Analyzing and Forecasting U.S. Immigration Trends

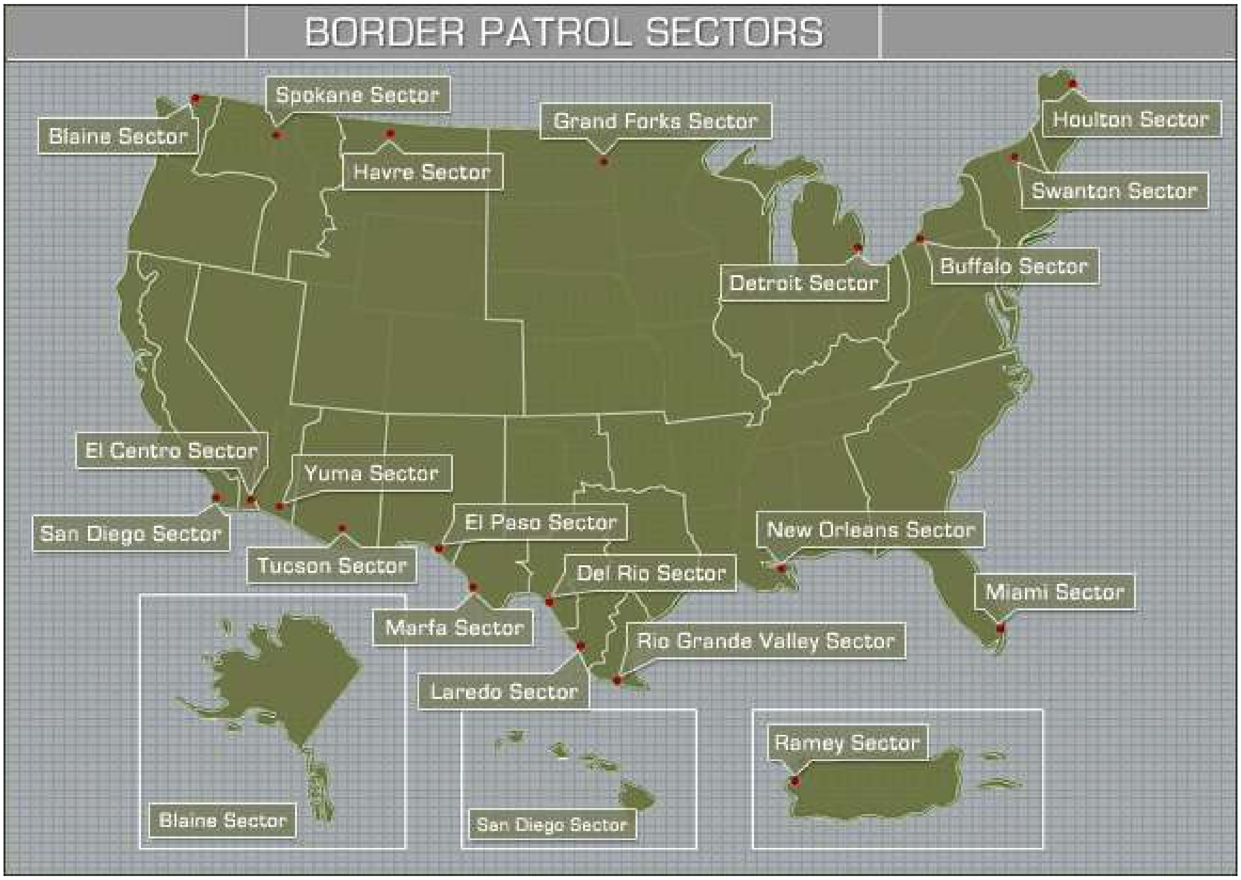
Group 7: Alan Lin , Trevor Petrin , Ganesh Kumar Ramar , Mingyu Chen

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

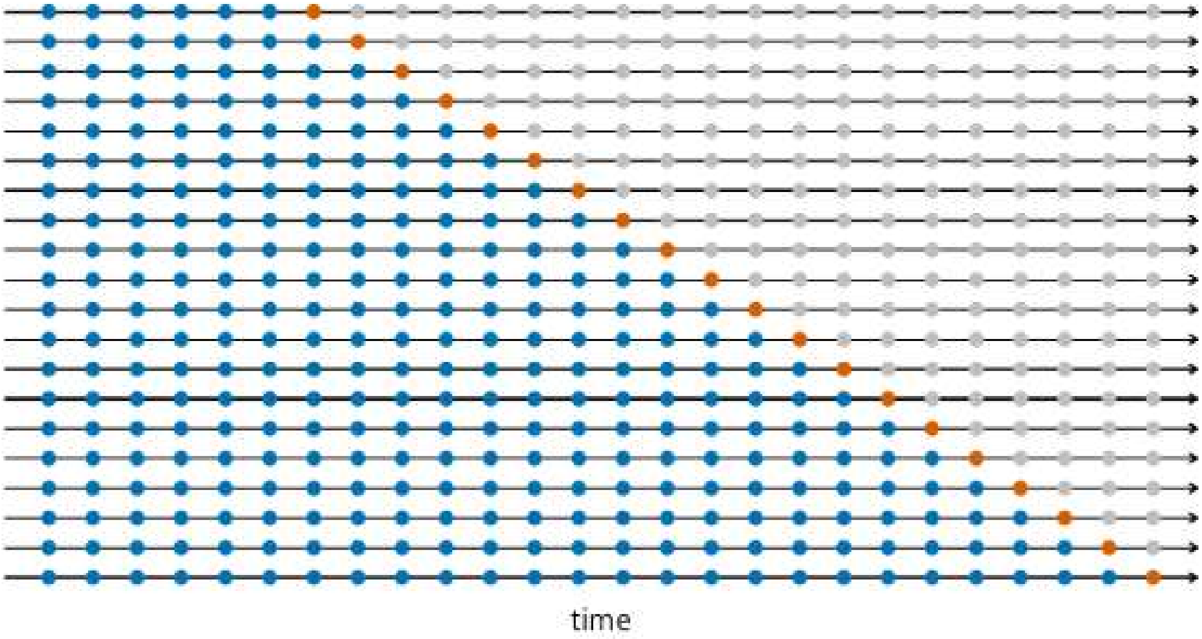
This week, Exploratory Data Analysis (EDA) will be performed upon the border encounters datasets at both the State and Sector level. What our team is looking to find are patterns within the data that may affect the model building process in forecasting the number of encounters of migrants in the calendar year of 2024-on. This will include graphical analysis to illustrate the data as well as provide a better idea for the origin of these migrants.

Last week, we performed an intensive dive into understanding our dataset as we began to think on how pre-processing would look like for our problem. One of the first steps we have taken is to create our train and test splits, for both State and Sector datasets. Due to the nature of time series, the usual randomized train/test splits using the sklearn package is out of the question. We have to keep the temporal nature of the dataset intact for our predictive models, so the best method to achieve this would be to split the dataset by date. Unfortunately for us, our datasets don't have a clearly defined date variable to easily split the dataset. It is necessary for us to perform our first feature engineering task - to create a variable that accurately represents the month and year in which the encounters were recorded. This was a simple endeavour as we can easily create a mapping for the months of the year and concatenate together a date string using the fiscal year variable. In addition, we had to account for slight discrepancies in calendar vs. fiscal year. As we may have mentioned before, the US government fiscal year begins on October 1st. So, while the data might say October FY20, it's really referring to October 2019. We reflected these minor changes in a custom conversion function to generate the correct month-year strings for each row. Finally, we could then split both datasets on our newly engineered feature. We decided that one year's worth of data would be appropriate for a test set, so we settled on using January 2024 - December 2024 as our test set, with the remaining dataset as the training set.

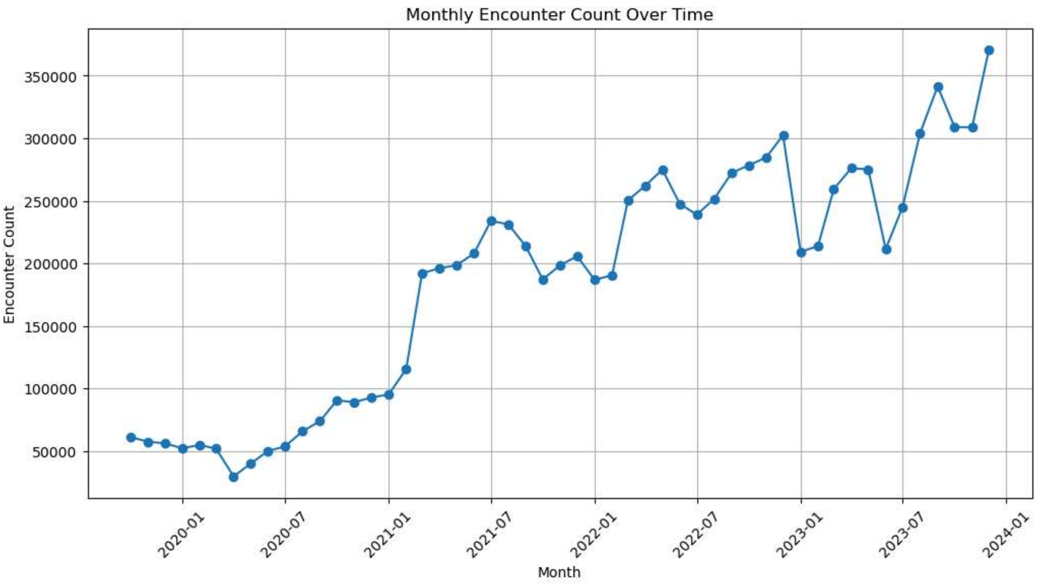
The next step would be to convert the fiscal year data into a straight time series data, which is where the 'Month Grouping' variable came into play. Though the initial definition we found in the Customs and Border Patrol dictionary was confusing, the main difference that the variable implied was ‘Remaining’ meant that the calendar year matched the fiscal year whereas ‘Fiscal Year’ implied that the calendar year was 1 more than the calendar year. How this was handled was simply through using a dictionary to subtract 1 from the year if it was ‘Remaining’. However, not all of the data was done this way, where the data definitions from FY2020-2022 only included ‘Fiscal Year’ and not remaining. To handle these cases, 1 was subtracted from the year of any month labeled October-December, as those are the months that are the next fiscal year but not the next calendar year.



The train/test splits would be performed on both the State and Sector datasets separately. The notion that the datasets be merged was brought up by the team, however, because sectors (courtesy University of North Texas web archive) do not necessarily break at the State lines, there is no possible way to map which encounters occurred at which state within each sector, or vice versa. Both datasets do still add value, though, as forecasting trends at the State level allows better for explaining in layman's terms where the migration is occurring, and good open-source packages exist for graphically representing these trends which would certainly benefit our project. It may also be more effective to forecast at the state level instead of the sector level. However, for keeping the audience of US Customs and Border Patrol, these findings would mean much more at the sector level, as it is how their funding is allocated and would provide more meaning for where they put their resources. For example, the State of Texas has 4 separate Sectors along its border alone, where just saying ‘Texas should expect more migrants’ will not add any value. There could be an influx at either El Paso or the Rio Grande Valley, for example, two of the Sectors with the highest encounter counts. As for how we plan to handle these separate datasets within our model building process, we plan to keep them separate and build models separately for both datasets because we cannot merge them, and both provide information that would add value to our final findings.

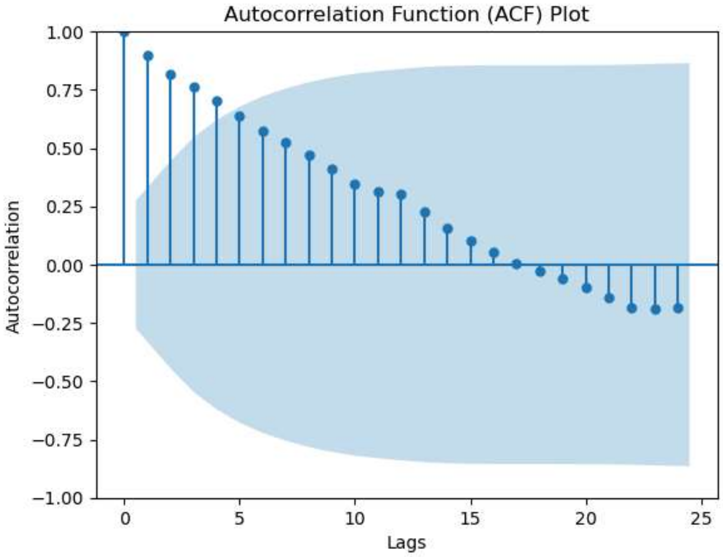


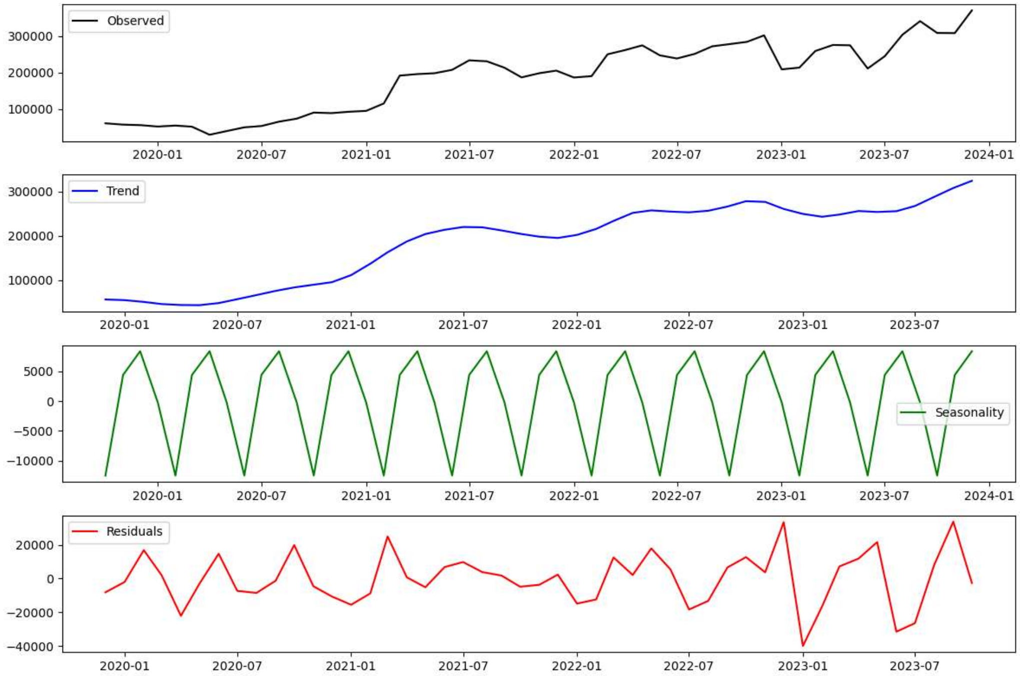
Our method for cross validation is detailed here (link: https://otexts.com/fpp3/tscv.html), where we are planning to validate our models based upon the principle that the observations must occur prior to the validation data. As is shown in the display from the book, the models are repeatedly trained on the observations (blue) before the validation set (orange) leaving out the grey spaces. This is done to have a better idea of how the model will perform upon unseen data.



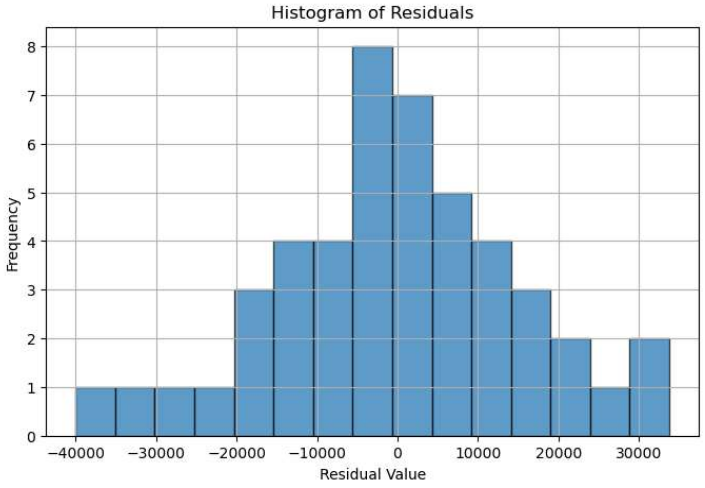
To begin, it would be wise to examine overall encounters by month, getting a general idea of the trend of the data. An additive decomposition will be performed to give insight to the trend, seasonality and error of the data. There is a positive linear trend within the data with potential seasonality that looks to be a roughly 6 month cycle. For outside factors that play a role in the number of migrants, there is a small dropoff in March 2020 when Title 42 and the Covid-19 Pandemic began, and a significant increase in May 2023 when Title 42 expired.

The ACF of the data is plotted to see which months are statistically significant at an alpha of 0.05 confidence level. Months with a lag of 1-4 are statistically significant, and to understand performance, a seasonality of 4 months will be used for the initial additive decomposition of the data. Multiplicative decomposition will be used as well to see if there is a more consistent result.



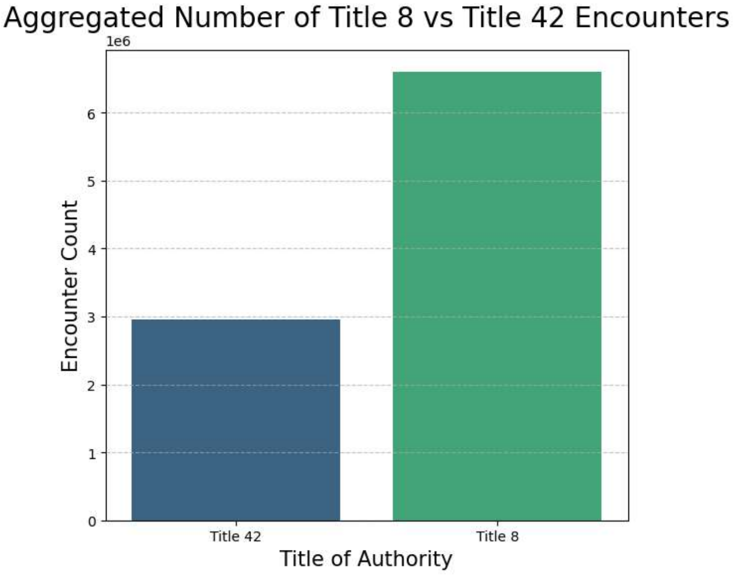


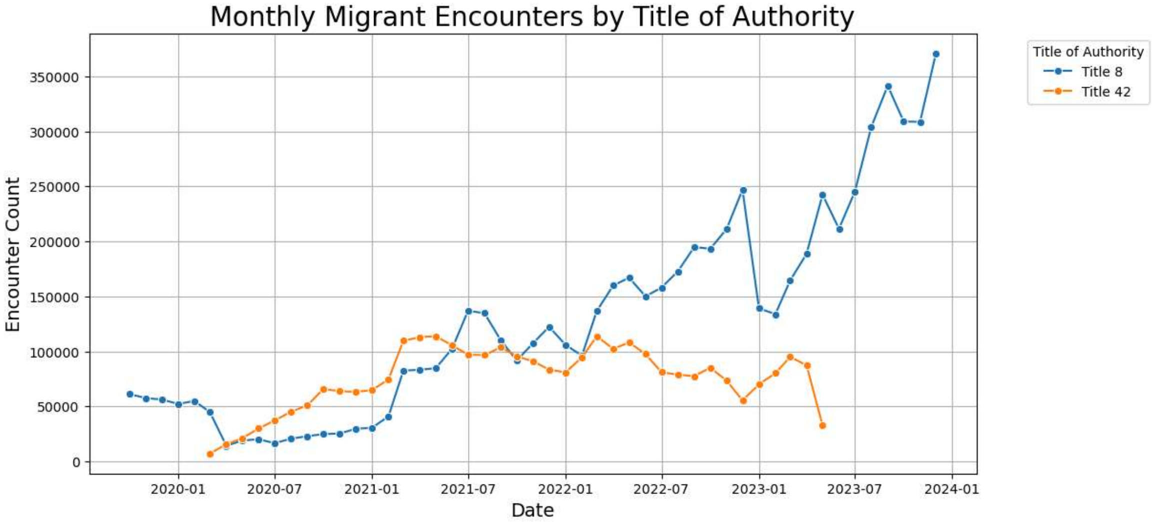
For additive decomposition, there is a positive linear trend within the data with a 4-month seasonality that peaks 5000 encounters above the trend in the second month before dropping down to 10000 below the trend in the 4th month. The residuals are normally distributed and do not seem to exhibit any kind of pattern, ensuring that the noise encountered when creating this model was random.



We can also explore if there are seasonal patterns that exist by month. Based on figure 4 in the appendix, it's clear that average encounter counts are at its lowest during the months of January-March, or during the winter months before increasing throughout the rest of the year. There appears to be a peak at the end of summer, in which September of each year experiences the highest average encounter count. Understanding these monthly/yearly patterns may help Customs and Border Protection expect and prepare for busier months.

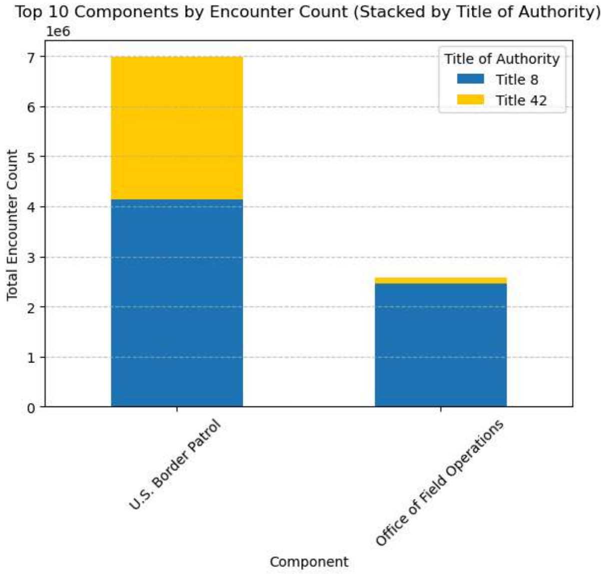
Though these general trends and seasonality are insightful to the dataset as a whole, how the migrants are being handled at the border with regards to immigration law, and when the individuals encountered are being processed under Title 8 or Title 42 may impact the forecasting of encounters. The policy's implementation and impact should be further investigated to see if it had a targeted impact upon specific migrants depending upon their country of origin.

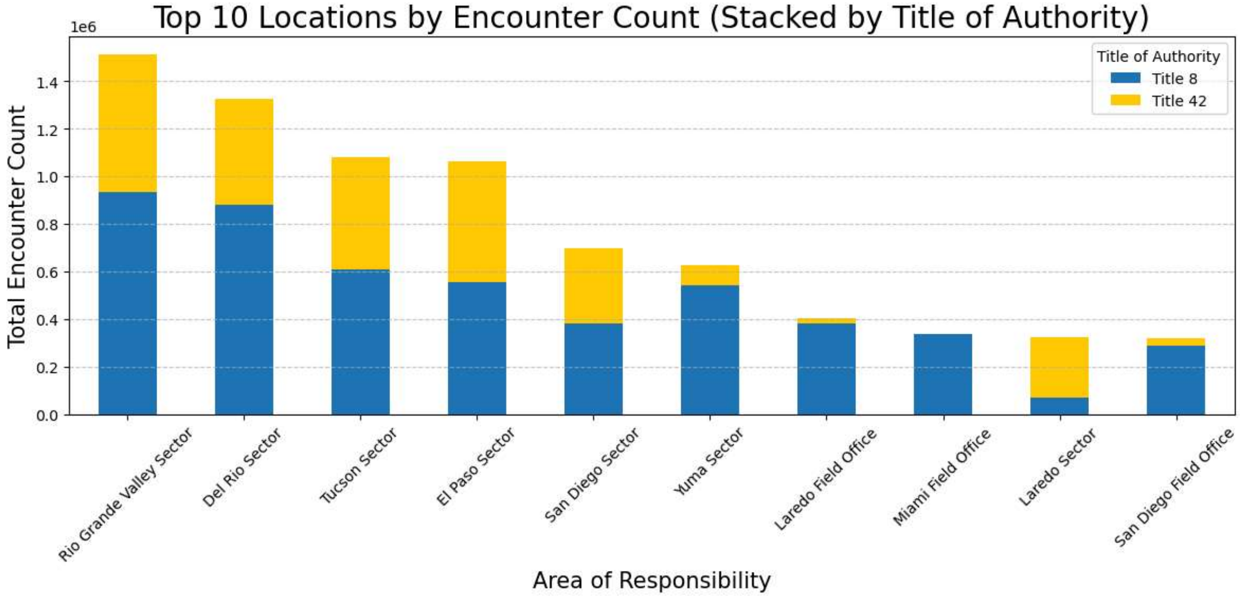




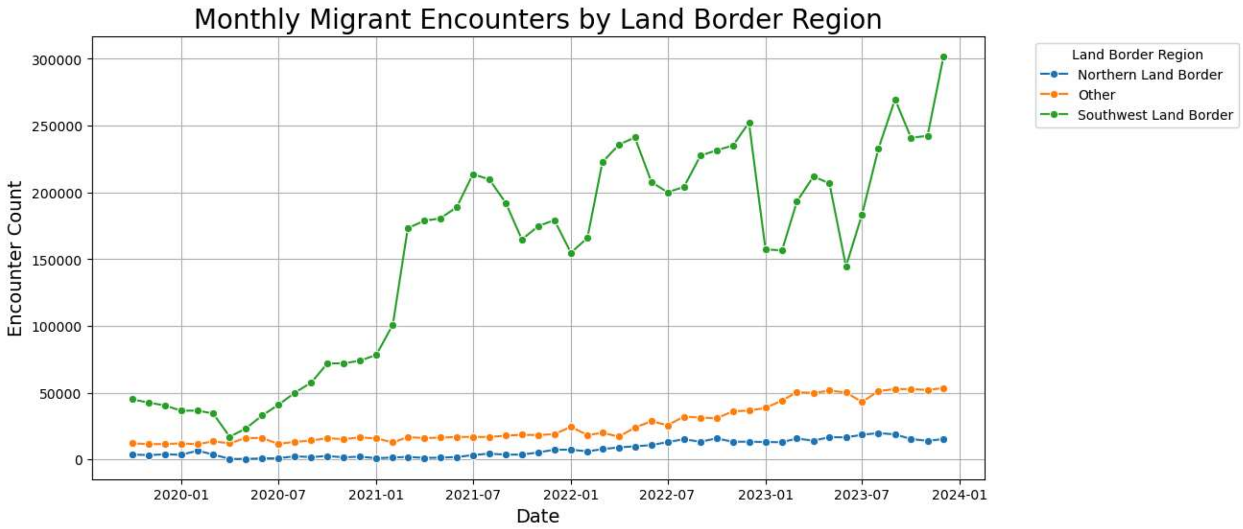
As to be expected, the number of encounters processed under Title 42 ran through the range March 2020 through May 2023 based upon the executive order placed upon the border to handle the public health crisis of Covid-19. However, it was only until 2022 that the number of Title 8 encounters and Title 42 encounters fluctuated with a significant rise in Title 8 cases after March 2022. Further background research will be conducted to see if there was a root cause at this point. For the trend of Title 8, it appears to continue its trend of linearly increasing without a sign of slowing down. Looking ahead to the model building portion of this project, positive, neutral, and negative scenarios will need to be considered to show multiple possible outcomes of the trajectory.

As for the use of Title 8 versus Title 42, the use of Title 8 is much more prevalent, especially within legal points of entry handled by the Field Offices. Most of the Title 42 encounters were reported along the border sectors, handled by Border Patrol. This trend is explained by the fact that Offices of Field Operations represent legal points of entry to claim asylum versus the Border Patrol sectors which are easier to turn migrants away at.

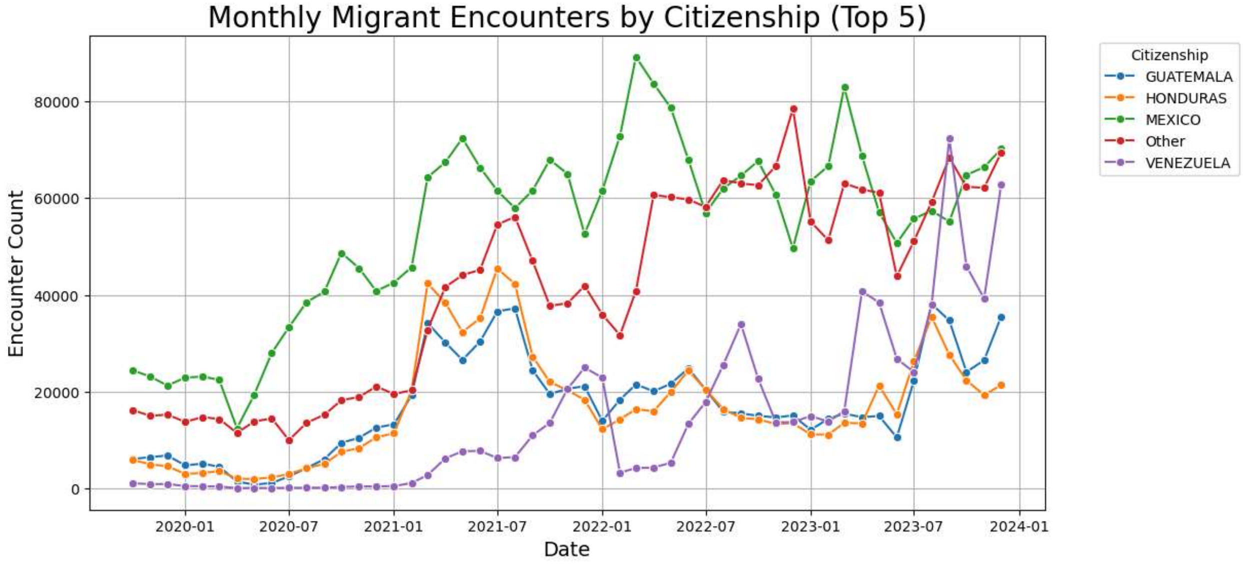




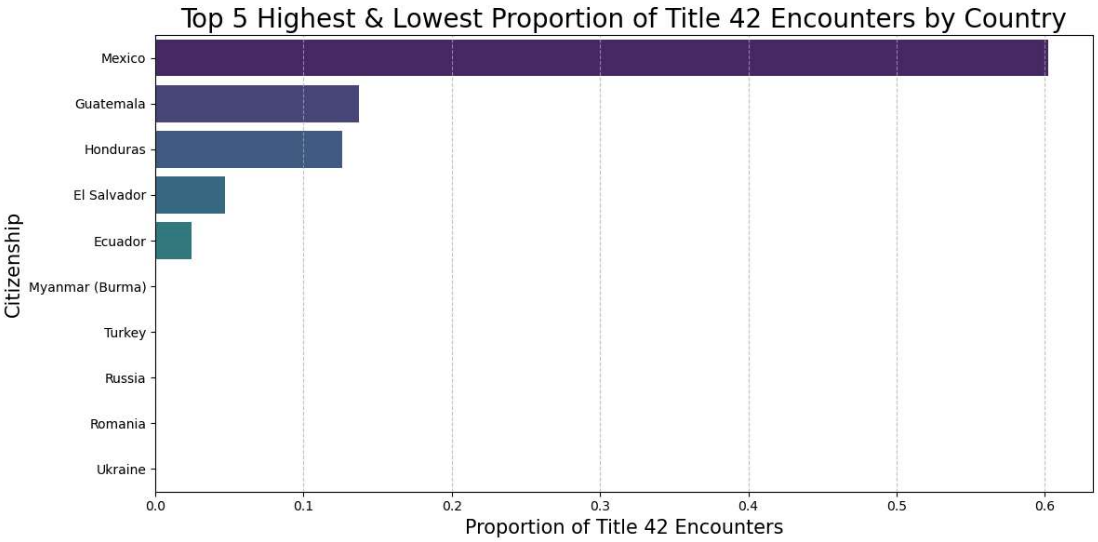
Aggregating where the regions that the migrants are being encountered by, geographically, borders along the Texas boundary are the busiest accounting for 3 of the 4 busiest regions measured. Arizona has busy regions as well accounting for both the Tucson and Yuma sectors. The disparity in handling Title 8 versus Title 42 cases by port of entry can also be seen clearly with the field offices taking barely any cases of Title 42, whereas roughly half of all of the Sector encounters are Title 42.



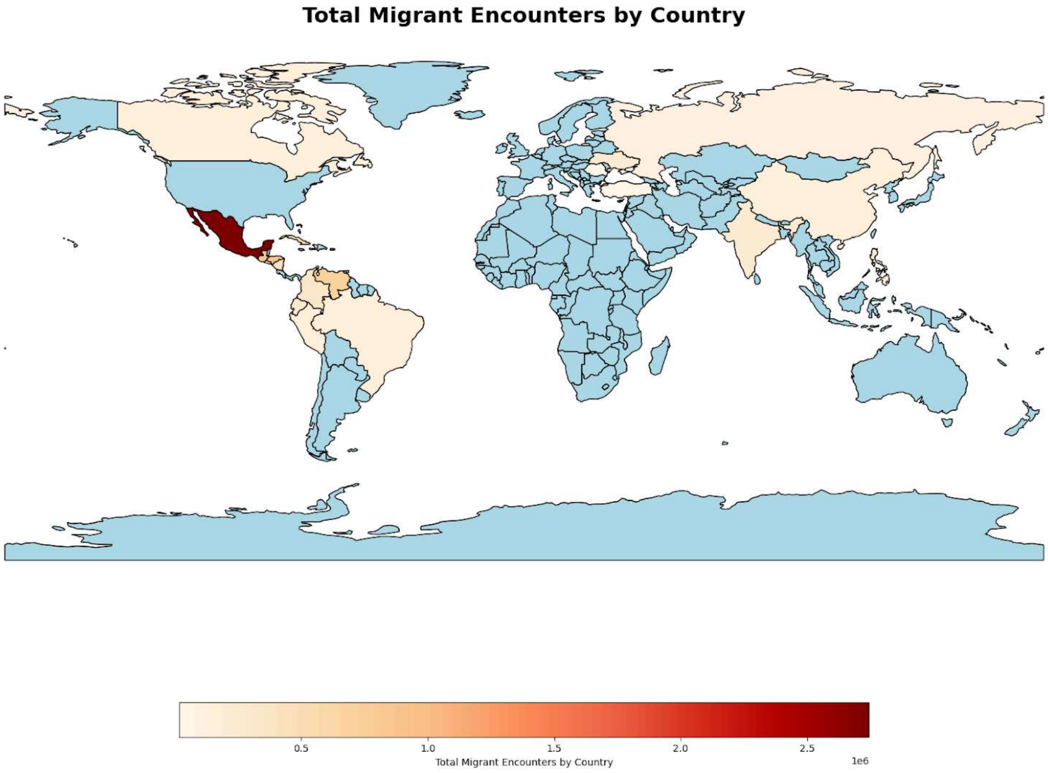
These encounter locations are consistent with the overall trend that most of the encounters are coming from the US Southern border as opposed to the Canadian border or a separate port of entry such as an airport or seaport.



It is just as important, though, to understand the underlying trends in the migrant arrivals as it is for their port of entry. For the top few citizenships encountered at the border, most of the individuals were from Mexico, with Honduras and Guatemala also in the top 5. This underscores the trend that most of the individuals encountered are from Latin and South American countries, with Venezuelan migrants increasing substantially in the past few years.



For Title 42’s use to expedite the removal of encountered immigrants, this policy appears to disproportionately affect citizens from Mexico, with Guatemala, Honduras, and El Salvador also making up a big chunk of all total Title 42 encounters. This is opposed to the lowest proportions from Eastern Europe and Southeast Asia, mostly countries with corruption or war in their current political situations. However, there is not a large difference between the 5th highest proportion of Ecuador (0.03) and last-place Ukraine (0.00).


Total Migrant Encounters by Country
0.5
1.0
1.5
2.0
2.5
1e6
Total Migrant Encounters by Country

The heatmap visualization illustrates the distribution of migrant encounters by country, highlighting the geographical origins of individuals seeking asylum or crossing borders. The heatmap of citizenship of origin for migrants supports the suggestion that a majority of the migrants are coming from Central and South America. However, it should be noted that this is not a uniquely Southern issue, with migrants claiming asylum from Canada as well. There are a significant number of individuals claiming asylum from Eastern Europe and Southeast Asia as well, with Russia, China, India, and Burma all having a significant influx of migrants. Certain major displacement events such as the Russian invasion of Ukraine at the start of 2022 have also contributed to a rapid rise in Ukrainian citizens under Title 8. (Appendix Figure 2). These external factors that are reflected in our EDA only further demonstrate how current events can have an extreme impact on immigration. These shocks will have to be accounted for in our modelling later down the road, and we will further discuss how we might appropriately add fluctuations in our forecasting to better resemble the volatility of US immigration trends.

As we've mentioned plenty of times before, immigration trends are a complex problem, and it's quite unlikely that a simple ARIMA model or STL forecasting model will be sufficient to accurately predict future values of encounter counts. The EDA we performed today, as well as all of the current news and policy changes that we've researched, all show that there are many external factors that go into this complicated system. A more advanced model using neural network architecture may be necessary to accurately forecast and answer our problem statement. Different forecasting methods such as LSTMs, and Transformer architectures will be used to apply more modern machine learning techniques to see performance against the more traditional aforementioned models. It will be interesting to see which models predict positive, negative, and neutral outcomes based upon our test data.

Going forward, the initial analysis performed in this report will be invaluable as we continue onto the model building process for understanding and interpreting the results of our forecasting. Further precautions and continued analysis must be done as well with our plan to aggregate the data to better forecast the number of incoming migrants, ensuring all missing values are imputed correctly and that no further problems within the data arise.

Appendix:

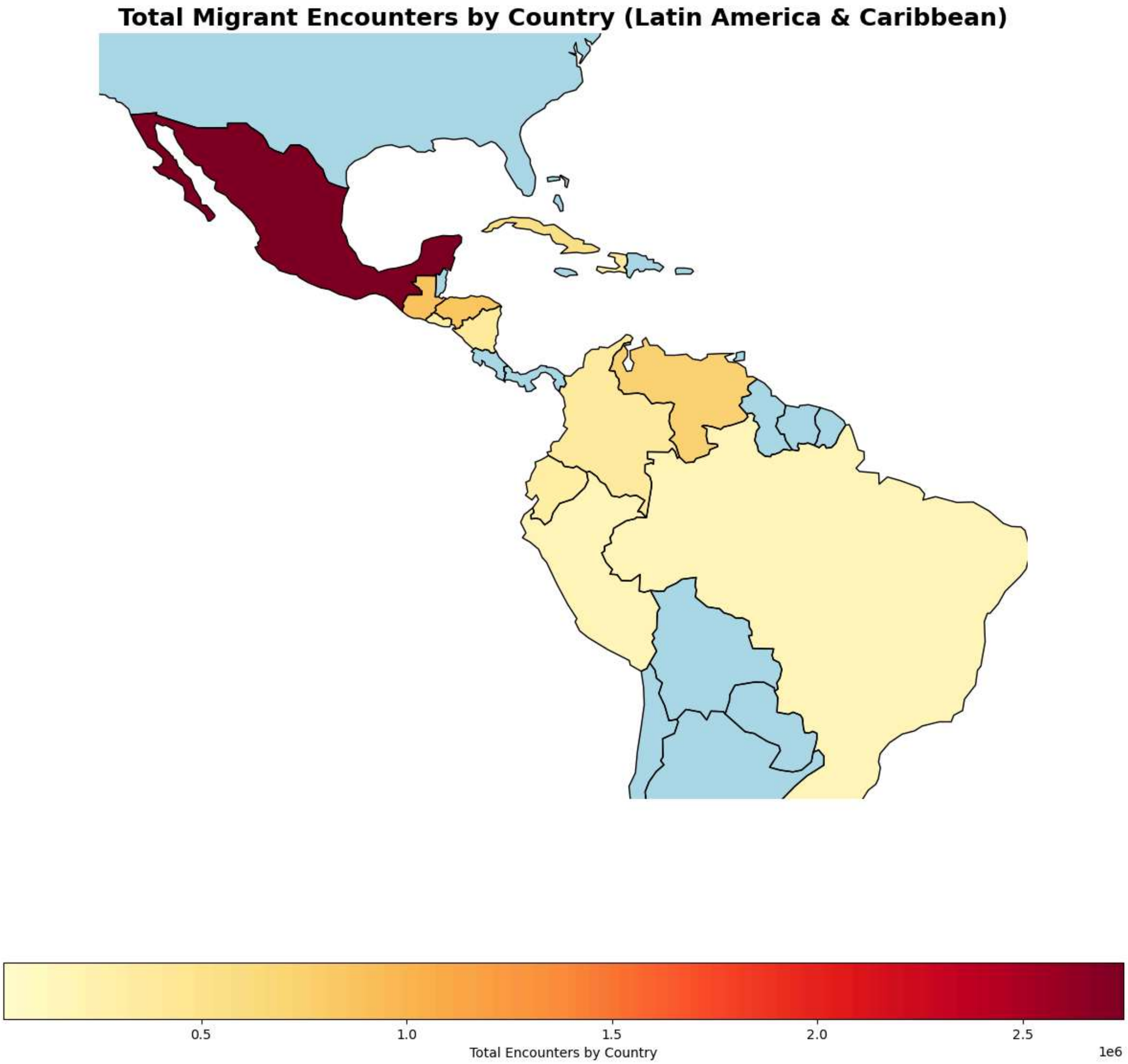

Total Migrant Encounters by Country (Latin America & Caribbean)
0.5
1.0
1.5
2.0
2.5
1e6
Total Encounters by Country

Figure 1: Total encounters focused upon Latin America and the Caribbean regions.

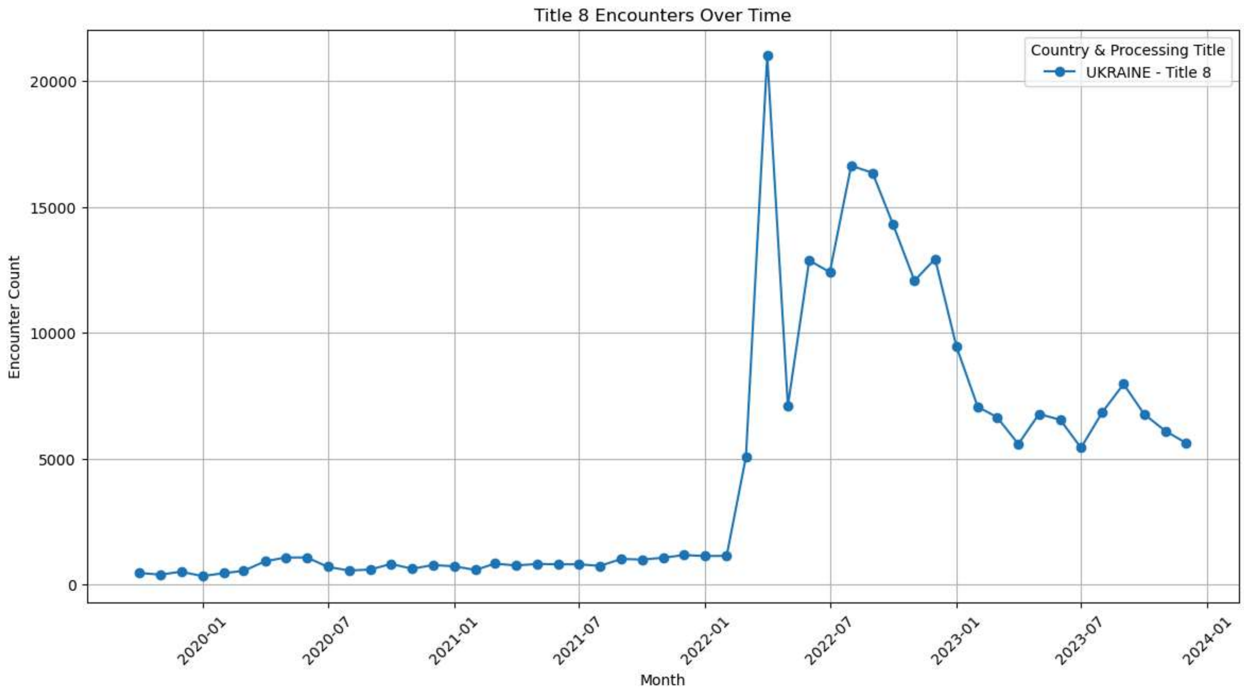


Figure 2: A huge difference in Encounter Count caused by the Russian invasion of Ukraine.

The graph in Figure 2 illustrates the number of migrant encounters from Ukraine over time, with a sharp increase starting in early 2022. This surge is directly correlated with the Russian invasion of Ukraine in early 2022. Before 2022 the number of encounters remained low and stable but following the onset of the war, there was an immediate spike in migration as Ukranians sought asylum and refuge, particularly in Western countries.

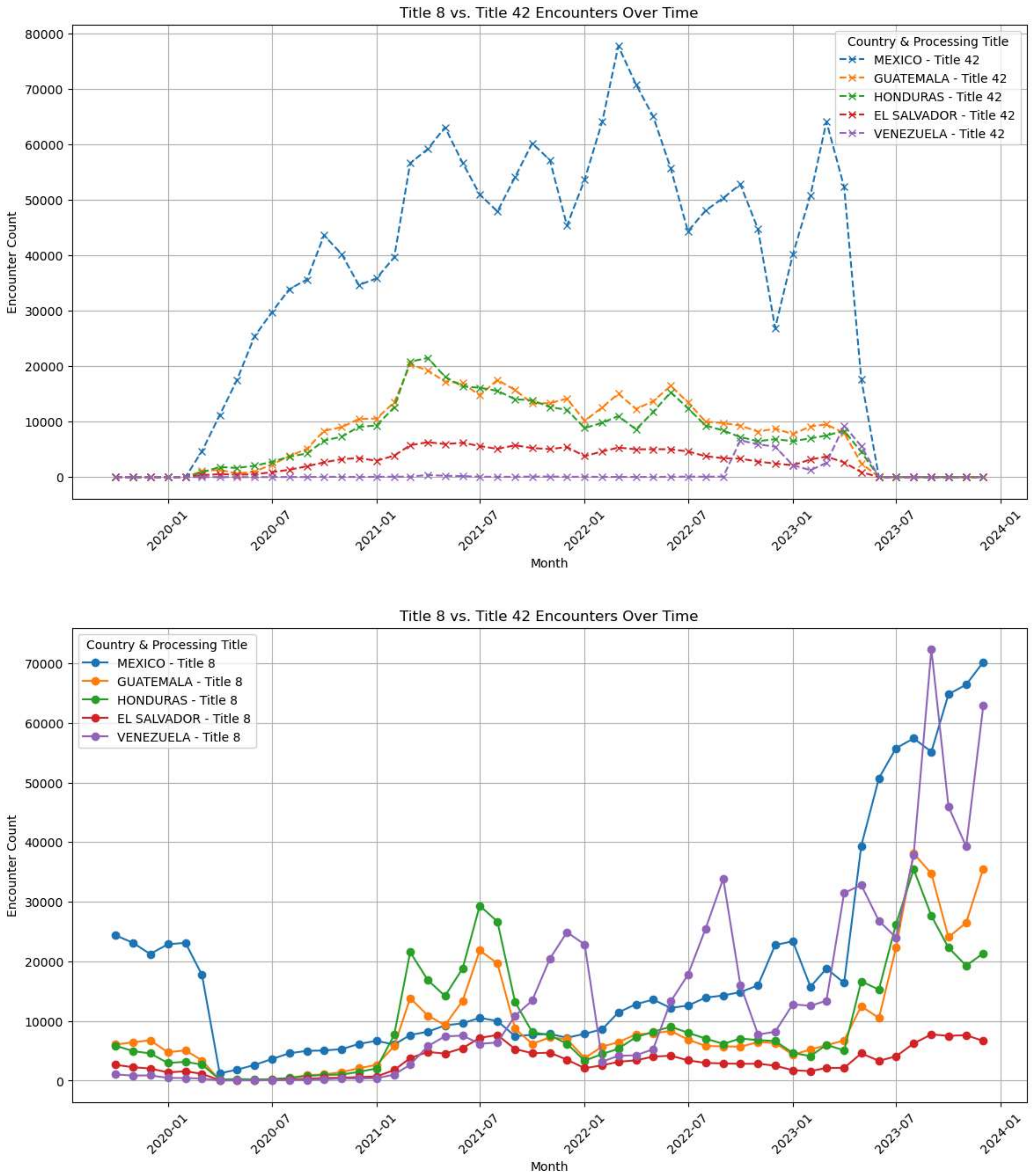


Figure 3: Countries heavily affected by Title 42, Comparison between Title 8

The graph indicates that Mexico had the highest number of Title 42 encounters in early 2020, followed by Honduras, Guatemala, and El Salvador. Venezuela was added to the figure to assess Title 8 encounters and the sharp influx in 2022 that contributes to the trend we see on Page 4.

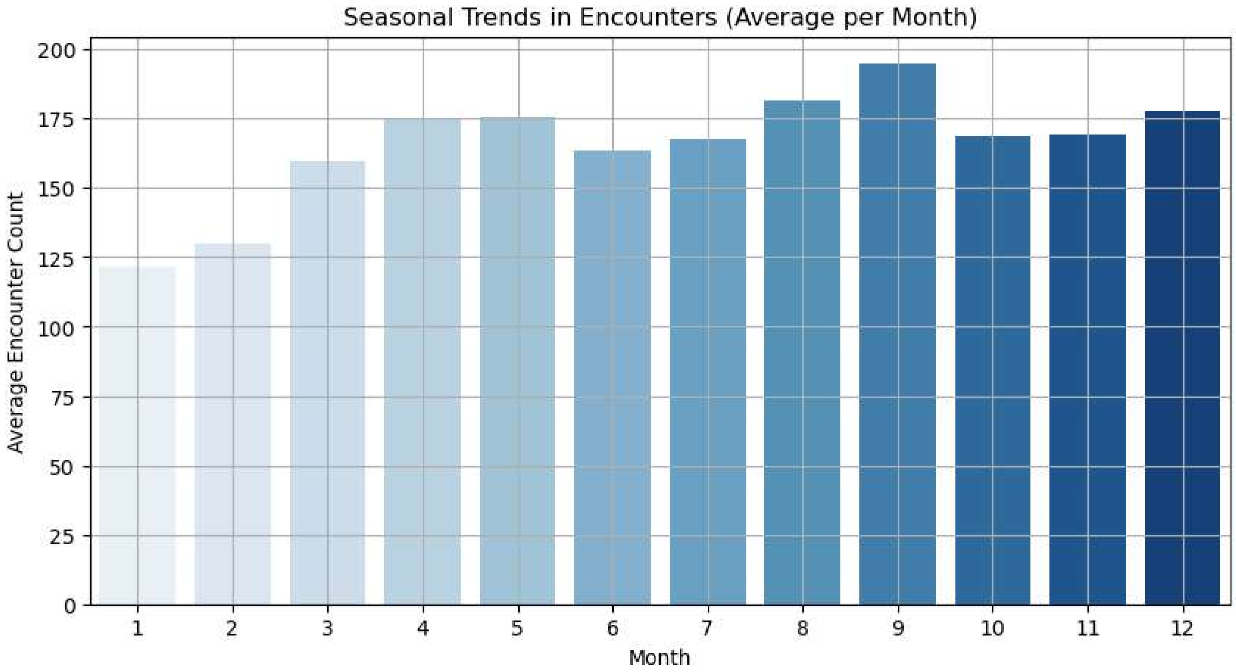


Figure 4: Average number of encounters by month

The average number of encounters per month is taken, showing a slightly higher number of encounters during the summer before dipping in the early months of the year. This will need to be further investigated using seasonal decomposition methods.

Sources:

Hyndman, R.J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 7 Feb. 2025.

U.S. Border Patrol: Locations & Job Opportunities in the United States, University of North Texas Libraries, 7 Nov. 2008, webarchive.library.unt.edu/eot2008/20081107215339/www.borderpatrol.gov/interactive_map.html.