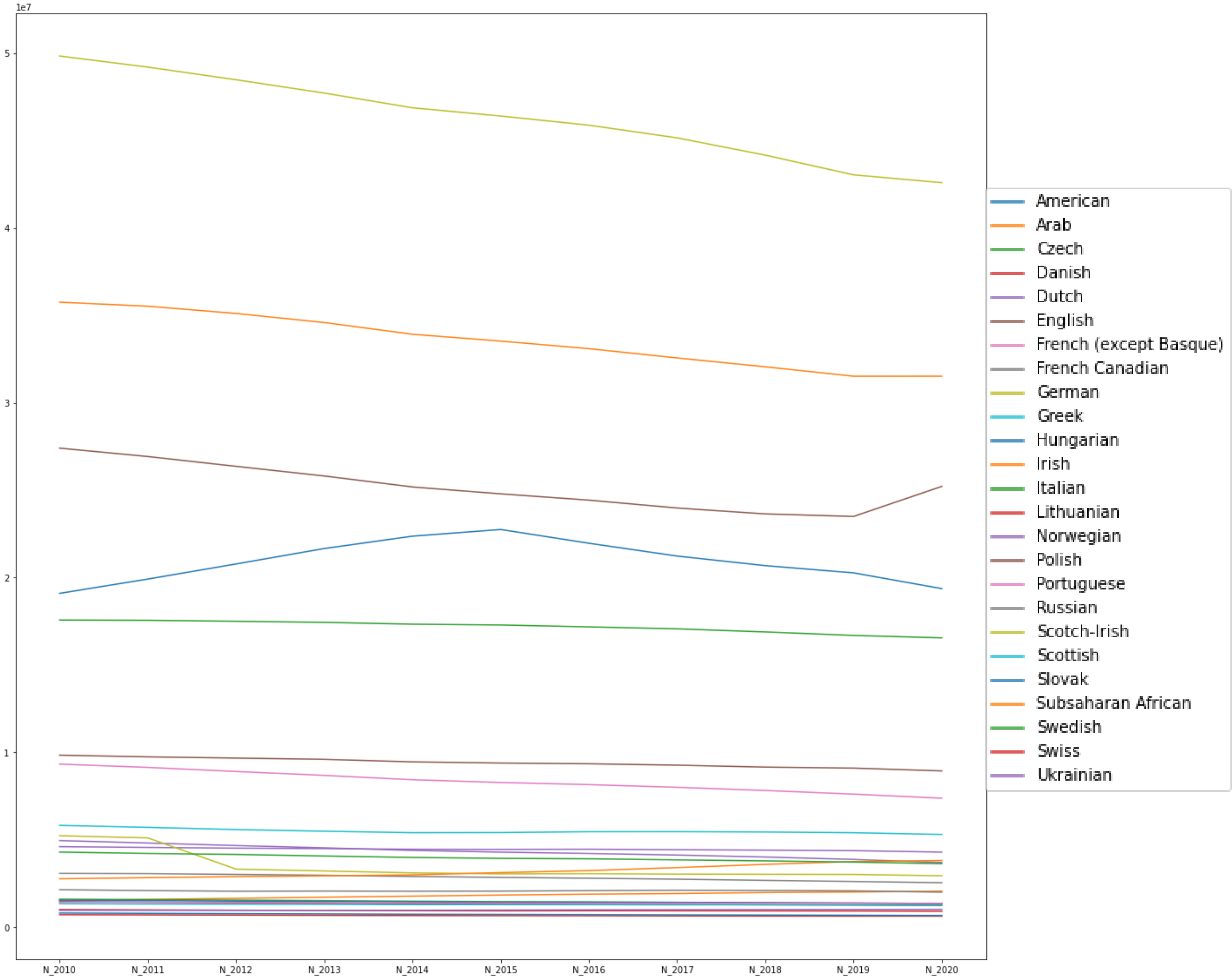
Project Final Report

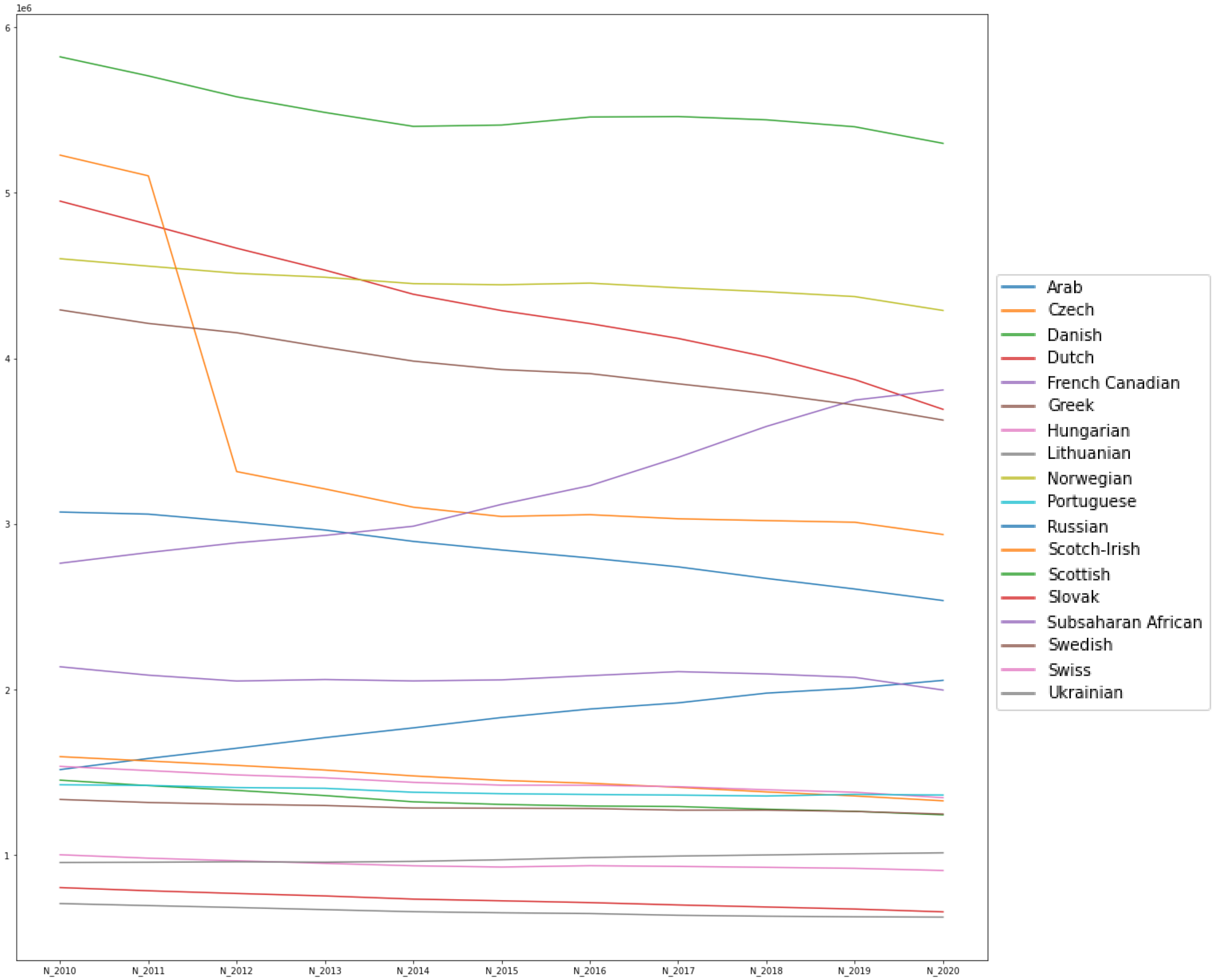
Arnaud Harmange
Luke Staib
04/25/2022

The United States is often portrayed as a “melting pot” of people and cultures coming together from across the globe, and forming unique communities throughout the country. However, for as long as people have immigrated to the United States from other countries, current citizens and residents of the United States often find reasons to complain about incoming populations. In recent years especially, claims of massive immigration and especially claims of immigration from specific countries have been used to polarize politics and create discord amongst people. This project aims to provide definitive answers to the following questions: What are current immigration patterns globally to the United States? Do these current trends differ from historical trends? Based on historical data, how might these trends continue looking forward? In order to answer these questions, this project will make use of data from the US Census and the US Department of Homeland Security (DHS) that details the origin country of individuals obtaining lawful permanent residence status in the United States from 1820 to 2020. The goal is to then utilize this dataset and create a prediction model utilizing techniques learned in class to help predict possible immigration trends in the future.

In working on this project, a significant amount of time has been spent researching the best means of tracking immigration trends over time. The first instinct was to look into Census data, which should have provided historical data on immigration reaching back past the 1900s. However, while data for the time period of 2010 to 2020 was readily available, it was difficult to find the same information for years preceding this time period. Fortunately, the Department of Homeland Security provided an acceptable substitute for the Census data. The Department of Homeland Security maintains a dataset detailing the origin region in the world from which people who obtain legal permanent residence status in the United States come from. This should allow for a more accurate model to be built, since the dataset will be much larger and should provide a better training dataset than the smaller, ten year span that was covered by the census data.

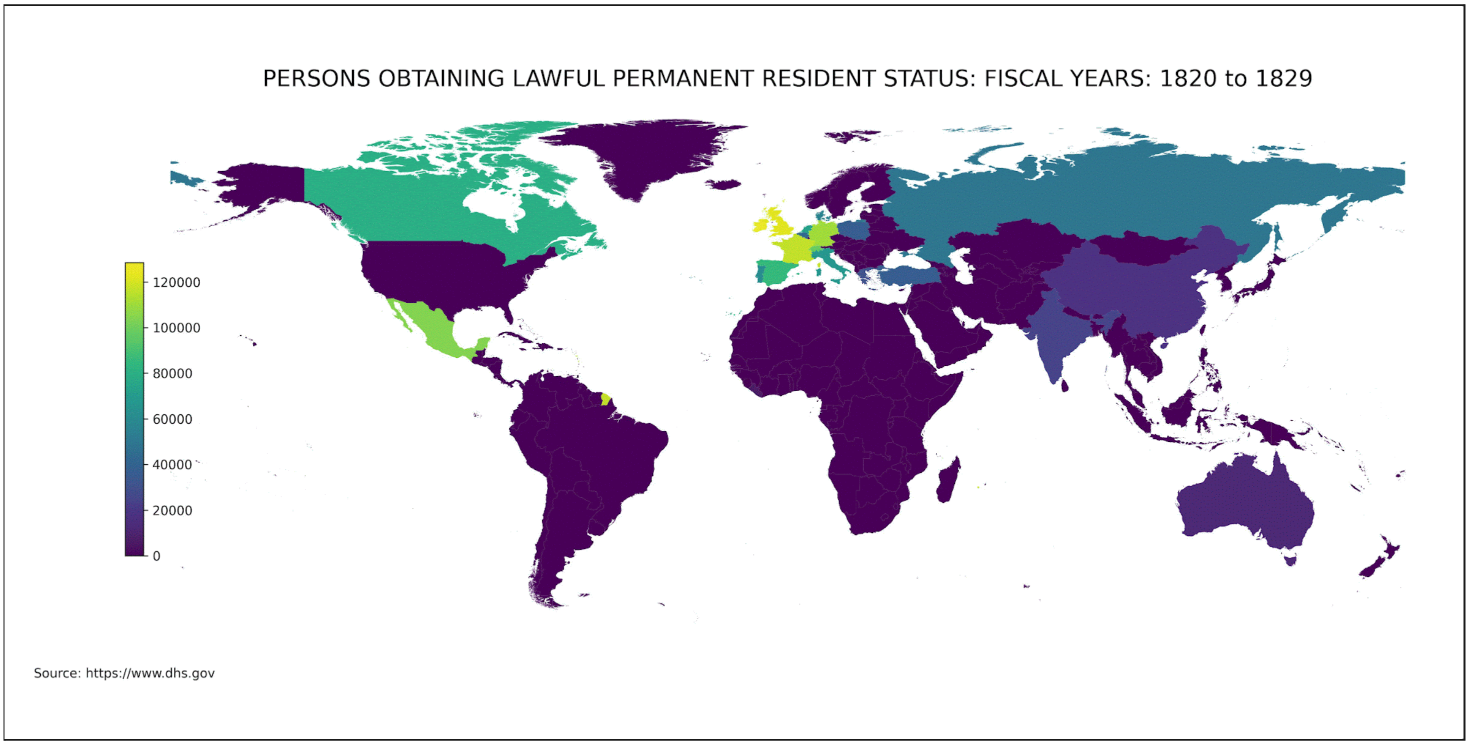
While the DHS data is more comprehensive from a time perspective, it does lack the resolution that the census data has. The census data provides a much more detailed picture of immigrants’ origin, since instead of simply breaking down the immigrant origins by region, the census data breaks them down by country. Despite the census data not covering the same time span as the DHS data, because of its more detailed nature, it can still be used to gain insight into immigration trends from specific countries in the last ten years. Below is a graph of the Immigration trends from the period of 2010 to 2020 according to US census data gathered in that time. Following the first figure is a second version of the same graph, but some of the countries have been removed so that the countries with smaller immigration numbers have more visibility and trends can better be analyzed. The X axis intervals are each year from 2010 to 2020, and the Y axis is the number of immigrants in millions.


1e7
5
4
3
2
1
0
N_2010
N_2011
N_2012
N_2013
N_2014
N_2015
N_2016
N_2017
N_2018
N_2019
N_2020
American
Arab
Czech
Danish
Dutch
English
French (except Basque)
French Canadian
German
Greek
Hungarian
Irish
Italian
Lithuanian
Norwegian
Polish
Portuguese
Russian
Scotch-Irish
Scottish
Slovak
Subsaharan African
Swedish
Swiss
Ukrainian

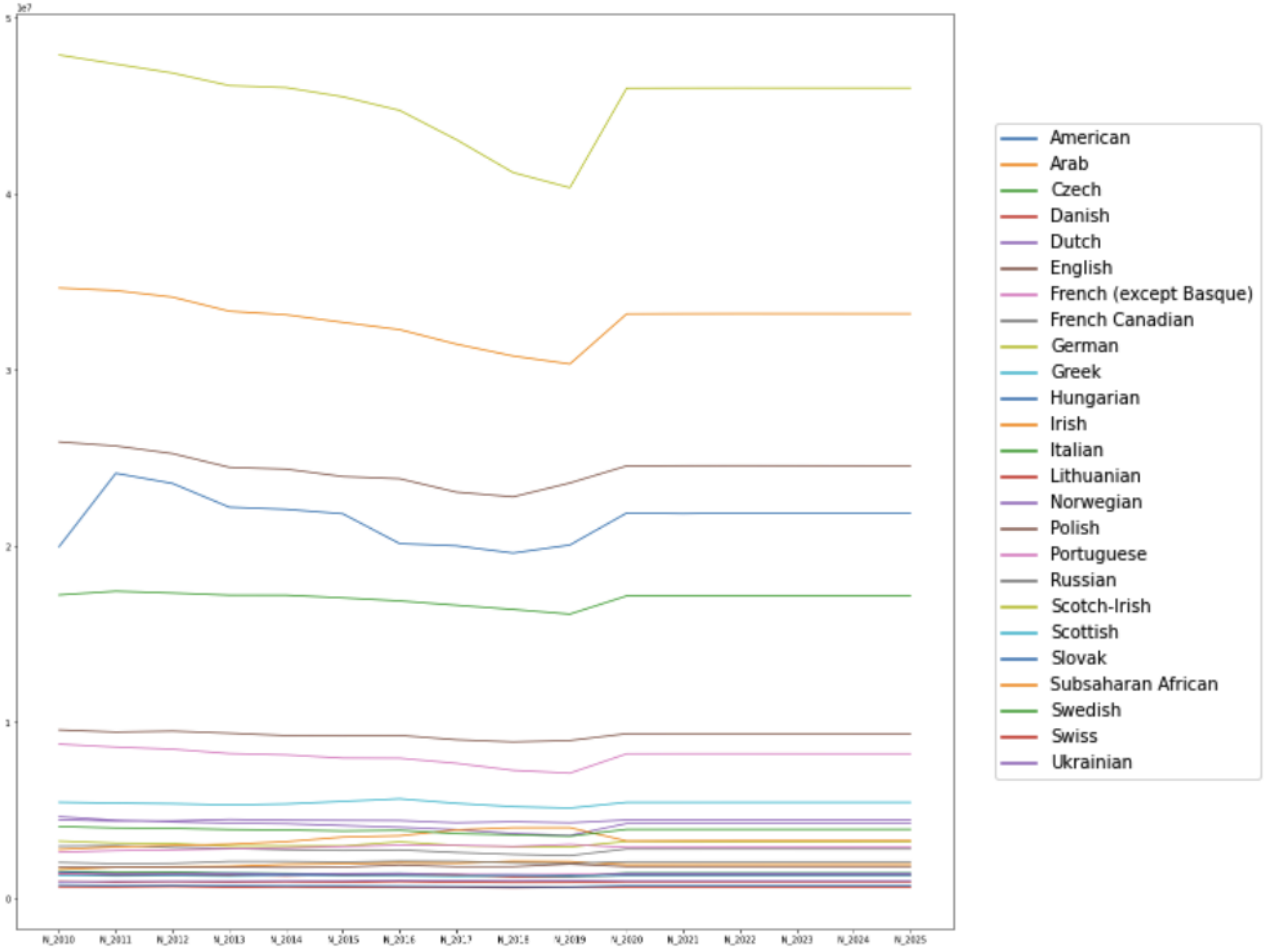


These figures show that immigration to the United States has not seen a massive increase, but rather immigration from most countries has remained fairly constant in the last ten years aside from some exceptions such as immigration from the Scotch-Irish dropping significantly, and immigration from Arabic and sub saharan African countries increasing. These clear visualizations are helpful in understanding the current trends in immigration to the United States, but historic trends can provide further context to this picture. There appears to be a means of obtaining census data like the data displayed above for years dating back to about 1900, but unfortunately the format in which it is made available made it difficult to work with, and ultimately there was not enough time for the data to be aggregated, parsed, and formatted to work for this particular project. In the future, it would be interesting to revisit these same goals from this project, but aim to utilize the historical census data and see if any notable changes in immigration from certain countries are more apparent, and whether this data would improve the accuracy of the prediction model.

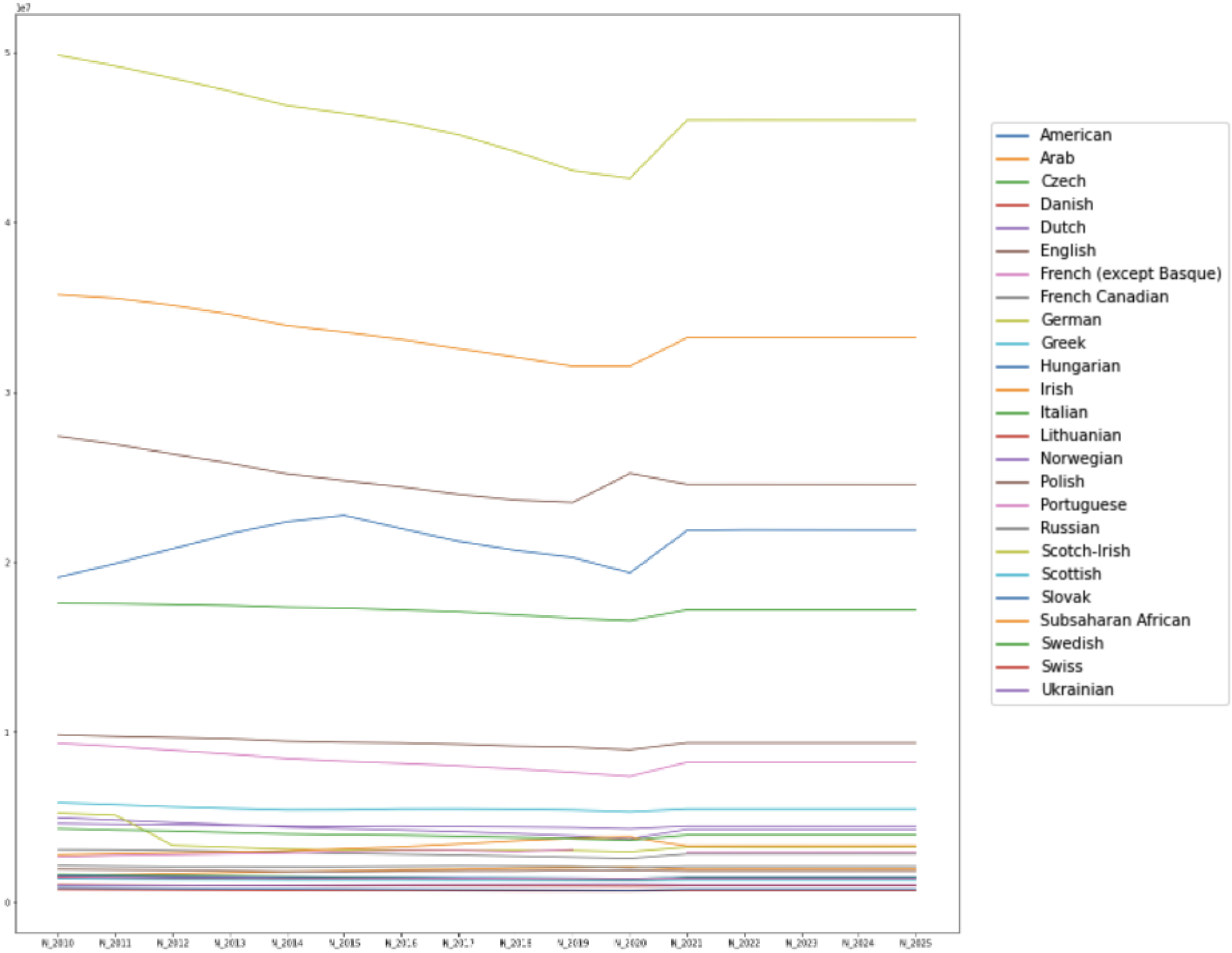
In order to see the trends in immigration over a longer period of time, the DHS data was used. This allows for a visualization of immigration by region to the United States to be created, which should help understand the trends in immigration over time from 1820 to present day. The visualization below was created using geopandas to create the individual images, and PIL to transform the multitude of images into an animated GIF for a more intuitive and engaging visualization. The separate images can be viewed in the project repository along with the animated GIF. The animation shows darker regions around the globe indicating fewer people from those regions acquired legal permanent residence status, and lighter areas on the map indicate regions from which more people acquired legal permanent residence status in the United States. By viewing the animation, it is clear that immigration to the United States over time has changed, and while in earlier years immigration seems to come almost entirely from European countries, Canada, and Mexico, as time progresses, these patterns change. While immigration from European countries and Canada continues to remain relatively high, Immigration from Mexico drops and immigration from China begins to rise. Over time, immigration from Russia, Australia, and some South American countries begins to increase as well. For some time, it appears that Canada, Europe, and Russia remain the largest contributors to United States immigration, while immigration from South America and Mexico remain stable. Over time, interestingly, Immigration becomes less concentrated in certain regions and is much more global, especially from 1960 onwards. From this point on, immigration from all over South America is higher, Immigration from India, China, and Middle Eastern countries increases. Finally, in 1990 to 2020 there is an increase in immigrants from countries in the African continent such as Egypt, South Africa, Ethiopia, and Morocco.


PERSONS OBTAINING LAWFUL PERMANENT RESIDENT STATUS: FISCAL YEARS: 1820 to 1829

Source: https://www.dhs.gov



PF1: Prediction Model using one year ancestry data



PF2: Prediction Model using five year ancestry data

The final component of this project is to create a prediction model that allows for the projection of immigration trends looking into the future. In order to achieve this, the one year and five year ancestry datasets were utilized. Since this immigration is considered time-series data, it made sense to choose a prediction model that is commonly used to predict things like stock market trends. Both Auto Regressive Integrated Moving Average (ARIMA) and Seasonal Auto Regressive Integrated Moving Average (SARIMA) are typically good choices for this kind of data. The key difference between the two prediction models is that SARIMA is often more accurate in predicting trends that follow a broader seasonal pattern over time. After doing some testing and parameter optimization on both models, however, it was clear that ARIMA was better suited for predicting immigration trends. After determining that ARIMA would be used, training and testing datasets were created, and the prediction model on both datasets was finalized. The results of the prediction models can be seen in figures PF1 and PF2 above. While the model does successfully predict trends looking into the future, there were some somewhat disappointing outcomes. The first is that regardless of parameter setting and optimization, it was incredibly difficult to get a result that returned a low root mean square error (RMSE) for any of the trends that were predicted. Second is that for some reason, many of the trends appear to have a very distinct “jump” transition from the real data to the predicted values. It is unknown why this occurs in this particular case, but it should be noted that the predicted trend that follows the jump does appear to be more reasonable and accurate.

Although neither of the prediction models appears to be incredibly accurate, the prediction model using the five year dataset did appear to produce a better model overall, with marginally lower RMSE values on the whole. This is most noticeable on the graph, where the jump from data to predicted trend (which occurs at year 2020 on this graph, rather than at year 2019 in PF1) is less jarring and more reasonable. There was also an attempt to utilize the much larger dataset that was used to create the animated map visualization, but there were issues with that approach. Although it is possible to use that dataset, there was simply not enough time to work with that quantity of data and its somewhat odd formatting to produce a model. The separation and then creation of predicted trends turned out to take a significant amount of time, and it would have taken easily another day to create a usable model from the larger dataset. Perhaps this is something that could be explored in the future. Another note about the prediction model is that the model does not factor into its predictions the other trends relative to the trend being predicted. This would mean that the model would be able to perhaps detect and factor into its predictions relationships between immigration trends from various countries. If it were possible to implement, this may have potentially greatly increased the accuracy of our model. Unfortunately this was not able to be implemented in our model.

This project attempted to answer several questions detailed in our project proposal. What are current immigration patterns globally to the United States? Do these current trends differ from historical trends? Based on historical data, how might these trends continue looking forward? On the whole, these questions have been answered. The prediction model that was created as a result of all this data collection does in fact produce trends that can be examined and compared to the real trends that take place in coming years. It will be interesting to see how closely or not closely these trends follow. In the end, this prediction model could never be entirely as accurate as hoped, since the factors that affect immigration are incredibly complex and cannot realistically be replicated by a simple model. That being said, the model should not be regarded as not useful. This model looks at historical trends and attempts to project them into the future. This can be helpful since any significant departures from the predicted trend may indicate a significant change in policy, legislation, or geopolitical event that leads to a significant change in immigration patterns worldwide.

TODO:

- All:
    - Visualizations, Results
        - Probably the bulk of our report
    - Reference README in report (refer reader to data, code, how to reproduce results, how to run, etc)

Mostly Done:

-
    - Data
        - Could just refer reader to “data” folder on repo
    - Code
        - Could just refer reader to our Jupyter notebook
- Documentation:
    - How to reproduce results/How to compile and use our codebase
        - Should just be:
            - 1. Download requirements.txt
            - 2. Open Jupyter notebook and run blocks of code
    - How to navigate our dataset
        - “data” folder on repo has different folders corresponding to different datasets used to obtain our results, can just discuss different folders briefly