Analyzing US Immigration

DATA 450 Capstone

Max Bilyk

February 26, 2024

1 Introduction

In an era characterized by unprecedented global mobility, the study of immigration patterns is crucial for informed decision-making and policy formulation. The United States, as a melting pot of diverse cultures and backgrounds, experiences a continuous influx of immigrants, shaping the nation's demographics and socio-economic landscape. The proposed project aims to delve deep into the intricate web of U.S. immigration through a comprehensive data science approach.

The objective of this project is to harness the power of data analytics and machine learning to analyze extensive datasets related to U.S. immigration. By leveraging advanced statistical techniques and predictive modeling, we aim to unravel hidden patterns, trends, and insights that can inform policymakers, researchers, and the public alike. This project seeks not only to provide a snapshot of current immigration dynamics but also to forecast future trends, facilitating proactive decision-making.

By undertaking this comprehensive data science project, I aim to contribute valuable insights that can guide evidence-based policymaking, foster a deeper understanding of the diverse fabric of U.S. society, and promote informed public discourse on immigration-related matters. The outcomes of this project have the potential to shape policies that align with the evolving needs and aspirations of both current and future generations of immigrants in the United States.

2 Datasets

The first dataset to be used will be a dataset that has the annual number of new legal permanent residents by country of birth, from the years 1999-2022. A single row represents a given country, and a column represnts the number of new legal immigrants from that given year.

The second dataset to be used will be a dataset that holds information on the immigrant population by state from 1990,2000,2010,2019,2021,2022. Each row represents a state in the U.S. in one of the given years, and the columns are features like, total population, immigrant population and share of total state population in those given years.

The third dataset to be used will be a dataset that holds information on U.S. annual refugee resettlement ceiling and annual number of admitted refigees from the years 1975-2024. Each row represents a given year, and that years annual resettlement ceiling and the number of admitted refugees.

The fourth dataset to be used will be a dataset that holds information on the National and State estimates of the unnathorized immigrant population from 2015 to 2019 (totals, not year over year). Each row represents a given state, their estimated number of unathorized immigrants and the state share of unauthorized immigrant population.

The fifth dataset to be used will be a dataset that holds information on the number of unaccompanied children released to sponsors by state from fiscal years 2014-2023. Each row represents a given state, and the columns represnt the years, and the associated values are the number of unaccompanied children.

The sixth dataset to be used will be a dataset that holds information on the educational attainment of recently arrived immigrants (ages 25 and over) by country of birth as of 2022. A single row represents a given country, and the different columns are Number of Adults, Less than 9th Grade, 9th-12th Grade,High School Diploma or Equivalent, Some College or Associate's Degree, Bachelor's Degree or Higher. The corresponding values are the percentages of the total number of adults.

3 Data Acquisition and Processing

All of the datasets come from the migration policy institute website."The nonpartisan Migration Policy Institute is an independent, nonpartisan think tank that seeks to improve immigration and integration policies through authoritative research and analysis, opportunities for learning and dialogue, and the development of new ideas to address complex policy questions." This organization has built a website that houses dozens of datasets relating to migration around the world. I selected a subset of these datasets that all pertain to answering questions I had about U.S. immigration. All of the datasets are in xlsx format with headings and pictures in all of the sheets. In order to pull these datasets into python, I will be converting them to csv's by essentially copy and pasting just the data into new sheets and saving them as csv's. The data is thoroughly recorded, so their will not be a ton of work imputing missing values, however the formating and naming conventions of some of these tables is not clean and I will likely be tweaking it in order to benefit the visuals I produce.

4 Research Questions and Methodology

1. How many green cards are issued annually and to what countries are most of the green cards issued to? I plan to answer this question by pulling the first dataset, and creating a line plot of green card issues year over year to analyze the pattern. Then I will total all the countries over all the given years, and come up with the top 5 countries with the most green cards issued to it's citizens.
2. What are the trends of state immigration? For this question, I will build a chloropleth where the color of each state will correspond with the state's immigrant share of total state population in 2022.
3. Does the U.S. have the resources to house refugees (Not Immigrants)? To answer this question I will pull the third dataset and build a line plot with two lines. One line being the number of admitted refugees and the other being the annual ceiling the U.S. government claims that year. It will be interesting to see if the admitted line ever surpasses the annual ceiling, and additional research on what was going on in the world at any of those points would lead to interesting insight.
4. Where are most unauthorized immigrants going when they make it to the U.S.? For this question I will pull the fourth dataset and build a chloropleth map where each state will be colored by it's estimated number of unauthorized immigrants.
5. Do unaccompied children get released to certain states at higher rates than others? For this question I will pull Dataset 5 and 4 and join them together. By doing this I can create a ratio, unnacompanied children over total estimated unauthorized immigrants by state from 2014-2023. It would be interesting to build a bar chart of the top 3 and bottom 3 states using the ratio I will create.
6. Are their differences in educational attainment of recently arrived immigrants based on what countries they come from? For this question I will pull dataset 6 and focus in on a countries immigrants attainment of a bachelor's degrees or higher. For this question, I will build a chloropleth world map, and color each country by their percentage of immigrants that attain a bachelor's degree or higher in the U.S. This will make it possible to quickly compare a large range of countries and make it easy to point out areas that find ample oppurtunity in the U.S. vs those who do not.

5 Work plan

Week 4 (2/12 - 2/18):

- Project Proposal Planning (10+ hours)

Week 5 (2/19 - 2/25):

- Project Proposal Planning (10+ hours)

Week 6 (2/26 - 3/3):

- Data cleaning of all datasets (4 hours)
- Answer Question 1 (3 hours)
- Answer Question 2 (3 hours)

Week 7 (3/4 - 3/10):

- Answer Question 3 (3 hours)
- Answer Question 4 (3 hours)
- Answer Question 5 (3 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).*

- Answer Question 6 (3 hours)
- Poster prep (4 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)

Week 10 (4/1 - 4/7):

- Review code (7 hours)

Week 11 (4/8 - 4/14):

- Review and finalize code (3 hours)
- Begin blog (4 hours)

Week 12 (4/15 - 4/21):

- Finalize Blog (7+ hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.*

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)

6 Links

All Datasets

UCIS

NILC

7 References

U.S. Immigration Trends. (n.d.). migrationpolicy.org. https://www.migrationpolicy.org/programs/data-hub/us-immigration-trends?gad_source=1&gclid=CjwKCAiAivGuBhBEEiwAWiFmYbxFPH7H1xkFCmwWR
NkOL9lygAPqFa7ppdZx1vOxxoC88YQAvD_BwE

Home. (n.d.). USCIS. https://www.uscis.gov/

How Immigration Has Affected the U.S. (n.d.). National Immigration Law Center. https://www.nilc.org/immigration-reform/?utm_campaign=SignUp&utm_source=Ad&utm_medium=Google
pOmABs4HJS07LOhscVOWphoCkIAQAvD_BwE