Analyzing and Forecasting U.S. Immigration Trends

Group 7: Alan Lin , Trevor Petrin , Ganesh Kumar Ramar , Mingyu Chen

Repository: alanklin/AA-Capstone: Boston College Applied Analytics Capstone Project

Following our comprehensive deep-dive into some of the background research on U.S. immigration, we're taking a closer look at the data that we have gathered from the U.S. Customs and Border Protection. We will perform exploratory data analysis on both datasets mentioned last week: Nationwide Encounters by State, and Nationwide Encounters by Sector (Area of Responsibility).

Nationwide Encounters by State runs from the start of the fiscal year 2020, which is actually October 1, 2019 as defined by the federal government, to the last updated date of the dataset, which is January 14, 2025. Through some cleaning and feature engineering, I confirmed that the dataset starts in October 2019 and ends in December 2024. In its current state, there are 57,780 rows of data with 9 columns. These variables and their definitions are as follows:

Variable	**Definition**	**Examples/Possible Values**
Fiscal Year	Fiscal year the encounter took place	2020, 2021, 2022, 2023
Month Grouping	Allows for comparisons between completed FY months vs. those remaining.	FYTD and Remaining
Month (abbv)	Month the encounter took place (abbreviated).	JAN, FEB, MAR, APR
Land Border Region	Border region in which the encounter occurred	Northern Land Border, Southwest Land Border, Other.
State	U.S. State in which the encounter occurred	AK, ID, TX, CA
Demographic	Categories for Individuals	Accompanied Minors, Single Adults
Citizenship	Citizenship of Encountered Individual	Canada, Peru, Nicaragua, Turkey

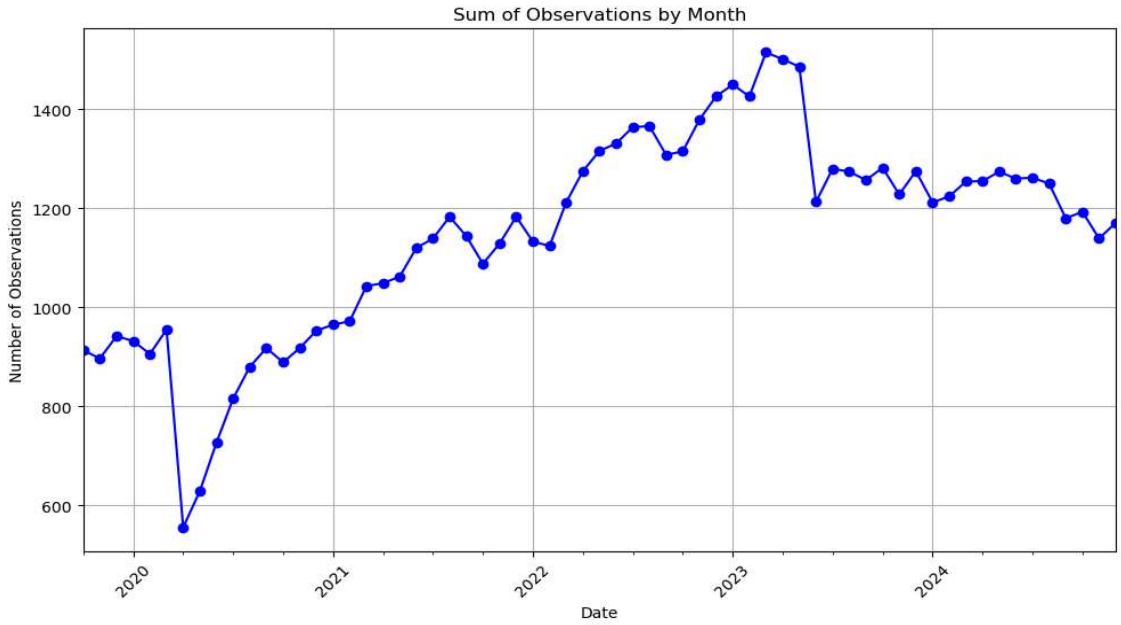
Title of Authority	Authority under which the noncitizen was processed	Title 8, Title 42
Encounter Count	Number of individuals encountered	Discrete values

We'll be using the encounter count as our target variable, as we'll want to track how many encounters the CBP is recording every month to then predict and forecast encounter counts in the future. Additionally, we have a second sister dataset for the migrant encounters which aggregates encounters by area of responsibility as opposed to by State. The benefit of doing so is it allows for a more refined look at the busier parts of the border such as in a state like Texas, where there are 4 different zones covering their border. The only main difference within the two datasets is State is transformed into the area of responsibility and component, making the analysis more granular. The benefit of having analysis at the State level would include making it more interpretable for the audience that wants to be informed but is not as familiar with the border regions. The benefit to having the more granular data with the Area of Responsibility (AOR) is it tailors the findings more towards the individuals who have the background about what the regions are as well as the field office operations and fundings.

Because the two datasets share many similar variables, we'll provide definitions for the unique variables in the Sector dataset.

Variable	**Definition**	**Examples/Possible Values**
Component	Who in CBP was involved in the encounter	Office of Field Operations, U.S. Border Patrol
Area of Responsibility	The field office or sector where the encounter occurred	Boston Field Office, Buffalo Field Office, Detroit Field Office, Portland Field Office, Seattle Field Office
AOR (Abbv)	Same as above - just abbreviated version	Boston, Buffalo, Detroit, Portland, Seattle
Encounter Type	The category of encounter based on Title of Authority and component • Title 8 for USBP = Apprehensions; • Title 8 for OFO = Inadmissibles; • Title 42 = Expulsions	Inadmissibles, Apprehensions, Expulsions

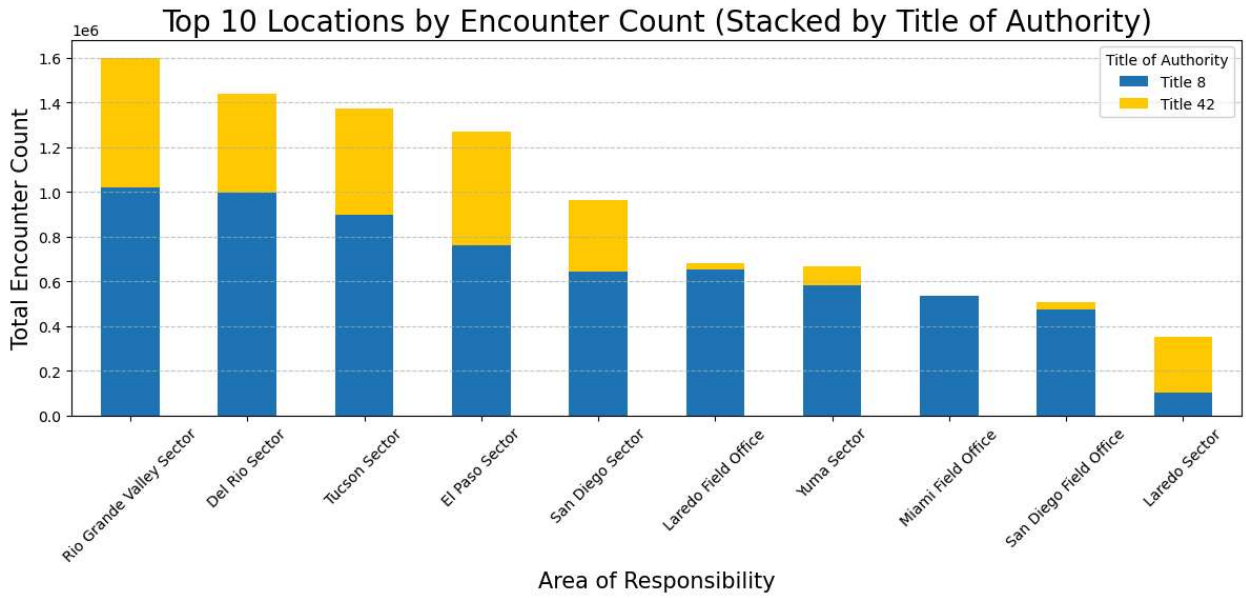
The Month Grouping variable was a bit confusing. Going off of the data dictionary, it seems like this is simply just a way for them to do analysis comparing months tagged with FYTD vs months tagged with Remaining. However, looking at the data itself, 'Remaining' implied that the Month recorded matched the current fiscal year (Jan, …, Sep). FYTD implied that the Month, though recorded in the next calendar year, for time series purposes, is in the previous year (Nov FY 2024 implies calendar Nov 2023). This was only the case from 2024-on, as the way that the data was recorded from FY2020-FY2023 was only done by Fiscal Year without the 'Remainder' tag. Thus, it's unlikely that this variable will be included in our model as it provides no information to the encounter count.



The graph above aggregates the number of observations by month regardless of encounters. This is to show potential variety in the data from encounter-specific columns such as citizenship, demographics, or title of authority. How this may affect our analysis includes that we may end up with 0-values when we are aggregating by country of origin by AOR. However, it is highly unlikely that we forecast at such a granular level for it to be considered an issue. By the same logic, we will need to pay attention if we do not aggregate by Title, as Title 42 is known to expire and can be seen when it is enacted in FY 2020 and expires in FY 2023. Overall, the main point of this graph is to show that if we are not careful when we aggregate, we could have ragged panel data, which may impact our model building process. Any values which would be

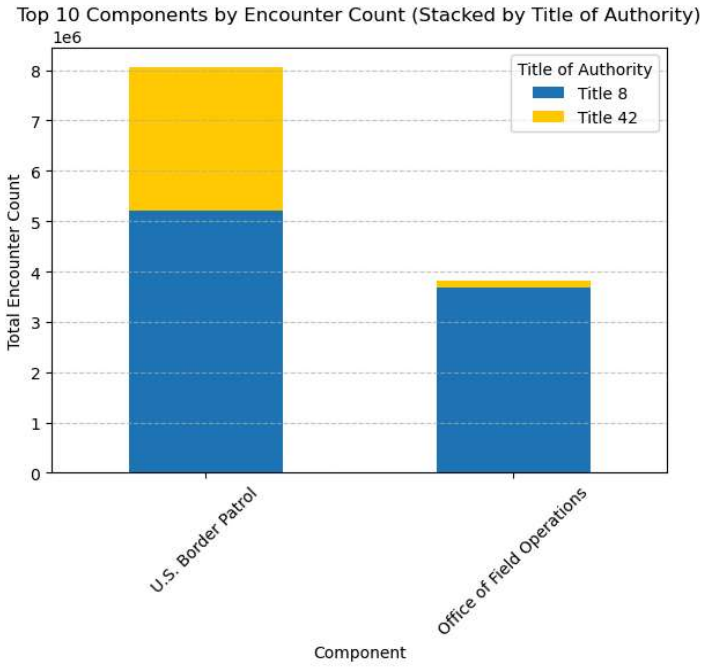
unaccounted for would be imputed with a known 0, as no individual matching that demographic passed the border at that time.

True to the timeline of Title 42 starting in March 2020, we see a large downwards tick in the number of Title 8 encounters (Appendix Figure 1), as Title 42 was often prioritized over Title 8. As such, the number of Title 42 encounters continued to grow until an average of 100,000 encounters each month was achieved in 2021, where it would hover and fluctuate until its eventual end in May 2023. A total of 2,960,908 encounters under Title 42 were recorded by the U.S. Customs and Border Protection. On the other hand, we can see a large increase in Title 8 encounters both towards the end of Title 42 and certainly much more after its termination (Appendix Figure 1). While encounter count doesn't represent actual migrant counts, it's interesting to note that there were multiple articles and officials stating migrant numbers didn't explode upwards as expected after Title 42 expired, although it seems that this graph would suggest otherwise. As of now, 8,904,261 encounters are recorded under Title 8.



Aggregating the locations by their sum of encounters, the top border sectors in terms of migrants encountered were the regions in Texas - Rio Grande Valley and Del Rio - followed by Arizona's Tucson, with El Paso coming closely in 4th. These results aren't too surprising, but we do want to draw some attention to the Title reported when dealing with the migrants versus where they are used, as all of the field offices in the top 10, Laredo, Miami, and San Diego hardly used Title

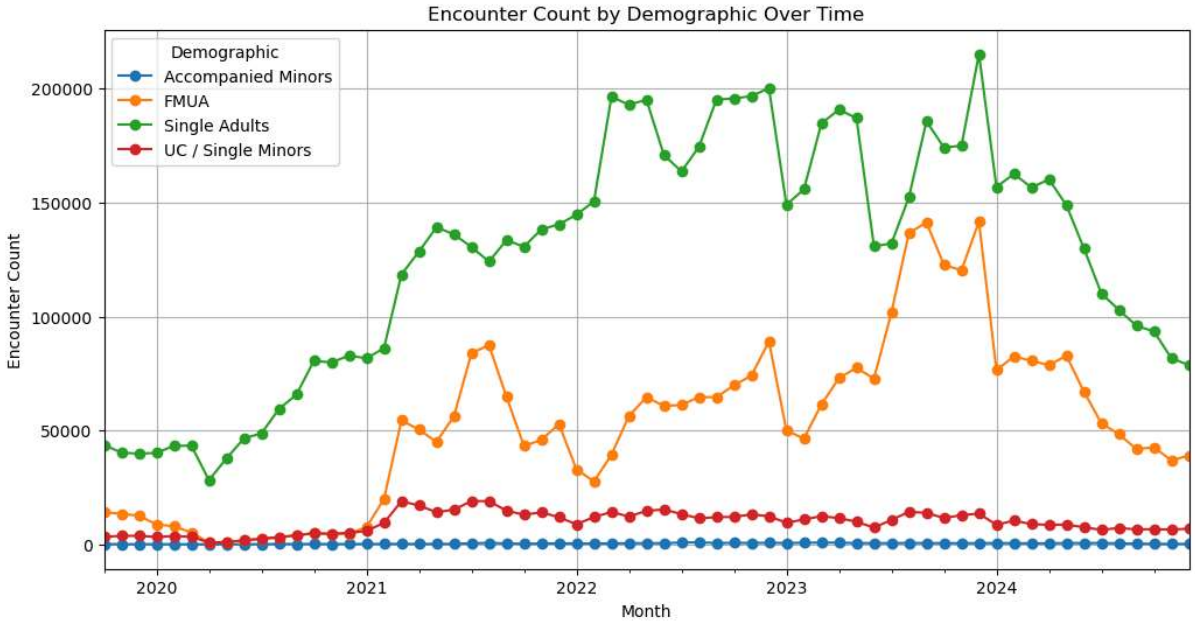
42, showing it likely to be more prevalent in the Border Sectors. The key difference between a sector and a field office is that a sector is a general border region such as Rio Grande Valley whereas the field offices are a physical address building with associated legal points of entry where migrants are encountered.



Aggregating the number of encounters by the component encountering them, it is clear that the trend of Title 42 being invoked along the border as opposed to at field offices held as an overall trend. Additionally, there were roughly double the number of migrants trying to cross along Border Control responsibility sectors as opposed to through the legal points of entry tracked at field offices. With background research, no source has definitively given a reason for this disparity in usage of Title 42 more between the border as opposed to at Offices of Field Operations. Additional resources will be found to address this shift.

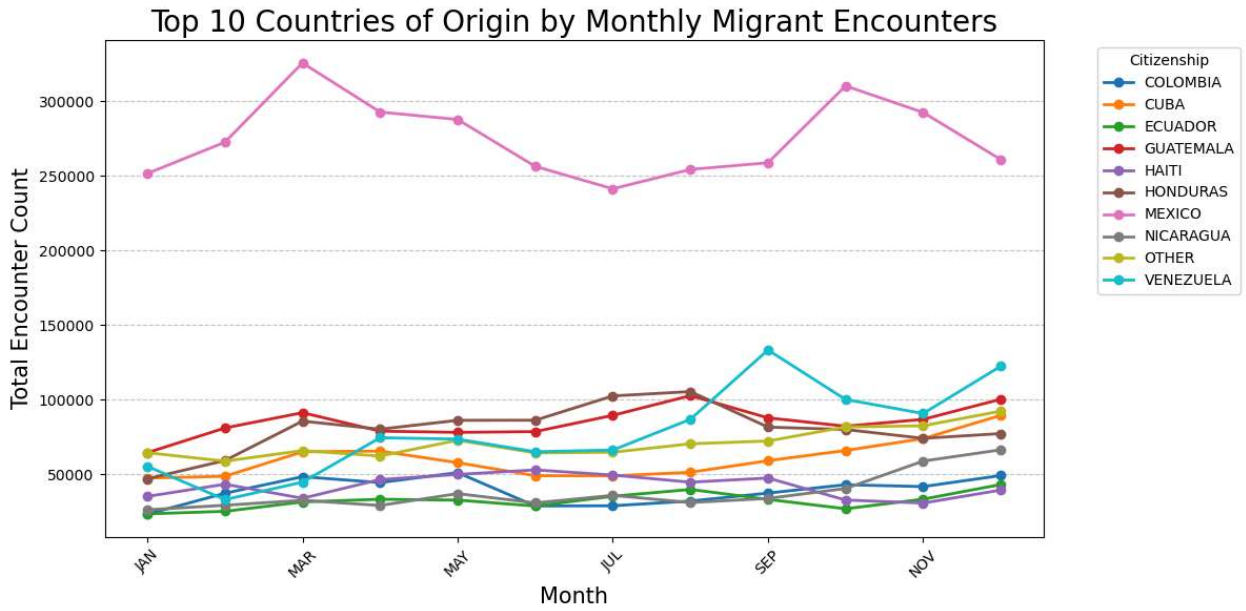
Unsurprisingly, the vast majority of migrants come through the Southwest land border, otherwise known as the US-Mexico border (Appendix Figure 3). The fewest migrants come through from the Northern land border (Canada). This reinforces why most of the discussion on border security and immigration focus solely on the southern border. We've already mentioned the top 10 locations earlier but a general visualization of the situation will certainly provide a glimpse at the immense amount of responsibility and pressure that rests on the CBP agents and offices near

the Southwest land border. One interesting tidbit that I will leave here is that in 2023, the Biden administration did attempt to open regional processing locations across the Western Hemisphere, particularly in Latin America countries like Colombia, Guatemala, Costa Rica, and Ecuador. The main purpose of these centers was to provide migrants a more local alternative to access screening for lawful immigration pathways. By reducing the need for migrants to travel all the way to the land border, crossing numbers might be more manageable, migrants might not have to rely on cartels for passage, and they certainly wouldn't have to endure the possibility of being stranded at the border in inhospitable conditions. Unfortunately, it would seem with changes in administration and policy that these centers are no longer accepting applications, which might push more migrants to undergo the grueling journey up again.



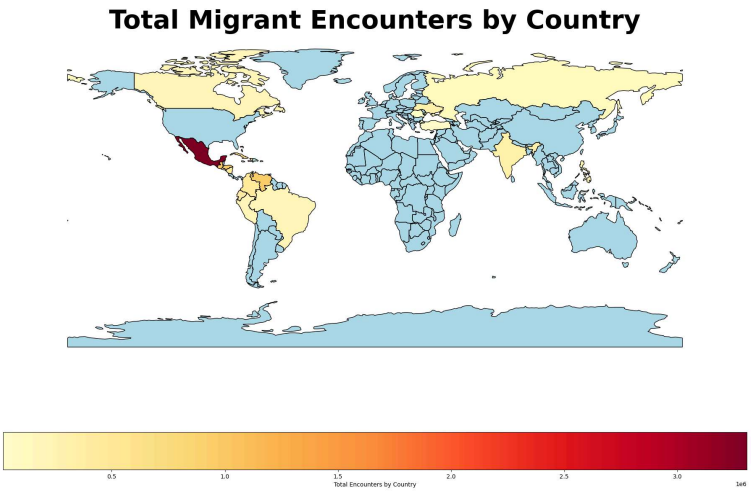
Demographic trends show that Single Adults and Family Unit Aliens (FMUA) are the most common type of migrants - which is unsurprising. A Family Unit is defined as at least one noncitizen minor and at least one noncitizen adult parent/legal guardian. Accompanied Minors, on the other hand, refer to a child accompanied by a parent/legal guardian that is either a U.S. Citizen, Lawful Permanent Resident, or admissible noncitizen, but the child is determined to be inadmissible. This variable might be very important, as we believe it's more likely for an individual (Single Adults) to take more risks, as opposed to endangering an entire Family Unit. Appendix Figure 2 looks closer at the two other demographic groups, and I wanted to take the time here to express concern in the number of unaccompanied minors. These are, in my opinion,

one of the most vulnerable groups, not just to health risks at the border but to their own safety for the sake of being taken advantage of.

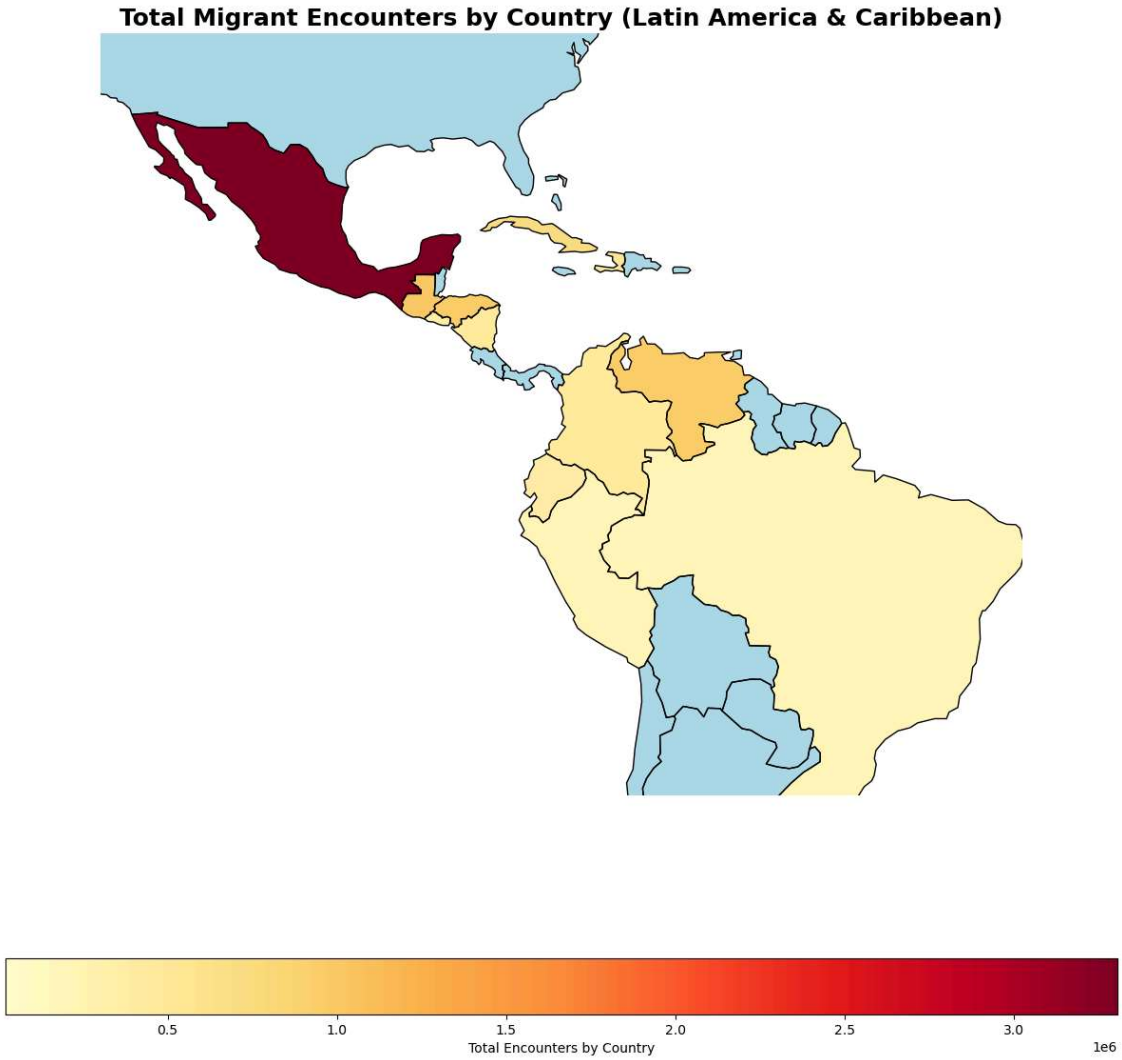


This figure shows that migrants of Mexican citizenship are coming to the US Southern border at roughly 5x the monthly rate of other common nationalities. Regionally, most migrants are coming from South and Central America, however, there is also a sizable population of migrants coming from the Caribbean from countries with harsh political circumstances such as Cuba and Haiti. When feature engineering is done, it would be wise to add a column describing the region of the country to better show where migrants are coming from.

The data clearly shows that a majority of the migrants are coming from South and Central America which corresponds with the literature researched on this topic. However, the Northern border cannot be ignored, with this and the previous graphs showing that there are a substantial number of migrants coming South from Canada as well. Outside of the Western Hemisphere,



there are significant numbers of migrants encountered from Russia, Eastern Europe, and India as well. In feature engineering, there will need to be a more comprehensive mapping created to bridge the divide between the keys of the mapping package, geopandas, and the USBCP dataset, as countries such as CHINA, PEOPLES REPUBLIC simply appear as China in geopandas.



Focusing in more upon the region of Central America, South America, and the Caribbean, it can be seen that most of the encounters are from countries geographically closer to the US Southern border such as Mexico, Guatemala, Honduras and El Salvador and gradually fade away the further South the country is. However, the exception to this is Venezuela, which was expected with the migrant by year graph, experiencing significant political turmoil and inflation within the country over the last several years.

Appendix:

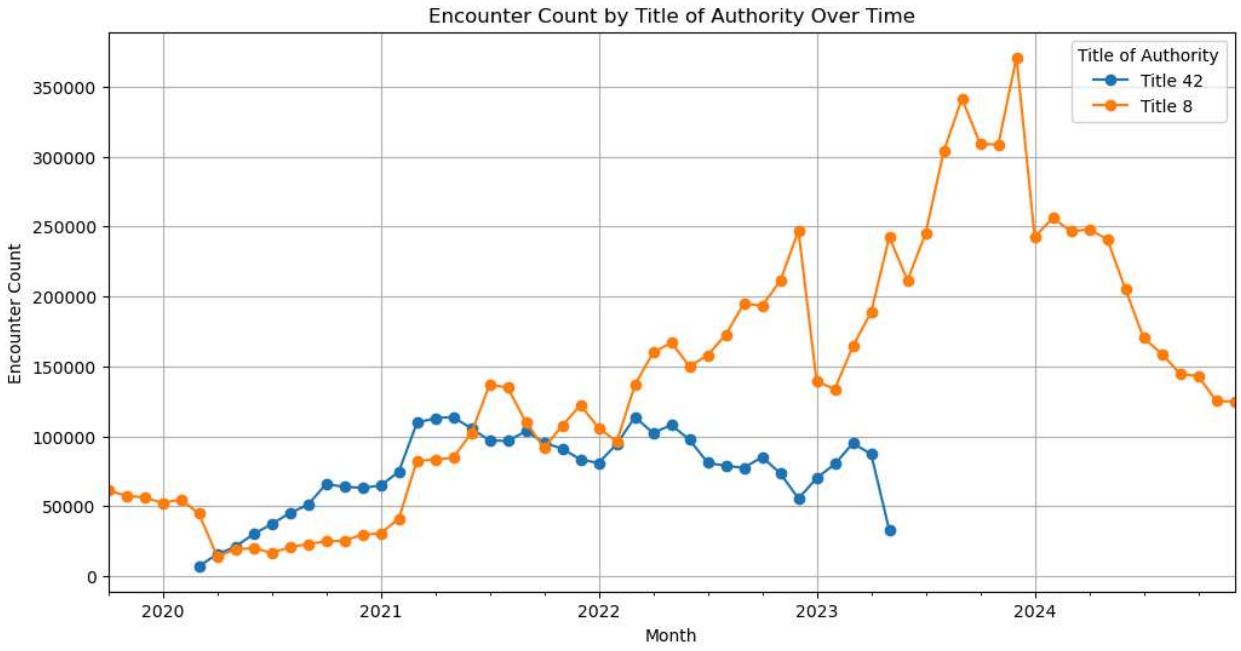


Figure 1: Relationship between Title of Authority and Encounter Count over time.

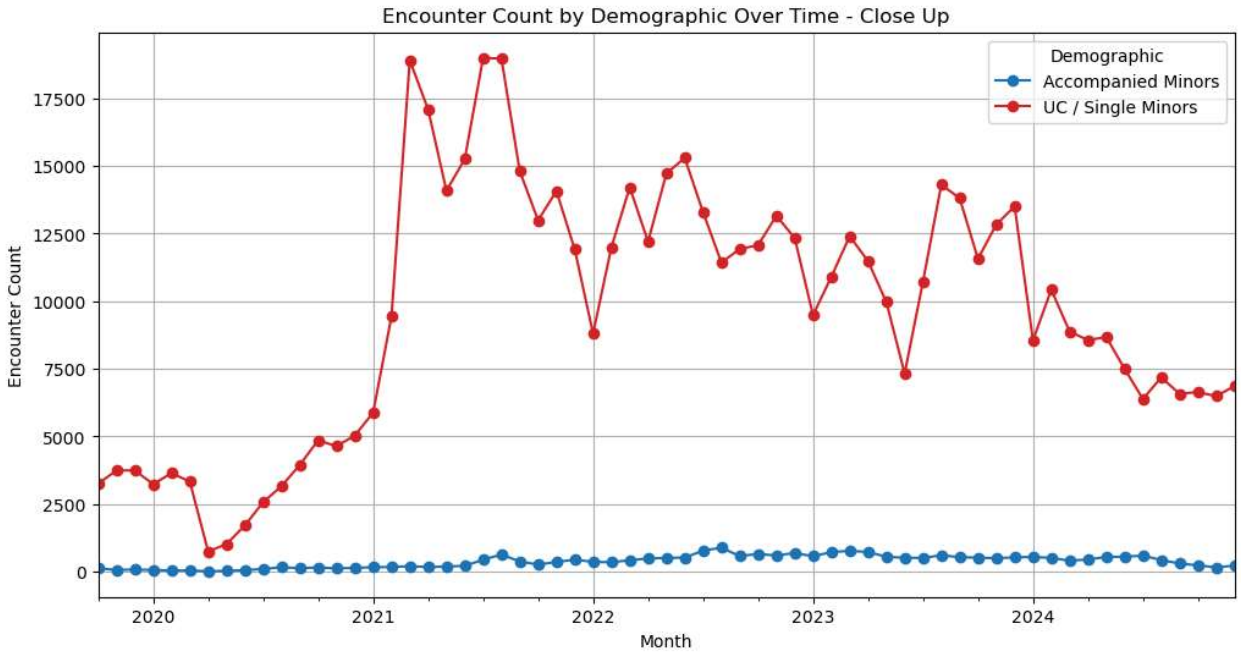


Figure 2: Relationship between demographic over time -

Wanted to zoom in closer on the two demographics that aren't as represented in the dataset. There is still a good amount of variability in the numbers of UC/Single Minors.

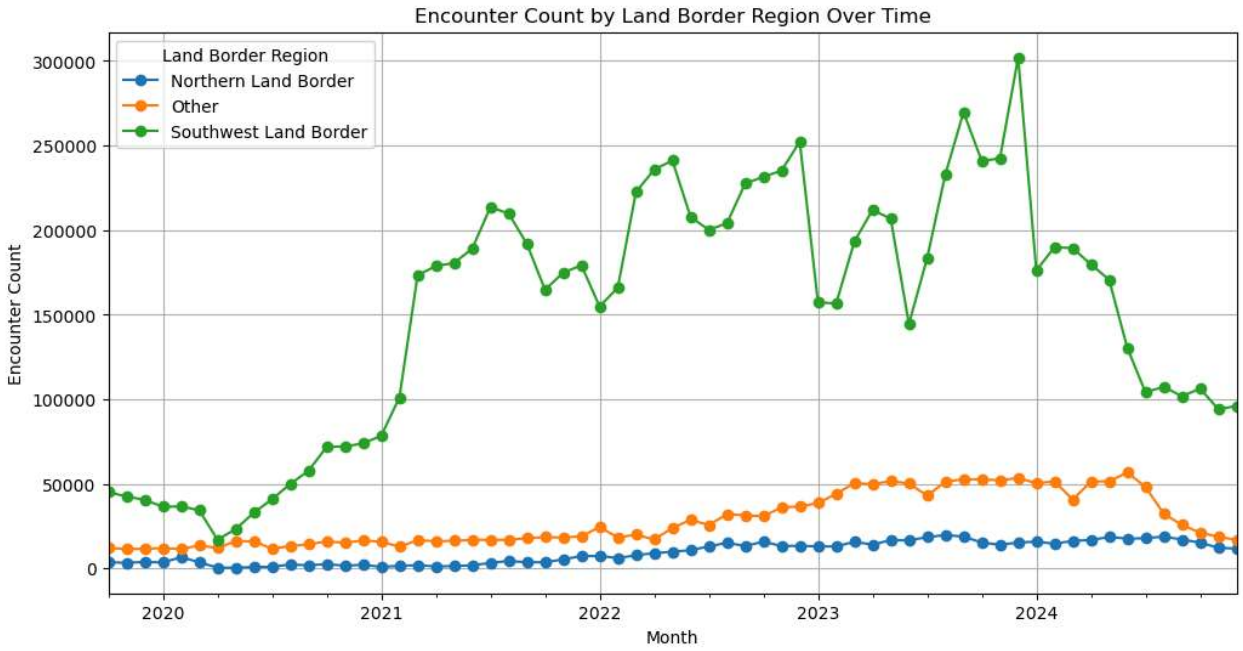


Figure 3: Graphing relationship between encounter count and land border region over time.