Teresa Lau
INF1340 Final Submission

1. Introduction

As a continuation of the mid-term submission where steps were taken to clean the UN Population Division international migrant stock data published in 2015, this report uses exploratory data analysis (EDA) techniques to identify and uncover trends and patterns of international migrants and refugees across the world. With a human-centred approach to data science, this report will discuss the implications of the results, and how this analysis could contribute to influencing decisions that benefit society and people in need.

2. Data and Methods

The cleaned dataset consists of eight tables as opposed to the original six, covering international migrant statistics from 1990 to 2015. Information in the dataset include *international migrant stock* (Table 1), *total population at mid-year* (Table 2), *international migrant stock as a percentage of total population* (Table 3), *female migrants as a percentage of international migrant stock* (Table 4), *annual rate of change of migrant stock at mid-year* (Table 5). The original table 6 consists of three sets of observation: (1) *estimated refugee stock at mid-year*; (2) *refugee as a percentage of international migrant stock*; and (3) *annual rate of change of the refugee stock*. These observations were split into three separate data tables in the mid-term submission.

It is noted that countries recorded their international migrant stock differently. The number could refer to one or more, or all of the following categories: (1) foreign-born population; (2) foreign citizens; and (3) refugees. Some countries have no data on international migrants.

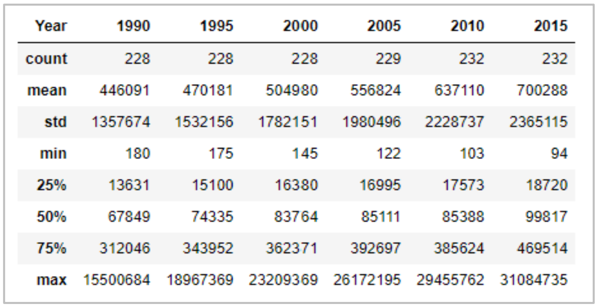
This report adopts methods of EDA which includes descriptive statistics and the use of visualizations to communicate findings. Key EDA principles proposed by Edward Tufte were used in this project, including (1) sort; (2) group; (3) subset; and (4) compare. Sort refers to re-ordering the data in ascendingly or descendingly to view smallest/largest data points. Group refers to categorizing the data by common attributes such as continents and regions. Subset means extracting the required data from the larger dataset to enable more focused analysis. Compare means using visualization to draw meaningful comparisons across categories, time, or other dimensions. In this project, the data is analyzed using EDA principles to unpack insights at the global, continental, and country levels.

3. Results

Summary Statistics of International Migrant Stock

Figure 1 is a table that details the summary statistics of international migrant stock of both sexes (table 1 in the UN dataset) at mid-year from 1990 to 2015. It shows that there is a general increase in international migrant stock as the median (50th percentile) grew from 67,849 people in 1990 to 99,817 people in 2015. As the means are larger than the medians across the years, it shows that there are outliers – countries with many international

Figure 1: Summary Statistics of International Migrant Stock (both sexes) at mid-year



Year	1990	1995	2000	2005	2010	2015
count	228	228	228	229	232	232
mean	446091	470181	504980	556824	637110	700288
std	1357674	1532156	1782151	1980496	2228737	2365115
min	180	175	145	122	103	94
25%	13631	15100	16380	16995	17573	18720
50%	67849	74335	83764	85111	85388	99817
75%	312046	343952	362371	392697	385624	469514
max	15500684	18967369	23209369	26172195	29455762	31084735

migrants – right skewing the distribution. In terms of dispersion, the standard deviation (std) of international migrant stock also increased from 1990 to 2015, showing that the distribution among countries in number of international migrants became wider. This is also evident in the changes of minimum and maximum values. The minimum values continued to decrease whereas the maximum values increased between 1990 to 2015. In sum, while this was an overall increase in the number of international migrants, some countries had gained more whereas others had experienced a decline, widening the distribution of international migrants across the world.

Trends of International Migrant Stock (1990-2015)

Besides the growth of international migrant stock, the world's population also increased from 1990 to 2015, totalling 7,326,091,061 people in 2015. Through sorting table 2 to look at the subset of 2015 data, it was discovered that the most populous country in 2015 was China with 1,376,048,943 people whereas the least populous one was the Holy See with 800 people. However, these two countries were not the ones with the highest/lowest numbers of international migrants. After sorting table 1, it was found that the United States of America had the largest number of international migrant stock with 46,627,102 people while Tuvalu had the smallest number with 798 people in 2015.

Figure 2 shows that after grouping the world's population by sex and comparing them in a line graph, both male and female's population increased between 1990 to 2015 at a consistent rate. Similarly, there is a growth in international migrant stock from 1990 to 2015 of both male and female migrants, reaching a total of 243,700,236 people in 2015. However, the growth rate of male international migrants outpaced that of female migrants starting from 2005, as indicated by the steeper slope of the line afterwards in figure 3.

While the growth in the total population could have contributed to the increase in international migrants as there were simply more people than before, it is noted that the upward trends of total population and international migrant were not the same. As seen in figure 2, the world's population increased at a steady rate from 1990 to 2005, whereas in figure 3, both male and female international migrant stock experienced an acceleration in growth rate after 2015, as the slopes of the lines became steeper beyond.

Figure 2: Line graph of the world's total population by sex (1990-2015)

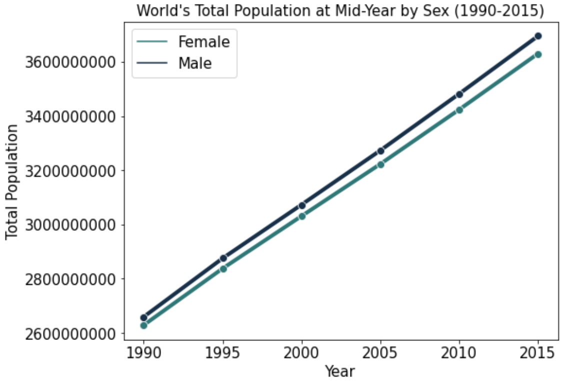
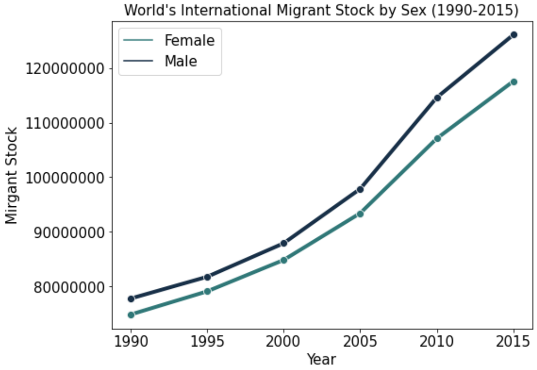


Figure 3: Line graph of the world's International Migrant Stock by sex (1990-2015)



The sudden increase in growth of international migrant stock prompted me to group the data differently by major areas (continents) to investigate in more detail. Comparing different major areas in figure 4, the growth of international migrant stock varied by continent. While Europe had long had the highest number of international migrant stock since 1990, international migrant stock in Asia had grown sharply since 2005, reaching almost at the same level as Europe in 2015. International migrant stock in other continents also increased from 1990 to 2015, with Europe and North America showing steady growth while Africa, Latin America and the Caribbean, and Oceania seeing minimal growth.

The substantial increase of international migrant stock among Asian countries piqued my interest in investigating the potential source of the growth by sub-region within the Asian continent. Using the principle of subset in EDA, I selected the Asian regions and plotted a line graph (figure 5) to compare changes in international migrant stock across 1990 to 2015. Figure 5 shows that the region of Western Asia was driving most of the growth observed in Asia, as indicated by a shaper increase in the slope of its line since 2005.

Figure 4: Line graph of international migrant stock (both sexes) at mid-year by major area from 1990 to 2015

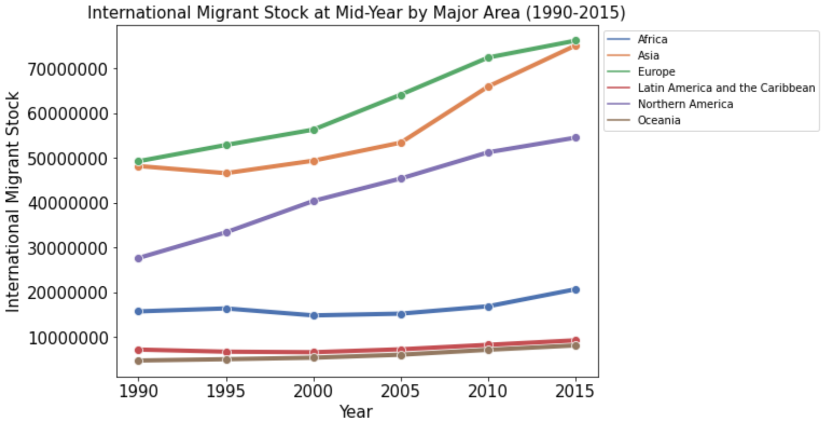
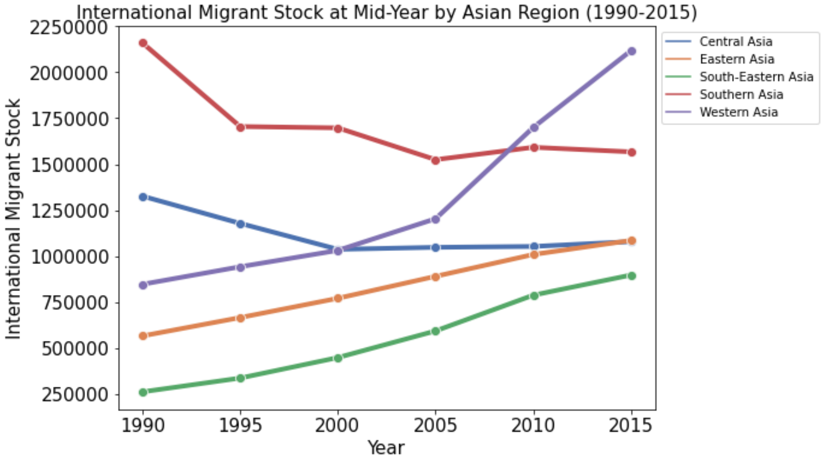
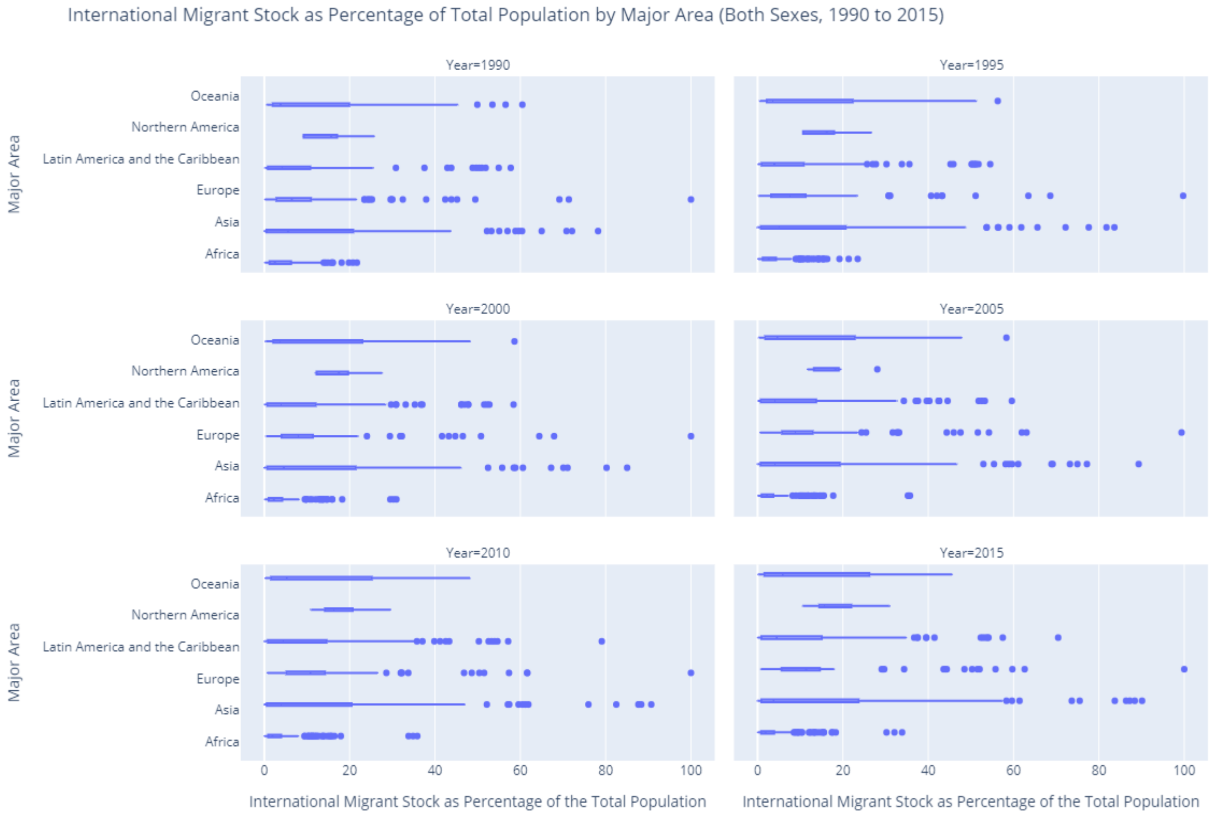


Figure 5: Line graph of international migrant stock (both sexes) at mid-year by Asian from 1990 to 2015



In addition to experiencing a strong growth in the number of international migrant stock throughout 1990 to 2015, Asia also had a wide distribution of percentage of international migrant stock as a share of its total population, as compared to other continents. Figure 6 contains a series of small multiple boxplots that indicate the distribution of the percentages across all major areas by year. Figure 6 shows that the distributions of percentage of international migrants as share of its total population were largely consistent from 1990 to 2015 across all major areas. In general, Oceania, Latin America and the Caribbean, and Asia had a wider variability of percentages of international migrant among their countries, as indicated by longer boxes and whiskers. On the other hand, the variations within North America's percentages were the smallest across the years, as indicated by the narrowest boxes and whiskers. It is worth noting that Europe has a relatively clustered distribution at the middle with its shorter boxes and whiskers in figure 6, however, its distribution has a long tail towards the right with outliers. One specific outlier present in all boxplots in figure 6 for Europe is as extreme as 100%, which means that all the country's population was international migrants. After sorting the data, I discovered that the country whose population was entirely made up by international migrants was the Holy See, also known as the Vatican City. The unique geopolitical characteristics of the country, where it serves as the governing centre of the Roman Catholic Church might have contributed to its population fully made up of international migrants.

Figure 6: Small multiple boxplots of international migrant stock as percentage of total population by major area (both sexes; 1990-2015)



Grouping the countries differently by developed and non-developed regions in figure 7, across 1990 to 2015, developed region had higher median percentages of female migrants as part of their international migrant stock. Countries in the developed region also had a smaller variability of percentage of female international migrants, as indicated by the shorter span of the boxes and whiskers in the boxplot. On the contrary, percentages of female international migrants varied more in the non-developed region, as shown by the longer boxes and whiskers and more outliers in figure 6. In sum, the developed region had a smaller variability in its distribution of female migrant percentage, with the data clustering between 45%-60%, whereas non-developed region had a wider distribution of percentage of female migrants, clustering between 35% to 60%.

Figure 7: Boxplots showing the percentage of female migrants as percentage of international migrant stock

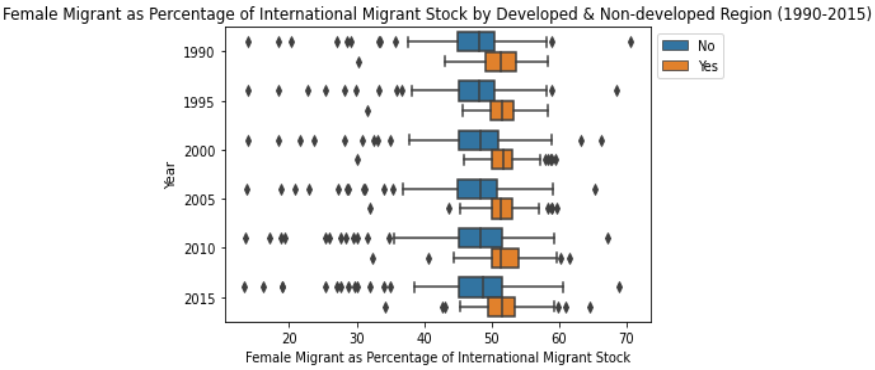
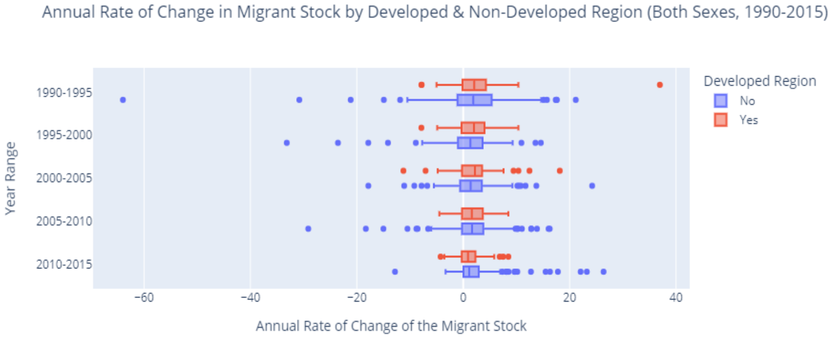


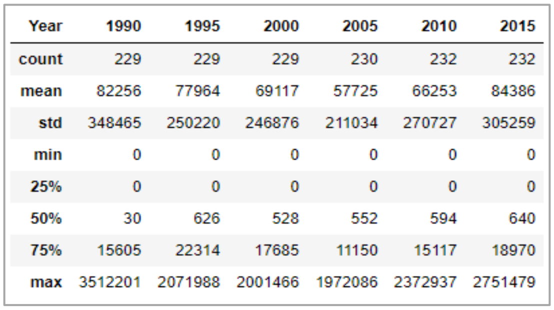
Figure 8: Boxplots showing the annual rate of change in migrant stock by developed and non-developed region from 1990 to 2015



Furthermore, comparing the annual rate of change in the number of international migrant stock (both sexes) between developed and non-developed regions in figure 8, it was discovered that the rate of change varied more for the latter versus the former. Figure 8 shows that while the 25th percentile, median, and 75th percentile of the annual rate of change for developed and non-developed regions are around 0% across the years, the boxes and whiskers of the non-developed region are longer and have more outliers. These characteristics indicate that the non-developed region had more substantial positive and negative fluctuations in the annual rate of change from 1990-2015. In contrast, developed countries experienced less of the drastic changes, except for Serbia between 1990 to 1995 with a 37.0% increase, which was discovered by sorting the data descendingly.

Summary Statistics of Refugee Stock

Figure 9: Summary statistics table of estimated refugee stock (both sexes) from 1990 to 2015



Year	1990	1995	2000	2005	2010	2015
count	229	229	229	230	232	232
mean	82256	77964	69117	57725	66253	84386
std	348465	250220	246876	211034	270727	305259
min	0	0	0	0	0	0
25%	0	0	0	0	0	0
50%	30	626	528	552	594	640
75%	15605	22314	17685	11150	15117	18970
max	3512201	2071988	2001466	1972086	2372937	2751479

The second part of the analysis focuses on refugee stock, which is a sub-group of international migrants recorded in the UN dataset. People are often dispersed as refugees because of warfare or natural disasters. Thus, some countries might have no refugees at all, resulting in a significant range between the maximum and minimum values across 1990 to 2015 observed in figure 9. In figure 9, the means of number of refugees are larger than medians across all years, indicating that some countries had substantially more refugees that skewed the distribution to the right. By sorting the subset of the data to focus on 2015, I discovered that the country with the highest number of refugees was Jordon with 2,751,479 people, followed by the State of Palestine with 2,051,096 people and Lebanon with 1,617,179. These three countries are all part of the Western Asia region and geographically close. The significant number of refugees observed in these countries in 2015 coincided with the political unrests in the Middle Eastern region around that time, demonstrating the adverse impact of social instability.

Trends of Estimated Refugee Stock (1990-2015)

After grouping the data by major area and comparing them in a line graph in figure 10, I observed that while most major areas had a minimal increase or decline in refugee stock between 1990 to 2015, Asia experienced a notable increase in refugees since 2005. Refugee stock in Africa was on a steady decline from 1995 but had an uptick between 2010 to 2015. Taking a subset of Asian region to compare in the line graph in figure 11, I found that Western Asia was driving the increase in refugee stock in Asia between 2005 to 2015. The growth in refugee stock in Western Asia counteracted the decline of that in Southern Asia, while other regions in Asia experienced minimal change in refugee stock.

Figure 10: Line graph showing estimated refugee stock by major area from 1990 to 2015

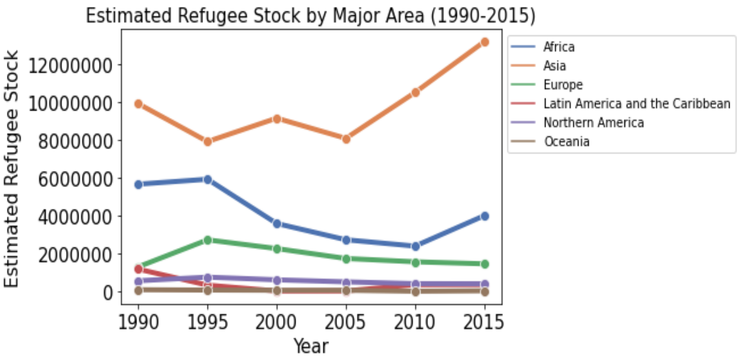


Figure 11: Line graph showing estimated refugee stock by region in Asia from 1990 to 2015

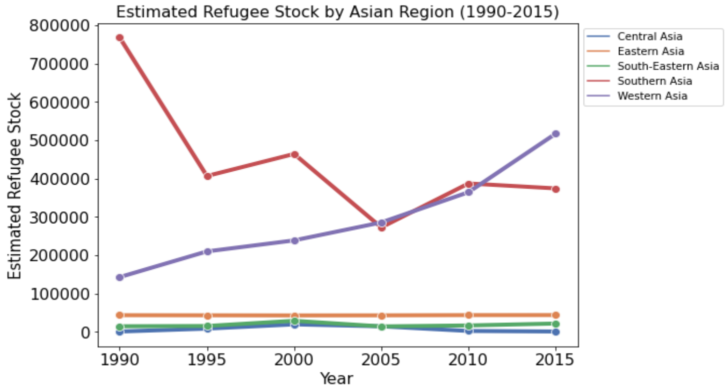
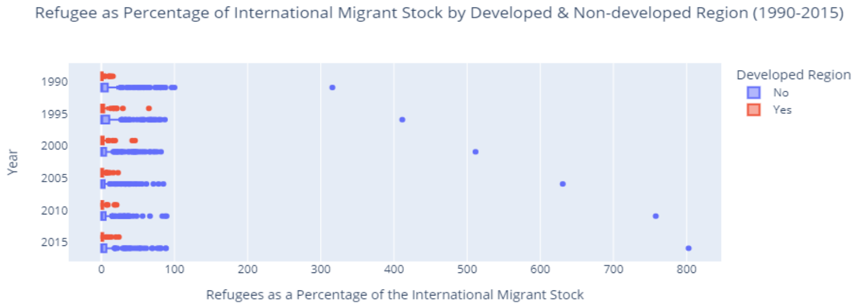


Figure 12: Boxplot showing the distribution of refugees as percentage of international migrant stock by developed and non-developed region from 1990 to 2015 (without removing outlier).



When grouping the data differently into developed and non-developed regions to observe the distribution of percentages of refugees as part of the international migrant stock, an extreme outlier was observed across all years (see Figure 12). After sorting the data in descending order, I found that the State of Palestine consistently recorded over 100% of its international migrant stock as refugees. It is unclear as to if this outlier was correct or a mistake made during the data collection process, but after cross-checking with the original dataset I can confirm that it was not a mistake made during the data cleaning process.

Therefore, data of the State of Palestine were removed from the dataset to create figure 13, which provides a clearer depiction of the distribution of percentages of refugee as part of the international migrant stock by developed and non-developed regions. Across all years, the non-developed region had wider variability in the percentage of refugee in their migrant stock, as

Figure 13: Boxplot showing the distribution of refugees as percentage of international migrant stock by developed and non-developed region from 1990 to 2015 (after removing outlier).

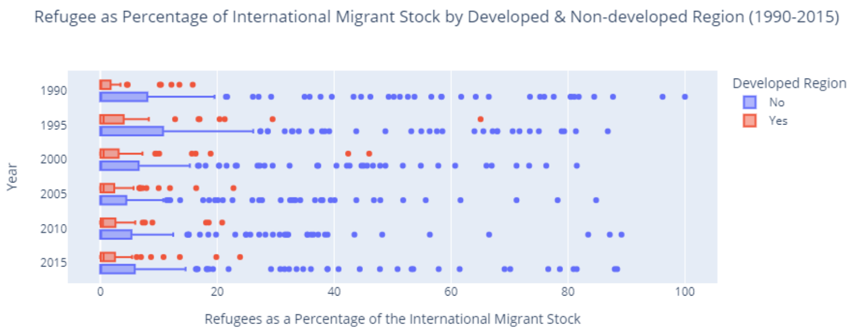
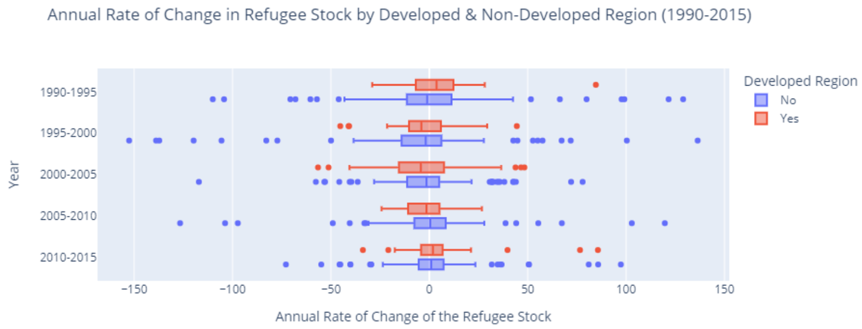


Figure 14: Boxplot showing the distribution of annual rate of change in refugee stock by developed and non-developed region (1990 to 2015).



shown by longer boxes and whiskers compared to those of the developed region. Moreover, the non-developed region had more outliers towards the higher percentages, meaning that there were more countries with substantially higher percentages of refugees in their international migrant stock.

As seen in figure 14, the medians of distributions of the annual rate of change in refugee stock were similar for both developed and non-developed regions across the year. However, the non-developed region had wider distributions of changes from 1990 to 2015, as evident in longer boxes and whiskers and more outliers on both ends. It means that the non-developed region experienced more fluctuations in the rate of change in refugee stock on both ends, and these fluctuations tended to be more drastic as compared to the developed region.

4. Discussions

Using Tufte's principles of EDA, this report identified major trends of international migrant by grouping the data by major areas (continents), isolating regions within the continents as a subset, as well as sorting migrant stock numbers to find the countries with the highest number of international migrants. By comparing the numbers across major areas and between developed and non-developed regions, it is discovered that the distribution and growth of international migrants differ across major areas and between developed and non-developed regions. Similar patterns were observed in the trends of refugee stock. The non-developed region continued to experience more fluctuations in refugee stock. Countries in non-developed regions were more likely to have a higher percentage of refugees in their international migrant stock.

These differences are not to be neglected, especially with the trends of growing number of refugees and the overrepresentation of refugees in non-developed region. In their book, Aragon et al. (2022) argue that learning the context of the data is imperative to the practice of data science. That means understanding the “information about users, groups, situations, and environments that are outside of the data but associated with [it].” (Aragon et al., p. 94, 2022). Putting the data back into context, data scientists must actively reflect on the implications of trends observed in the analysis. The trends of growth in the number of refugees and the overrepresentation of them in certain regions should concern data scientists. The dataset tells a sobering story of how in some parts of the world, people were disproportionally impacted by warfare, social instability, and natural disasters that they had to escape from their home countries to seek refuge. Rather than stopping here with the analysis, data scientists’ work continues with telling this story effectively to concerning parties (Aragon et al., p. 130, 2022). Thus, taking a human-centred approach to data science, the results of this report should be communicated effectively to influence governments and not-for-profit organizations to direct resources to regions and people that need help.

Reflecting on my own approach to data science in this project, I went through a few iterations with the data cleaning process to arrive at a curated set of data that could be used for analysis. Throughout the analysis process, various approaches were used to attempt at analyzing the data by grouping the data differently or taking a different subset. The most impactful insights were chosen to be included in this report to tell the story about changes and variation of international migrant stock and refugees in different parts of the world from 1990 to 2015. The next step of this project would be to share it with others and continue to explore other approaches to analyzing the data, as data science is an iterative process that involves collaborations.

5. Conclusion

Using the principles of EDA proposed by Tufte – sort, group, subset, and compare – this report shows the trends and distribution of international migrants and refugees with summary statistics and visualizations. The report highlights that the growth in international migrants and refugees was more prominent in some parts of the world versus the other, and the distribution of such growth rates varied across continents as well. While international migrants could refer to people who were born outside of their current residing countries or foreign-citizens, refugees are people who seek asylum elsewhere because of the unfortunate events that happened in their home countries. Therefore, with a humen-centred approach to data science, the findings of this report are not to be dealt with negligence. Instead, they should be used to influence decisions that direct resources to provide additional support to people living in these countries.

6. Reference

Aragon, C. R., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-centered data science: An introduction*. The MIT Press.