LISA-HONG FINAL WRITE UP

Due: Wednesday, Dec 14, 2022 for INF1340H

Introduction

Analysis on the tables from the UN dataset ‘Trends in International Migrant Stock: The 2015 Revision’ were done using EDA (Exploratory Data Analysis) which included sorting, grouping, subsetting, and visualizing the data to expose potential patterns. The three tables were subsetted to view the data for International Migrant Stock, Annual rate of change between migrant and refugee stock, and Estimated refugee stock.

After looking into the initial descriptive statistics of each table, visual analysis was performed using the following techniques:

- **Horizontal bar chart**
  - To view counts of international migrant stock across countries and by sex
  - To view counts of estimated refugee stock across countries
- **Line graph**
  - To compare annual rate of change between migrant and refugee stock
- **Box plot**
  - To better view differences in sex distribution for international migrant stock
- **Small multiples**
  - To compare granular changes in international migrant stock country to country by variances in sex

These visualizations surfaced high-level findings across the dataset which include:

- Clear differences in international migrant stock versus refugee stock based on general geographic location
- Unequal distribution of sex for international migrant stock over the years 1990-2015
- Different rates of change between migrant and refugee stock

These findings sparked further research questions to better understand why there are such stark differences in count data for migrants and refugees for specific countries. And what factors may be involved in impacting the unequal distribution of migrants on a global scale.

Methods & Results

Mid-term Data Clean up

Before beginning the exploratory data analysis, I did some slight clean up and edits to my UN dataset from the mid-term. The following changes were made:

- I double checked that the **data types were correct** for each variable (e.g. Annual rate being attributed as an float data type versus an object)
- I **dropped any duplicate rows** that may have carried over through merging
- **Added back in the country code**; in order to subset just countries or a global view later on
- I double checked that there were **no missing rows** through the merging of various tables

Approach

After cleaning up the mid-term data, I reviewed the general steps for Exploratory Data Analysis (**Sort, Group, Subset, Compare**). I decided to approach the analysis using a **'free-form' method** to cover all EDA steps versus going through each step one by one.

1. Review descriptive statistics about each table

To recall, my three tables that I generated for the mid-term subsetted data on:

- Table 1: International Migrant Stock
- Table 2: Annual Rate of Change between International Migrant Stock and Refugee Stock

- Table 3: Refugee Stock

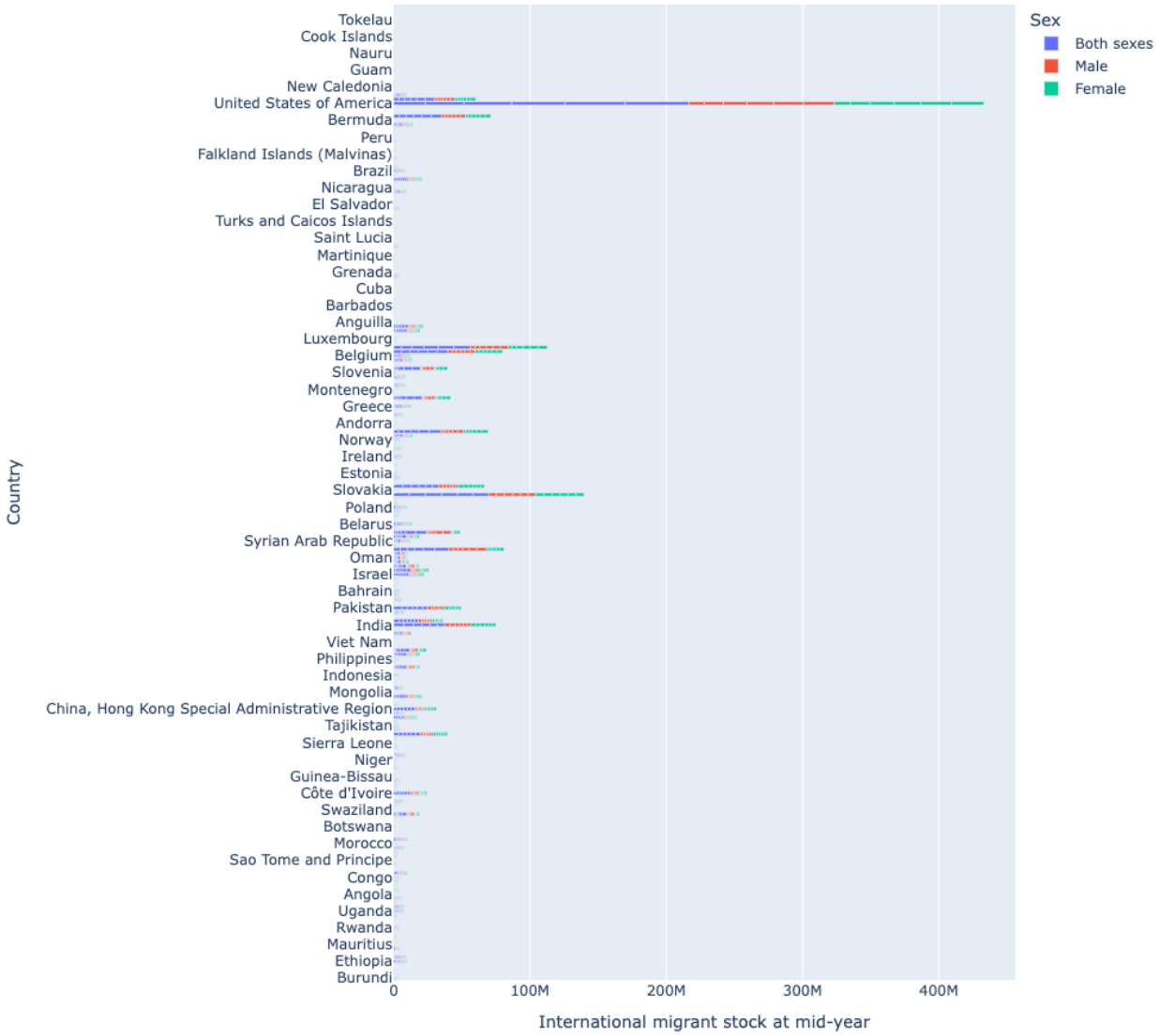
By just reading the mean, min and max values I could tell that there would be some interesting analysis that could be done and shown through visualization later on.

2. Sort and visualize countries by ‘International migrant stock at mid-year’

To see if there were certain countries that had the largest international migrant stock, I sorted the table and found that the United States of America had the first to fifth positions as the most migrant stock.

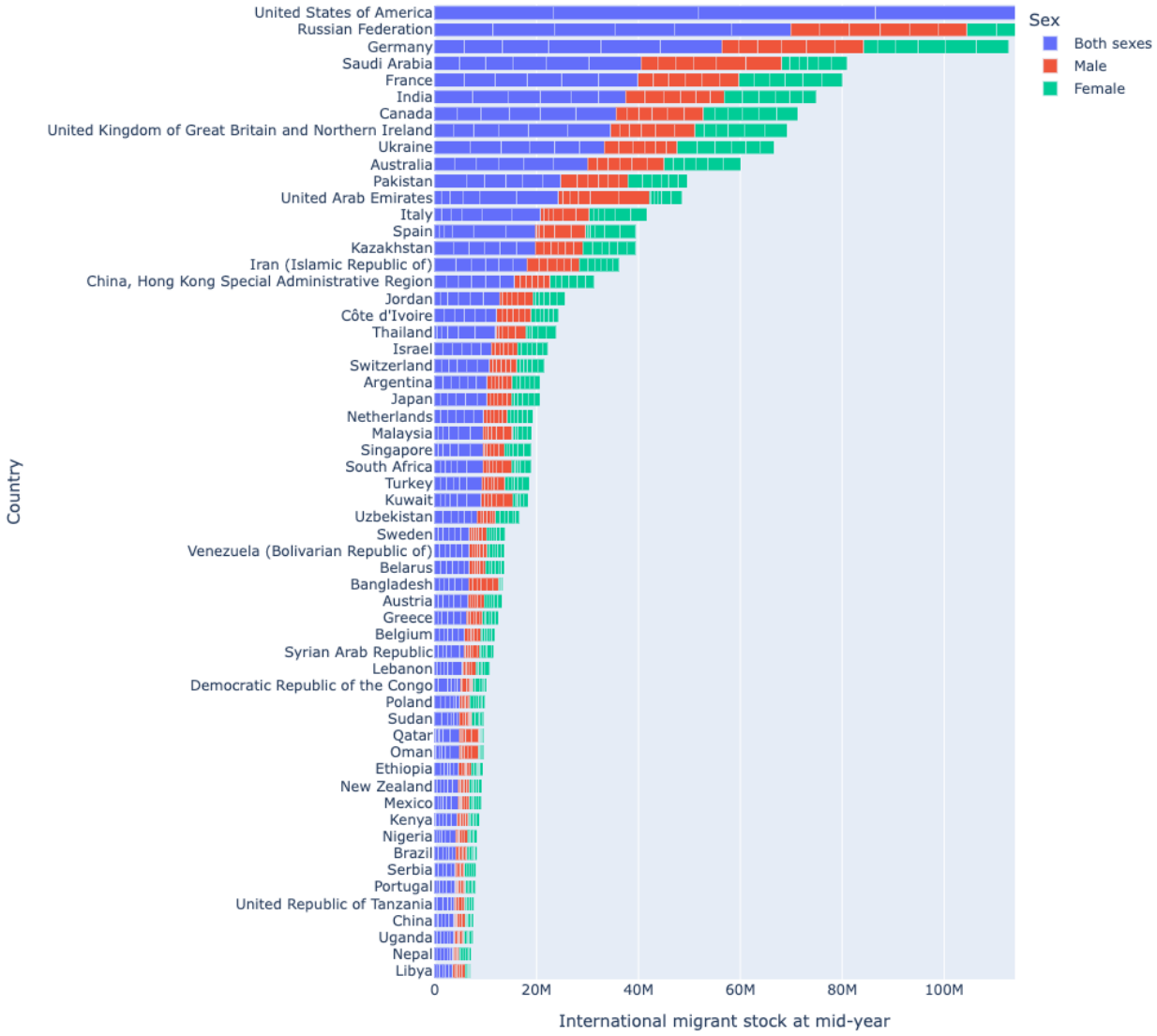
To explore how much more international migrant stock the United States had above other countries, I decided to create a horizontal bar chart using Plotly pulling in data from Table 1 (International Migrant Stock), with a view on difference in Sex for hue: **Fig 1: Global International Migrant Stock from Years 1995-2015**

Fig 1. Global International Migrant Stock from Years 1995-2015



Viewing this chart, one can see that the United States of America by far has had many more international migrant stock compared to others countries. When comparing sexes, it seems to be generally equal on a visual scale. To view the top countries that had the most migrants, I decided to sort the data to view by descending order:

Fig 1. Global International Migrant Stock from Years 1995-2015

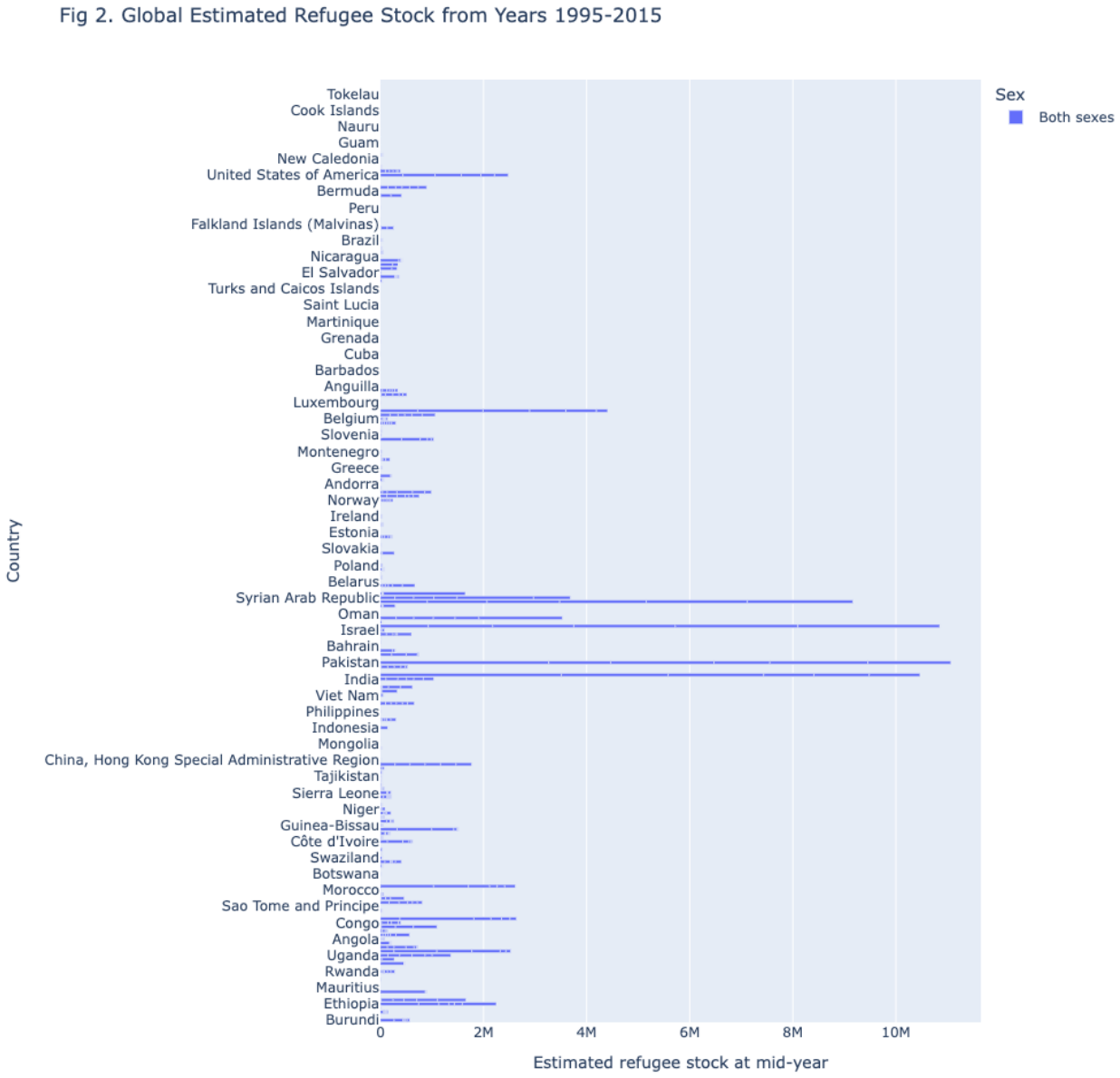


Here, it is much easier to view the top countries that have had the most migrants year to year. With the Russian Federation, Germany, Saudia Arabia, and France coming in 5th.

3. Sort and visualize countries by ‘Estimated refugee stock at mid-year’

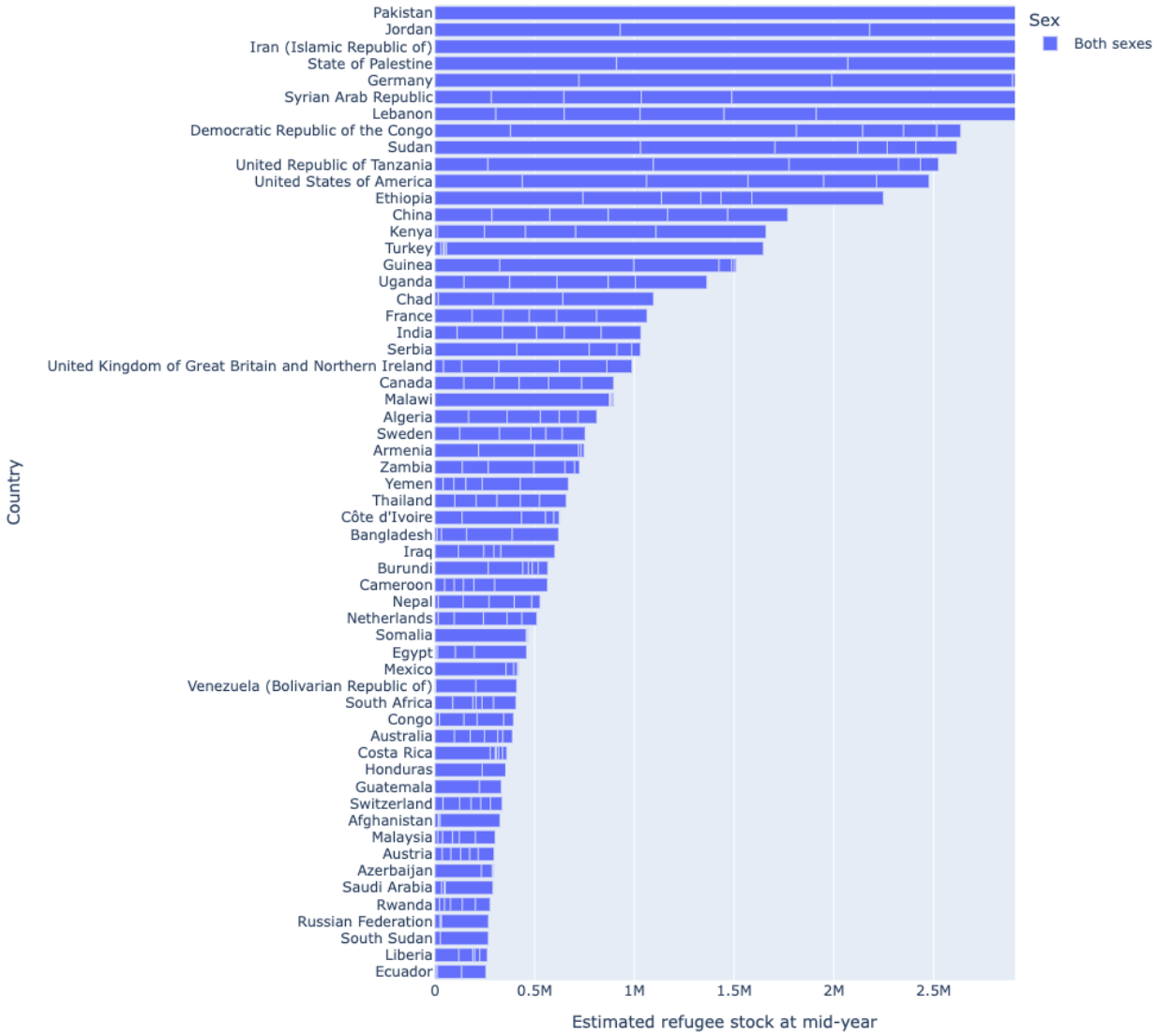
This sparked some curiosity as to whether the same or similar pattern would occur for the estimated refugee stock by country. So I created another Plotly horizontal bar plot to

analyze and visual patterns with that subset from Table 3 (Refugee stock): **Fig 2. Global Estimated Refugee Stock from Years 1995-2015**



At a high-level one can can see that there are different countries now showing the greatest refugee numbers. To have a better view; I also sorted this to show the top countries:

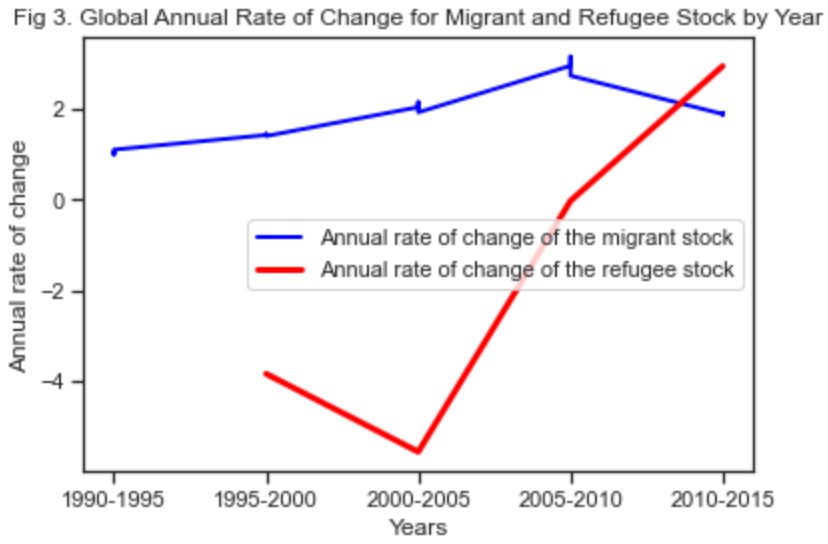
Fig 2. Global Estimated Refugee Stock from Years 1995-2015



4. Compare the annual rate of change between international migrant stock and estimated refugee stock

One can see here that there is now a clear pattern with westernized countries having the most migrants, and Asian and Middle Eastern countries having the most refugees. This sparked some interest to compare whether the annual rate of change between international migrant stock and refugee stock was similar or different.

To do this, I created a line plot using Matplotlib to map the two annual rates of change against each other in different colours across the years 1995-2015. Because the years were set as ranges, I changed these to a single numeric year to map it on the x-axis. For a cleaner holistic view, I pulled in the data for the whole world versus by country:
Fig 3. Global Annual Rate of Change for Migrant and Refugee Stock by Year

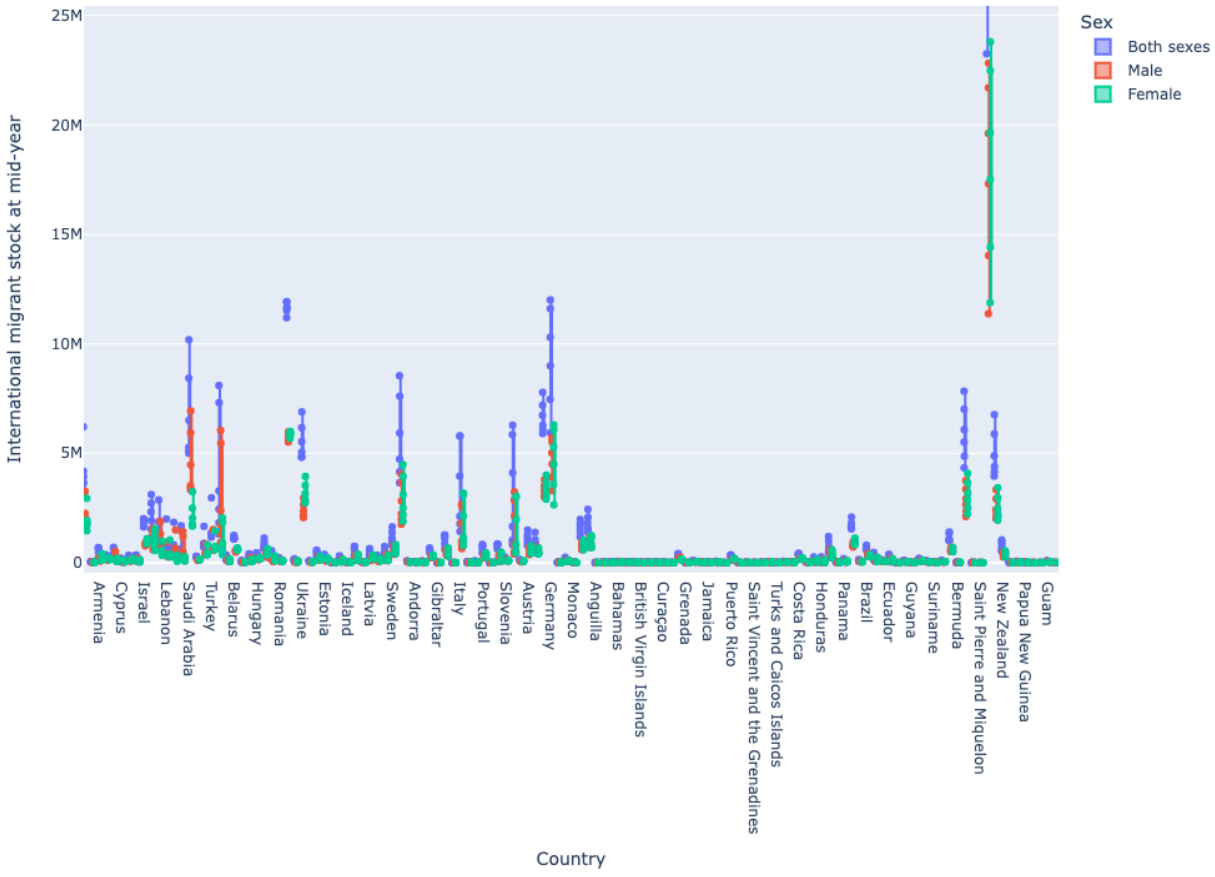


This was quite interesting because it shows that the rate of change for **migrant stock has more or less remained consistent** since 1990 to 2015. Yet the **refugee rate has heavily increased** since 2000 to 2015. This makes me wonder what kinds of events have been happening at a global scale since 2005 that is causing such a sharp increase in refugees.

5. Analyze whether there is a variance in Sex for International Migrant Stock

Returning back to international migrant stock, I decided to plot a boxplot using Plotly to see if there was a difference in Sex when it came to international migrant stock using data from Table 1. This was because the horizontal bar chart made it difficult to visualize variances: **Fig 4. Variance in Sex for International Migrants**

Fig 4. Variance in Sex for International Migrants

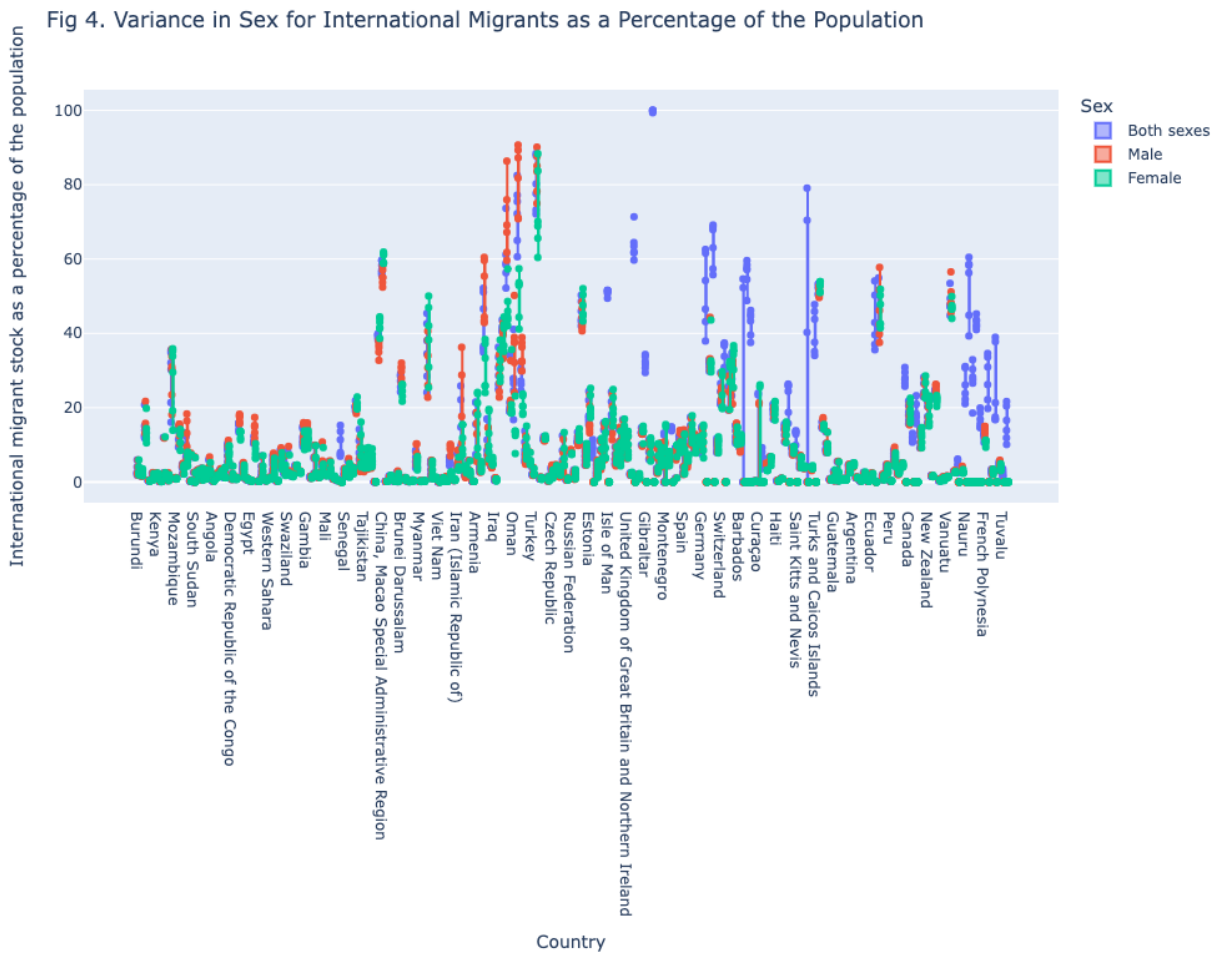


This graph (once zoomed in) shows that there are certain countries that have huge variance between migrants who are male versus women. Examples would be:

- Saudi Arabia
- Qatar
- Kuwait
- United Arab Emirates
- Ukraine

To see if this pattern still shows up when looking at international migrants as a percentage of the population; I decided to replot the same graph looking at the column

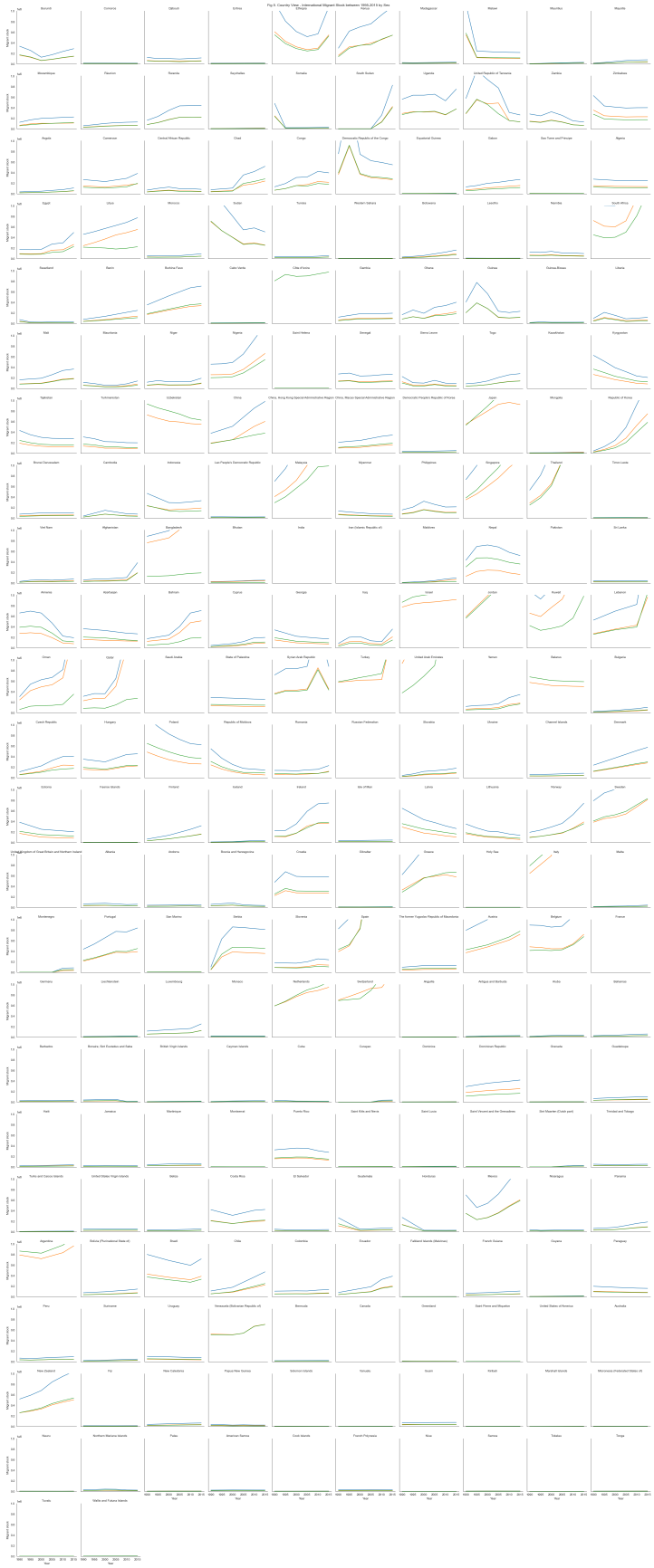
‘International migrant stock as a percentage of the population”: **Fig 5. Variance in Sex for International Migrants as a Percentage of the Population**



Interestingly, when viewing the same countries zoomed in, this view showed **an even greater inequality** when comparing male and female migrants.

6. Further analyze country differences using small multiples

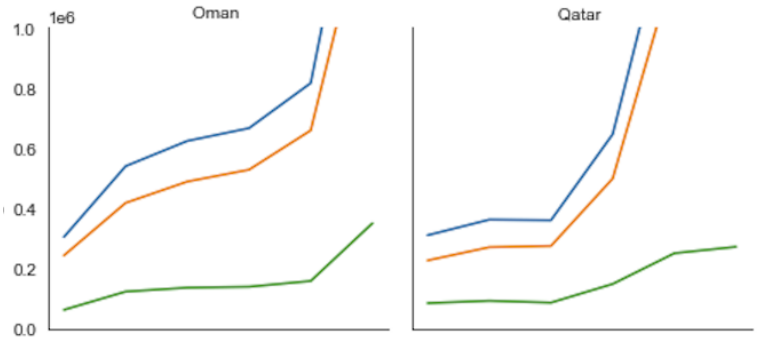
To take a deeper look at the countries with the greatest inequality between Male and Female international migrant stock, I decided to visualize a small multiples chart using a Seaborn Facet Grid with the same data from Table 1: **Fig 6: Country View - International Migrant Stock between 1995-2015 by Sex.** *(Legend: Blue = Both Sexes, Orange = Male, Green = Female)*



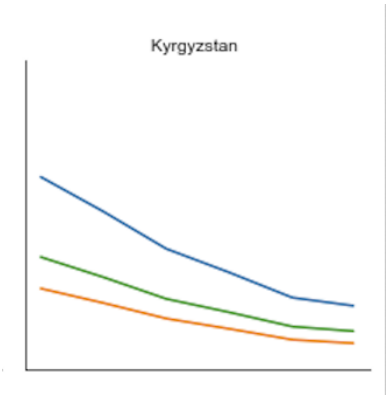
This view allows an even easier way to see which countries has the most unequal distribution across sex for international migrants. Or countries that show an equal distribution as increased or decreased. Examples:

Increase in unequal distribution (Oman and Qatar)

The **orange** (Male) and **green** (Female) lines are shown closer together in 1990, but are shown increasingly further apart in 2015



Decrease in unequal distribution (Krygyzstan)



Overall when we look at the small multiples chart, we see overall, if there is a large increase in international migrant stock over the years, there seems to also be a correlated increase in inequality between sexes. Countries with little variance seem to also see little variance in sex distribution of migrants.

Discussion

There are a number of findings that are surfaced by the visual analysis done to the three tables (International migrant stock, annual rate of change, refugee stock). These include:

- Clear differences in migrants and refugees based on geographic location (country)
- Variances but overall unequal distribution of migrant stock based on Sex
- Low variance in rate of change for annual migrant stock
- High variance in rate of change for annual refugee stock

These findings spark further research questions that could potentially surface more interesting patterns and analysis. These include:

- What variables or events are triggering such a high increase in rate of change for refugees?
- Are there patterns in the groupings of countries where there are bigger increases in migrant stock versus refugee stock?
- Are there any cultural variables that we can keep in mind for these patterns?
- What are some of the reasons as to why some countries have increased variance in sex distribution for migrants
- Is there also variance for sex when it comes to estimated refugee stock?
- Are there certain countries that have similar increases in migration numbers when it comes to migrants and/or refugees?

These questions can offer a more robust understanding of the patterns that are occurring for migrants and refugees on a global scale. Migrants who leave the country based on their own whim will always have more time and space to plan their move. Refugees on the other hand are often forced to leave their country on an assumed basis of urgency. Based on these assumptions, could it be that the sex variance for refugees is not as prominent as it would be for migrant patterns since migrants are dependent on the culture and rights of all humans. It would be an interesting study to look into.

Noted Limitations

Based on technical limitations, I was not able to visualize some aspects of the data as I wished. This includes a view of a world map where I could compare at the same time international migrant stock numbers (using a heat map for example), versus the refugee stock (using circles).

This same view would have been interesting to see as a small multiples view as well to quickly compare countries and see which have both high counts of migrants and refugees, low counts, and differing counts.

Conclusion

Overall, the exploratory data analysis that was done on the UN Migrant Dataset shows that are clear geographical differences when it comes to migrant and refugee stock. We're able to see these differences by sorting, grouping, and subsetting the data in various ways to view it from different visual perspectives.

Using visualization techniques such as **line graphs** seem to be best when wanting to compare data on a continuous scale (such as rate of change). Whereas **bar plots and box plots** show distribution of data at a quick high-level scale which makes it easier to see which countries are worth looking more deeply into. **Small multiples** was helpful at looking at the data at a more granular scale; country to country.

These exercises made it much easier to explore the data and find potential patterns and research questions to take the analysis further.