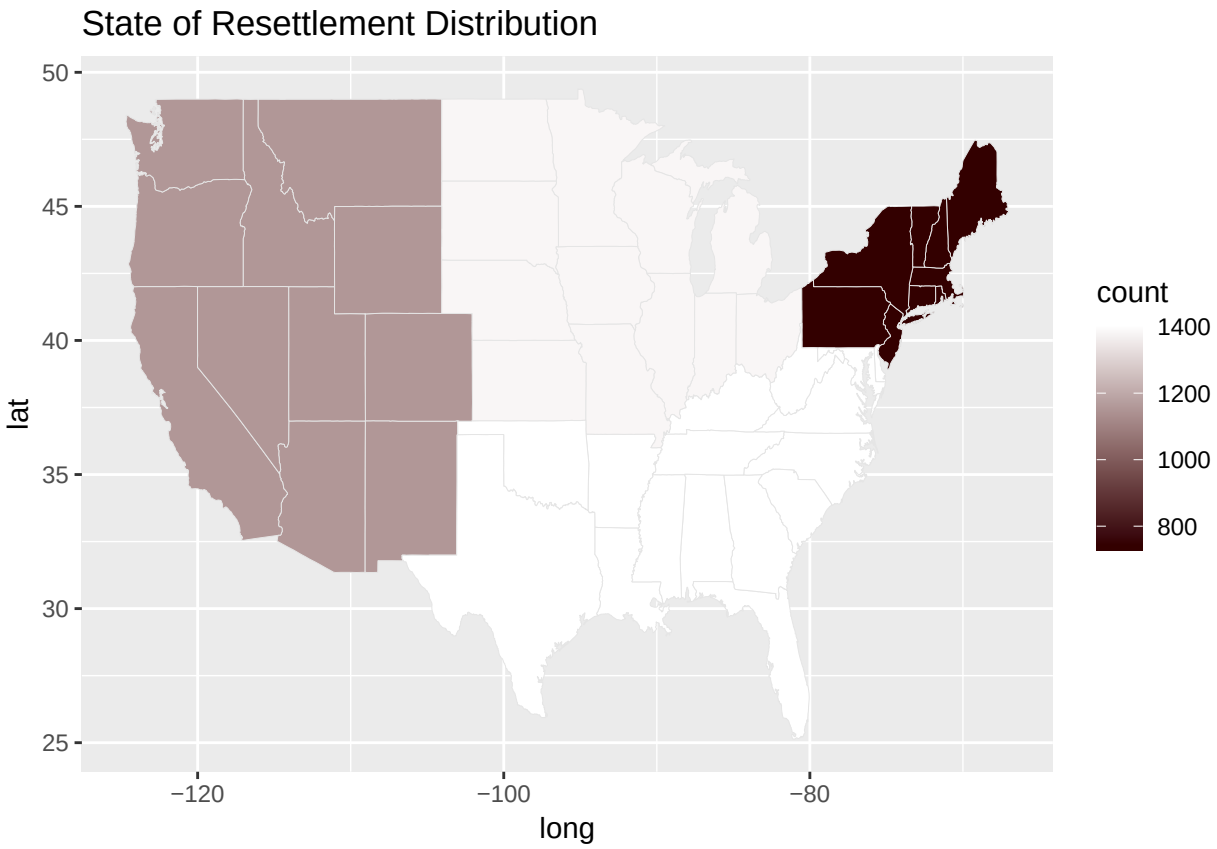
We have decided to articulate our discussion of results around three questions that capture the main doubts we had going into the process of data exploration:

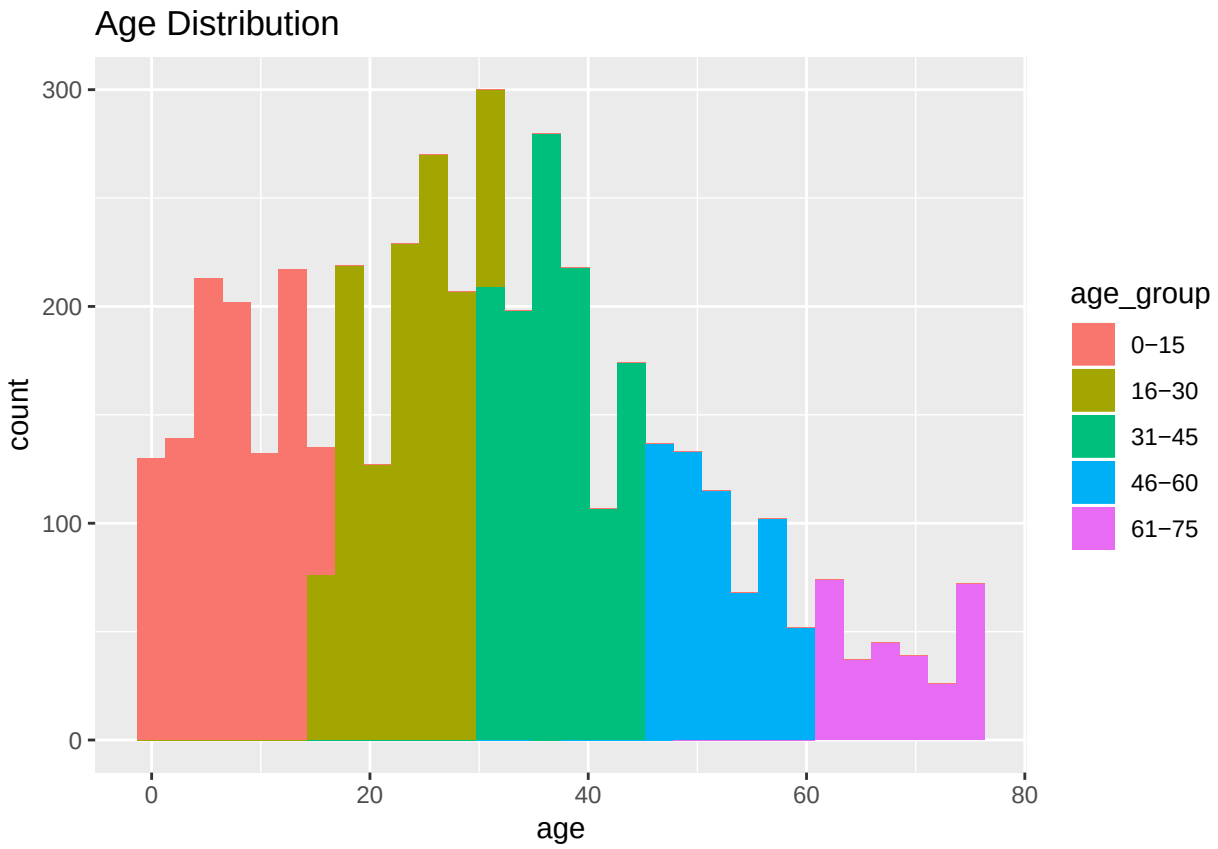
1. Who exactly are the refugees and where are they?

We reviewed the demographic composition of the refugee survey data; our analysis found that:

- The largest group of refugees in the sample were born in Iraq, totaling 1567. They were closely followed by Bhutan (611) and 513 for Burma/Myanmar. We can notice the absence of large numbers of Syrian refugees, perhaps indicating that their arrival ocurred after the sample was taken.
- Refugees in age groups 16-30 and 31-45 represent the majority of the age distribution. When we factor in gender, we find that there is no significant difference between male and female refugees.
- The South and the Northern central region of the United States seem to have very similar resettlement counts; the Northeast has fewer number of refugees, compared to the other regions.


Age and Gender Distribution
1
2
density
0.09
0.06
0.03
0.00
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
1 − Male, 2 − Female
age_group
0−15
16−30
31−45
46−60
61−75


State of Resettlement Distribution
lat
50
45
40
35
30
25
−120
−100
−80
long
count
1400
1200
1000
800

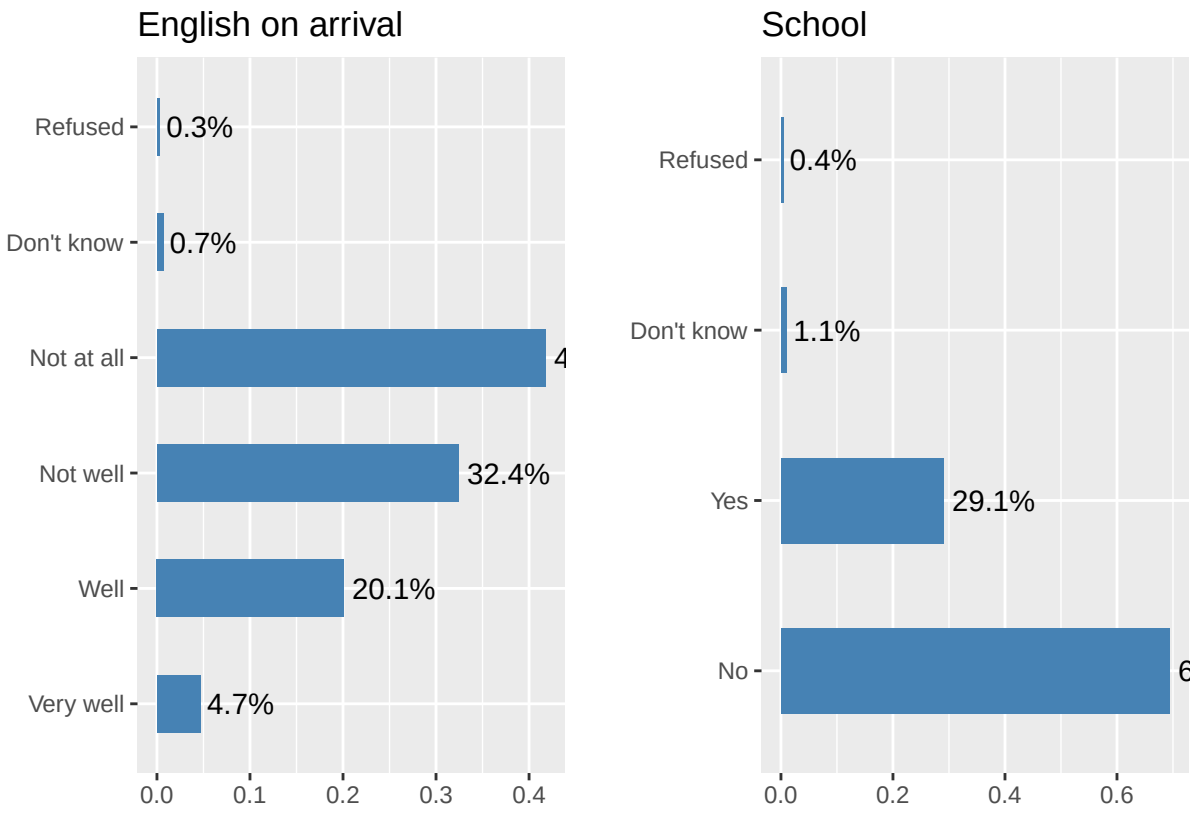


2.What do we know about the education and employment outcomes of refugees?

A common question surrounding refugees is their ability to make decisions about their employment and education that can increase their financial and human capital. Our analysis found that:

- 1 in every 4 refugees said they spoke English well or very well upon arrival to the U.S. However, close to 70% of them said they spoke little to no English, which naturally presents challenges for integration efforts. Low English skills create entry barriers to schooling, employment, housing, and more.
- 25% of refugees sampled had obtained a high school diploma prior to arrival, with a similar proportion reporting they had completed primary school. Only 1 in 10 said they held a university degree. While this speaks of the presence of a largely low-skilled population, these figures do not reflect work experience or training received.
- The second graph shows that the majority of refugees are not in school. A deeper analysis should examine the breakdown of this proportion by age and gender. For those who were in school, close to 50% are seeking to obtain a high school degree or equivalent.

- Only 21% of refugees reported having worked since their arrival to the United States. This speaks of the presence of barriers – either in terms of their own skills or the ability of the system to absorb them as workers – that prevent them from accessing job opportunities.
- As expected, most refugees are employed by the private sector (63%). It is interesting to notice that close to 10% of them have joined the public sector, either at the federal, state, and local level.



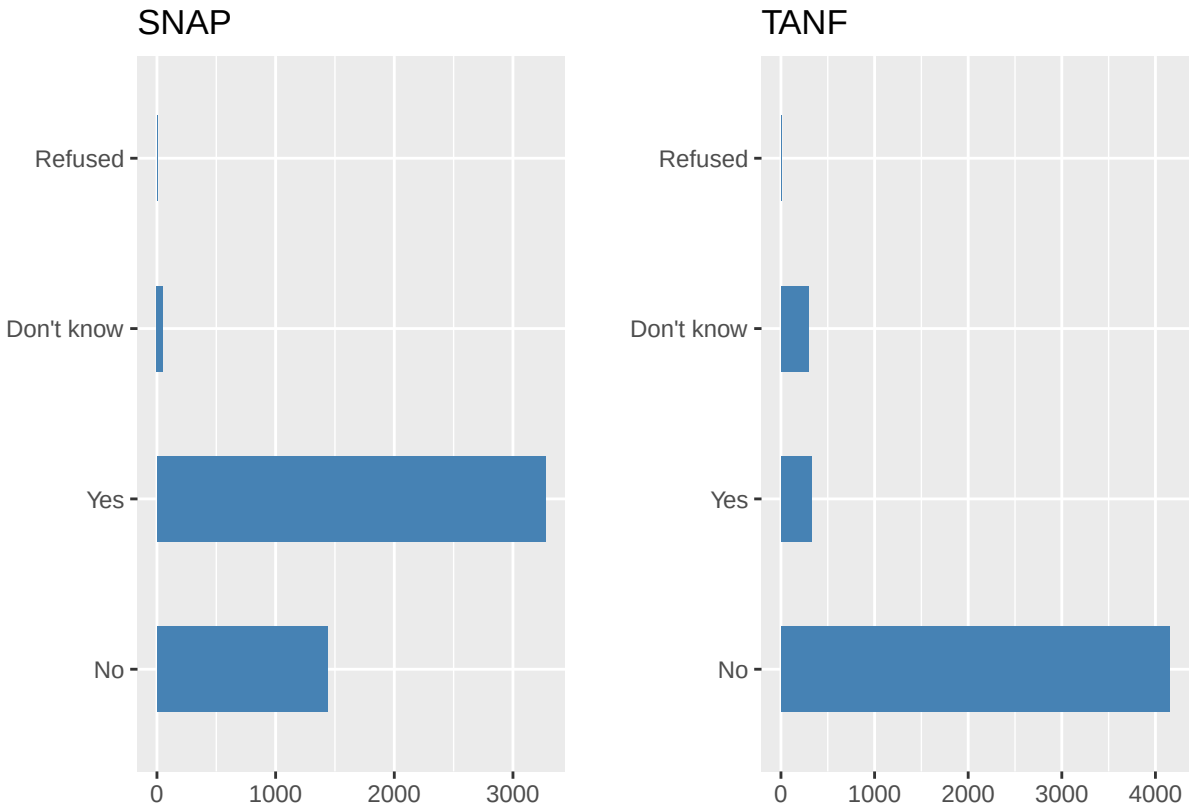
3. Do refugees represent a burden on the tax system (through benefits)?

The policy debate in the United States emphasizes the economic burden refugees put on the tax system. We created visualizations showing how many refugees in our sample have chosen to use government benefits in the past 12 months. We found that:

- Few refugees, generally less than 5% except for SSI, receive or take advantage of cash assistance programs.
- The "I don't know" category is more pronounced for cash asssistance programs, alluding to limitations of the data. It is possible that survey takers knew about assistance for SSI and SNAP, but not about

cash assistance programs, such as TANF and RCA.

- Generally, cash assistance programs also have eligibility barriers than non-cash assistance programs; e.g. RCA is only available for the first five months, and TANF is limited to a cumulative lifetime of up to 5 years.
- While acceptance of SSI is more pronounced than the other cash assistance program, those that do not receive SSI far outweight those that do.
- SNAP is an outlier, where more refugees indicated they receive food stamps than those that do not. This is likely the case because income elibigilty thresholds are much lower and are further available for households with children.



Measuring success

For the purpose of this project, we measured success by asking: were we able to build a platform that allowed us (and other users) to explore data related to refugee resettlement? Furthermore, was the platform effective in leading us to answers to the initial research question(s) we set? We have determined our project was successful because it managed to (i) identify relevant, high-quality data from a trustable source; (ii) apply

the data science tools we learned in the course to find answers; and (iii) communicate the results of our finding to a general audience using an appropriate platform (dashboards).

At the same time, we recognize the limitations of our approach and our project as a whole. For one, EDA is an open-ended process which can arguably lead to different conclusions depending on the perspective (and personal agenda) of the researchers. In this sense, more statistically driven approaches could be perceived as more rigurous in reaching conclusions backed by evidence.

Moreover, we relied on a single dataset which contained information on a single cohort of refugees, thus pushing us to think about the external validity of our conclusions. This means questioning whether the refugee resettlement outcomes have changed substantially for refugees who arrived after 2015 (last year in our survey).

Thinking about next steps, we believe our project could be a starting point for research that seeks to compile data from the U.S. Census and NGOs into a larger "refugee resettlement outcomes" dataset that can help build the knowledge base on this important policy issue.