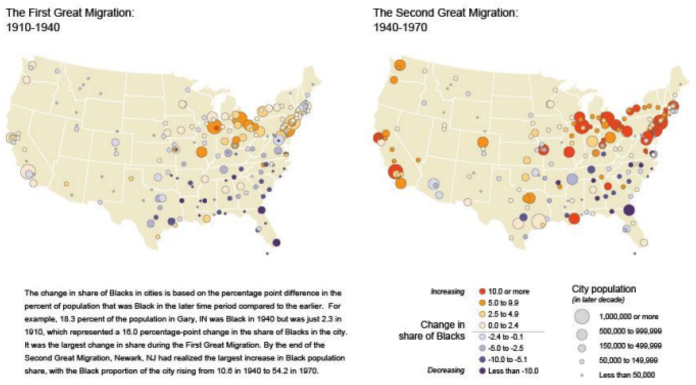
Immigration Tracking Over Time in the United States

Contact	Names: Arnaud Harmange, Luke Staib Emails: arnaudh@bu.edu, ljstaib@bu.edu Cell Phone: 978-809-1682, 516-408-8668
Organization	N/A
Organization Description	N/A
Project Type	Data Science
Project Description	We would like to create a visualization that shows the origin countries of immigrants coming to the United States over time. We'd also like to use these datasets to create a prediction model. This prediction model might help show where immigrant populations are likely to come from in the future based on past trends. Our intent is that our visualization will allow for an easy to understand history of immigration to the United States. The prediction model that we create could be used by lawmakers, historians, and analysts as a supplemental tool.
Data Sets & Sources	We can use datasets from the US Census Bureau and from the Department of Homeland Security (DHS) to compile necessary immigration data. For census data (*https://www.census.gov/data/datasets.html*), use the filter "Population Estimates" under the "Population" section on the left of the page. Many datasets over the years are available for the kind of data we want to use. For DHS data (*https://www.dhs.gov/immigration-statistics/yearbook)*, choose a year from 1996 to 2020 using the sidebar on the left. Each year contains a set of immigration tables along with a description of each table.
Suggested Steps	1. Utilize the datasets made available by the US Census Bureau and by the Department of Homeland Security. 2. Aggregate all historical datasets and refactor them for use in one or more pandas dataframes.

	3. Use this aggregation to create an animated visualization of historical immigrant population movement to the United States over time. 4. Once completed, the historical data can then be used to train a population movement model, most likely similar to the one utilized in this paper (source below):  The First Great Migration: 1910-1940 The Second Great Migration: 1940-1970 The change in share of Blacks in cities is based on the percentage point difference in the percent of population that was Black in the later time period compared to the earlier. For example, 18.3 percent of the population in Gary, IN was Black in 1940 but was just 2.3 in 1910, which represented a 16.0 percentage-point change in the share of Blacks in the city. It was the largest change in share during the First Great Migration. By the end of the Second Great Migration, Newark, NJ had realized the largest increase in Black population share, with the Black proportion of the city rising from 10.6 in 1940 to 54.2 in 1970. Increasing Change in share of Blacks Decreasing 10.0 or more 5.0 to 9.9 2.5 to 4.9 0.0 to 2.4 -2.4 to -0.1 -5.0 to -2.5 -10.0 to -5.1 Less than -10.0 City population (in later decade) 1,000,000 or more 500,000 to 999,999 150,000 to 499,999 50,000 to 149,999 Less than 50,000 Source: http://snap.stanford.edu/class/cs224w-2015/projects_2015/Analyzing_and_Predicting_Internal_Migration_Patterns_in_the_USA.pdf
Questions to be answered in Analysis	Where have immigrants coming to the United States come from historically? What are current immigration patterns by group in the United States? What are likely trends in immigration looking into the future? Do current immigration trends differ drastically from past immigration trends?
Ideal Output + Final Deliverable	We hope to deliver: 1. An immigrant group prediction model using census and DHS data that can freely be used for any future project. 2. An animated visualization of the past and predicted immigrant origins in the United states. 3. Our implementation of a population immigration model that is reasonably accurate.
Additional Information	N/A

List of Limitations	1. It may be difficult to compile the extensive amount of data from the US Census and the US DHS. 2. We may have to look deeper into whether any data is being shared between the census and DHS in order to avoid duplicate data. This could end up being a difficult and tedious manual step. 3. A great deal of data cleaning will likely need to be done before working with the previously mentioned datasets. 4. DHS data spans from 1996 to 2020 while Census data spans from 1970-2021. Less data in earlier and very recent years (1970s-1980s and 2021-22) could potentially be an issue when creating an accurate model. 5. (New Limitation, Deliverable 1) There are many datasets. We will need to take a look through each relevant dataset to see which ones are best suited for our project.
Answer 1 Key Question: Deliverable 1	Q: What are current immigration patterns by group in the United States? How have immigrants moved throughout the United States historically? A: Please see the graphs in “exploration.ipynb”. We have created basic models for: - Number of people immigrating into the United States per year (1820-2020) - Immigration into the United States by continent (1820-2020) - Ancestry in the United States (2020-2010)
Deliverable 2	- Data from deliverable was used to make better visualizations - More data was aggregated and collected and made into easy to understand visualizations as well - Question answered: Where have immigrants come from to the United States and has this changed drastically over time? - Answer: See visualizations for where immigrants have come from. Most immigrants come from European nations. Trends in immigration at least from 2010-2020 appear to have remained relatively similar year over year. - Refining Project Scope: - We may not be able to create a prediction model as hoped. Gathering the data and making it usable has proven to be more challenging than anticipated. We have recently found a new source of census data that may solve this problem, but we have not had sufficient time to implement changes utilizing this data, so we are not sure yet if this will solve our problem. We hope to figure out

	how to utilize the census data that is in this new format since it seems to include more historical data than we have previously had access to. III. List of Limitations with Data a) Lack of older data, lack of relevant data/missing fields - There is a shortage of data preceding 1990. In much of the Census and DHS data that we have worked with, there is little to no ancestral data. b) Missing country data - Data for many countries is not present in the datasets that we were able to work with - Data for countries that no longer exist such as Yugoslavia, Czechoslovakia, and Austria-Hungary is difficult to use (in, for example, our world map timelapse GIF) c) Older data is unorganized - 1970s, 80s Census data is stored as a plaintext file. Even with a helper file describing the layout, this is still very difficult to work with. IV. Potential Risks of Achieving Project Goal a) Current Limitations with Data - As discussed in Section III, we currently have several limitations with the datasets that we have available to us