INF1340 Final Project Write Up

YUETONG JIANG
Student #: 1004339343

DECEMBER 15, 2022

Table of Contents

Table of Figures

Introduction

Data visualization plays an essential rule in data science cycle where the visualizations create graphic representation of the numerous and complex data and contributes to both the initial exploration phase of the project and the result communication phase after analysis. As proposed by John Tukey since 1970, visualizations and statistical graphics assist the exploratory data analysis (EDA) to summarize the main characteristics embedded in the data (Leinhardt, 1980). The process of EDA can lead to new valuable research questions and promote further data collection, experiments and analysis.

Data visualization helps communicate the analysis process and insights discovered to the masses in need. Then it's important to avoid generating 'chartjunks' as coined by Edward Tufte, referring to the non-informative and confusing visualizations. By Tufte's visualization principles, the data visualizations should pay extra attention on: Comparisons, Causality, Multivariate, Integration, Documentation, and Context (Tufte, 2001). Moreover, Tufte highlights the terms of graphical integrity, lie factor, data-ink ratio, layering, small multiples, multifunction, and the data density of a graphic to emphasize the illustration of the substance of the graphic and data where every data point contributes effectively at both the Marco and Micro levels, instead of blindly overlaying excessive decoration in visual displays (Tufte, 2001).

In this paper, I will apply the concept of EDA and Tufte's visualization principles to assist data exploration and visualization on the data set 'Trends in International Migrant Stock' which contains five tables after tidy cleaning.

Method and Result

Method

The process of data exploration and visualization will be described from the macro to the micro, from the global dimension, gradually refined to the region, continent, and the individual countries. Three kinds of variables are taken into consideration, which are Years (1990-2015), Migrant Types (All, Refugee, Female, Male), Destinations (Main Regions, Continents, Individual countries. And four different types of data are visualized: International Migrant Stock, Total Population, Migrant as a Percentage of Total Population, Migrant Stock Annual Rate of Change. Python visualization libraries, such as *Matplotlib*, *Seaborn*, *Datascinece*, and *Plotly*, are used for creating graphics in an aesthetically pleasing and comprehensible manner. Six types of visualizations are applied, which are: Bar Plot, Box Plot, Violin Plot, Scatterplot, Histogram, and Line Plot. Each diagram has a different variable emphasis and help to answer and tell the story of the following two questions:

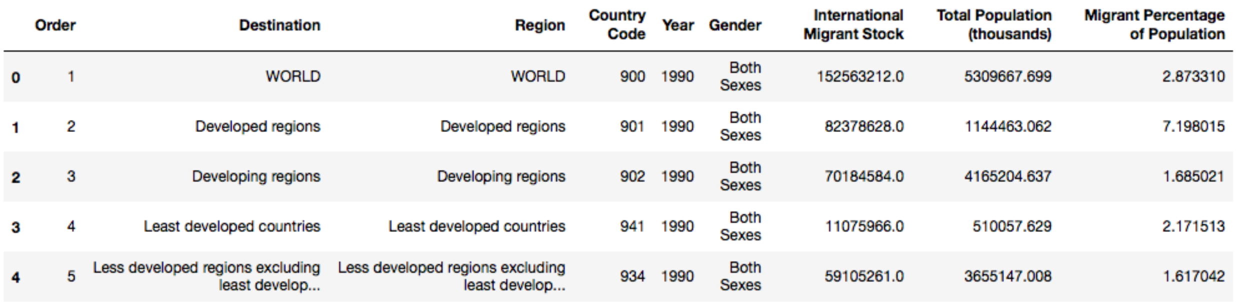
Question 1: Is there a gender gap among immigrants?
Question 2: Are there continent differences in immigration?

Besides, the data sets and tables were furtherly cleaned, rearranged or partitioned for the purpose of preparing and making suitable visualization.

Results

Data Overview

Recall that the joined_table contains the data of International migrant stock, Total population, International migrant stock as a percentage of the total population, at mid-year by sex and by destination, which is the combination of Table1, Table2, and Table 3.



	Order	Destination	Region	Country Code	Year	Gender	International Migrant Stock	Total Population (thousands)	Migrant Percentage of Population
0	1	WORLD	WORLD	900	1990	Both Sexes	152563212.0	5309667.699	2.873310
1	2	Developed regions	Developed regions	901	1990	Both Sexes	82378628.0	1144463.062	7.198015
2	3	Developing regions	Developing regions	902	1990	Both Sexes	70184584.0	4165204.637	1.685021
3	4	Least developed countries	Least developed countries	941	1990	Both Sexes	11075966.0	510057.629	2.171513
4	5	Less developed regions excluding least develop...	Less developed regions excluding least develop...	934	1990	Both Sexes	59105261.0	3655147.008	1.617042

Figure 1 First five rows of joined_table

For the purposes of data exploration, let's first take a look at the overall global migration stock to get a sense of trends. For easier comparison, we create a bar plot which overlaid the migrant stock onto the total population and a line plot for the migrant percentage cross years, such that the values are all in the same scale and we can easily detect the overall trend within each type of data.

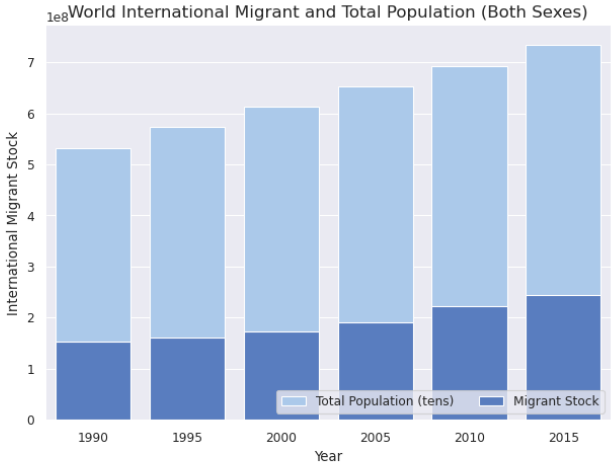


Figure 2 Bar plot for World International Migrant and Total Population (Both Sexes)

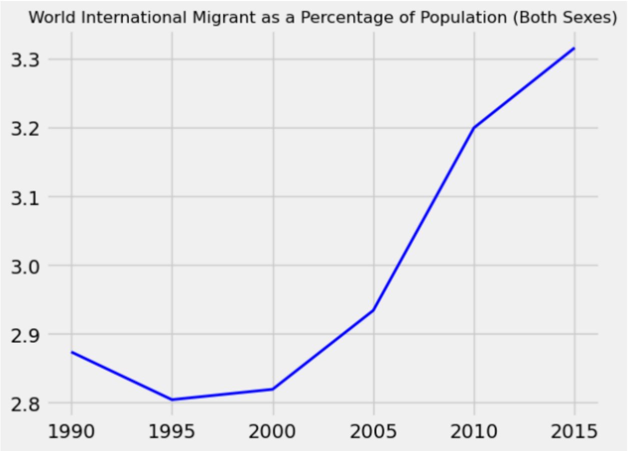


Figure 3 Line Plot for World International Migrant as a Percentage of Total Population (Both Sexes)

From Figure 1, we can find that both the amount of migrant stock and total population has increased every year, while the total population has a larger increase. But when we look at the migrant percentage trend in Figure 2, we can find a decrease between 1990 to 1995 that the proportion of migrants to the total population decreased by about 0.8 percent. Later, from 1995 to 2015, the percentage of migrants of the population kept increasing rapidly, especially during 2005 to 2010.

Question 1: Is there a gender gap among migrants?

To answer this question, let's also observe and explore the data at different levels. Figure 3 shows the histogram for international migrant percentage of population in 2015. As what we can find from this interactive plot, the distribution for Male and Female migrants are almost the same, where most of the countries have a migrant percentage between 0 to 17.5%, and there exists some outliers that the countries have migrant percentages close to 100%.

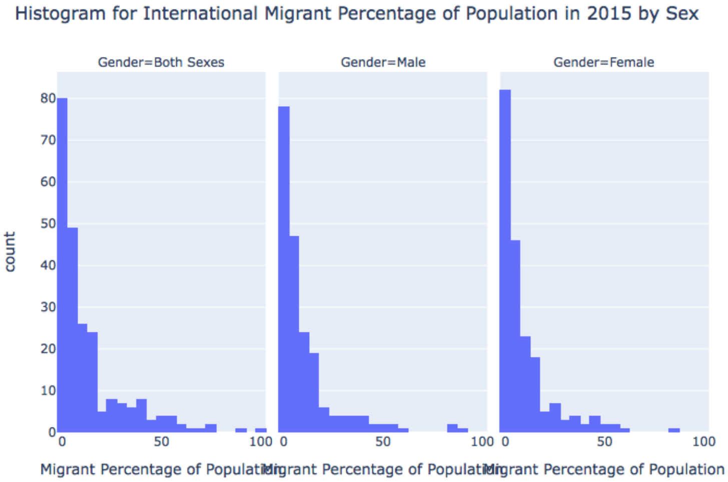


Figure 4 Facets Histograms for International Migrant Percentage of Population in 2015 by Gender

By Figure 5, we can find that the overall trend of international migrant stock is growing, where the amount/percentage of male migrants is always higher than which of the female migrants. Especially in 2015, the difference of Migrant Percentage between male and female is about 0.2%.

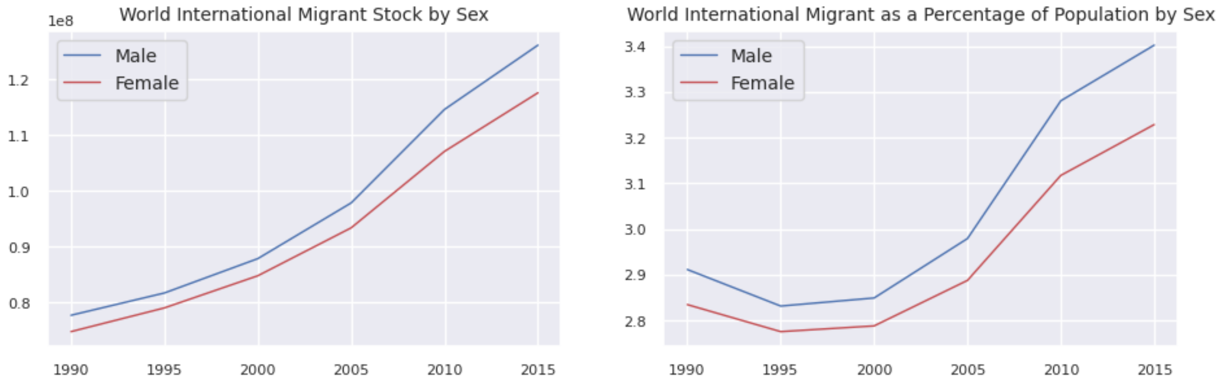


Figure 5 Line Plots for World International Migrant Stock and Percentage by Gender

Someone can easily claim that there is a gender gap among since the male migrant was much higher than that of women. However, when zooming in and examining within the smaller units, it appears a different result. In Figure 6, we partition the data set by the destination's developing state and calculate the average migrant stock cross years. The main two categories are: developed regions and developing regions, where developing regions contain least developed countries and other less developed regions. In Figure 7, we split the data more specifically into years based on the Tufte's Small Multiple principle. In these two plots, we found the migrant stock in developed regions is higher than which in developing regions. Moreover, female migrants is always higher than the male migrants in developed regions, and the other way around in developing regions.

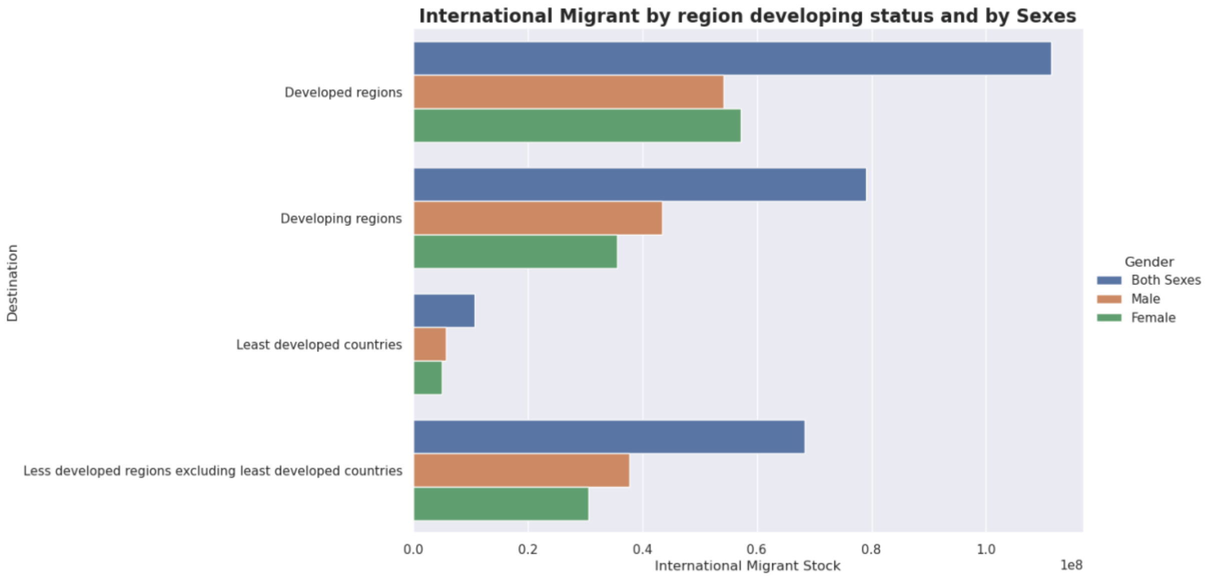


Figure 6 Bar plot for International Migrant by Destination State and Sex

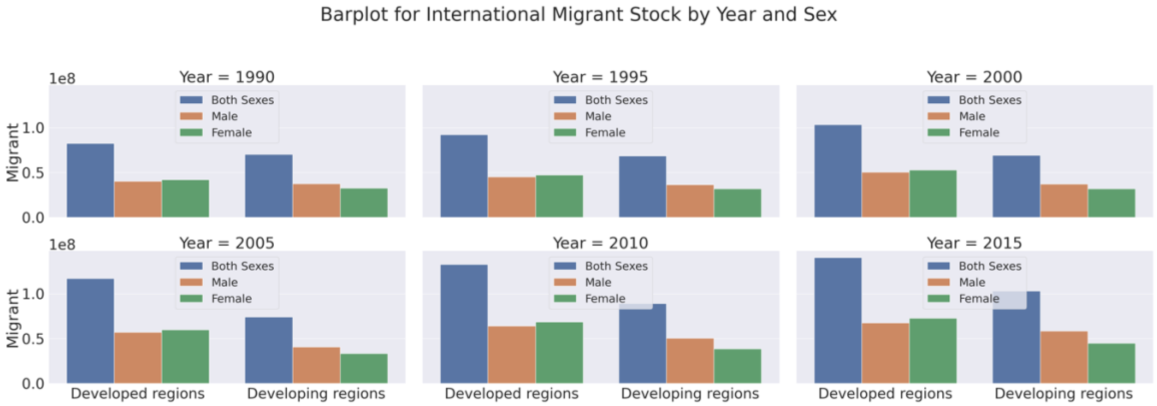


Figure 7 Small Multiple Bar plot for International Migrant Stock by Year and Sex

Now let’s take a closer look at the differences between 6 main continents by years. In Figure 8, it can be found that the migrant stock in Africa, Latin America and the Caribbean, and Oceania are quite low and stable, while migrant stock in the rest three continents increase progressively year by year. Besides, in Africa, Latin America and the Caribbean, Northern America, and Oceania, there is little different between gender. In Europe, the female migrant stock is slightly higher than the male migrant stock. However, in Asia, there is a more and more obvious differences between male and female migrant stock.

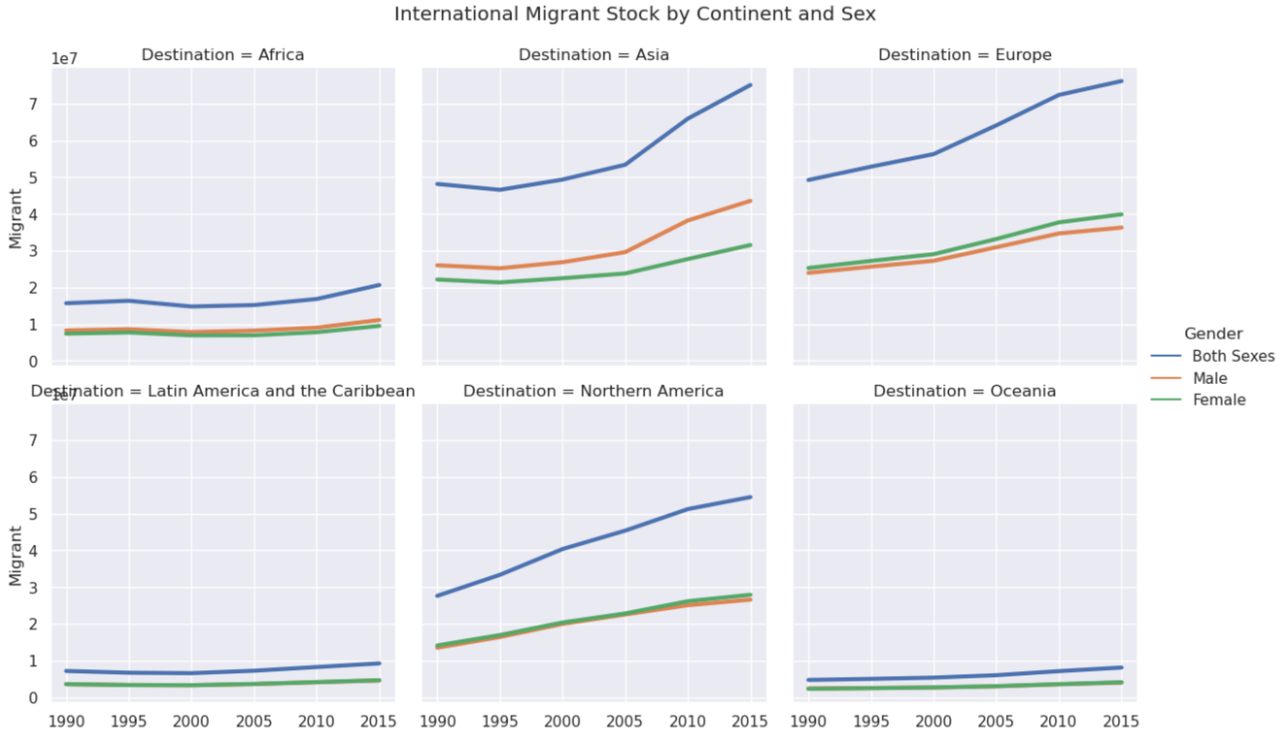


Figure 8 Small Multiple Line plots for International Migrant Stock by Continent and Sex

Furtherly, by the figure 9 below, we can find female migrant trends in different continents with more details by looking at the female migrant percentage of all international migrants. The boundary here is 50%, where the lines above 50% means the female migrant is higher than male migrant in the current year and current continent. We can find that in Africa and Asia, the female migrant percentages are always below 50% and showed a downward trend. In the rest four continents, the female migrant percentages are always above 50% with a relatively stable and slight upward trend.

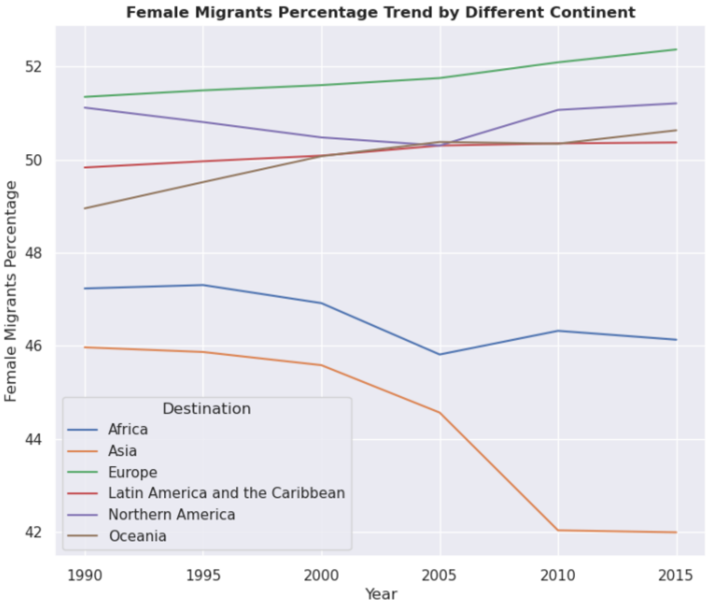


Figure 9 Line Plot for Female Migrants Percentage Trend by Continent and Year

From the scatterplot below, the small multiples for annual rate of change of female migrant stock, we can observe some yearly changes and trend in detail. The benchmark here is 0, where any points above 0 means an increasing trend, and any points below means a decreasing trend. In Europe, Oceania, and Northern America, the annual rate of change is always positive, while the rest continents have negative rates before 2000.

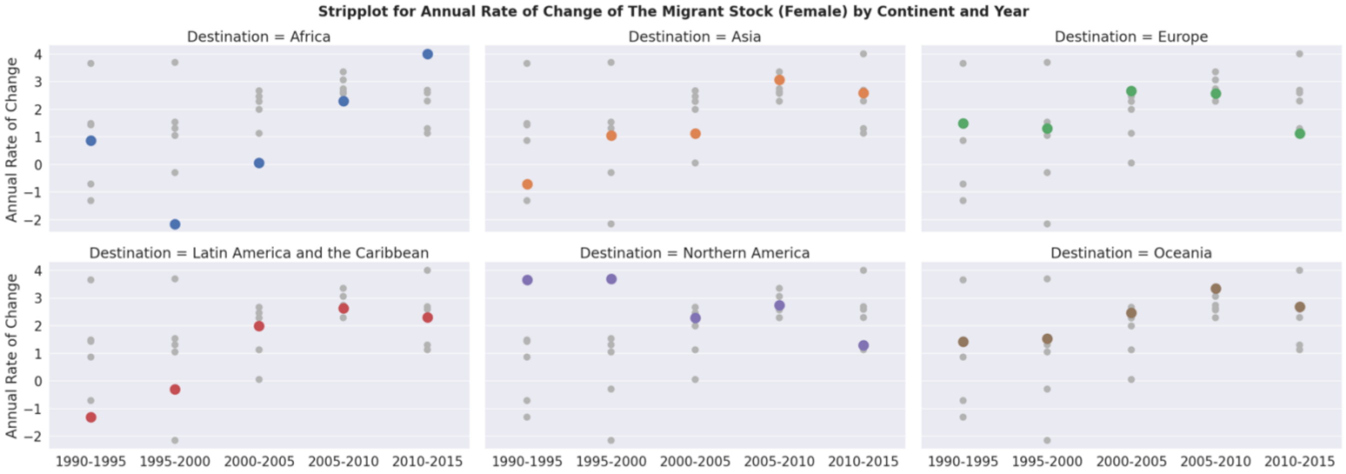


Figure 10 Small Multiples Scatterplot for Annual Rate of Change of Female Migrant by Continent and Year

By creating a lists of box plots in Figure 11, we can observe the distribution of Female migrant as a percentage of population for each continent. Here we averaged the values in each individual country by year. This plot provide some information that confirm with our previous plots. For example, the middle half countries in Africa has female percentage lower than 50%; middle half countries in Europe, has percentage above 50%. However, it can be found that Asia has a heavy tailed distribution that there are many countries have relatively large or small values. Even though the overall trend in Asia is downward, almost half of the countries in Asia still have female percentage above 50%. Similarly, even though the overall trend in Oceania is increasing, 75% countries have female percentage below 50%.

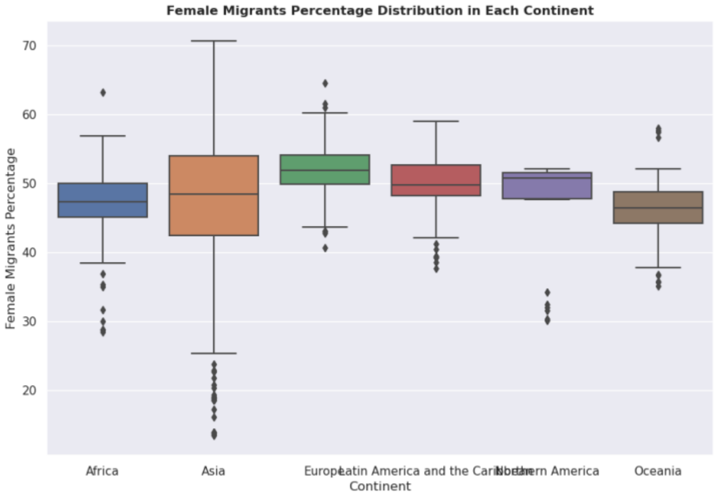


Figure 11 Box Plots for Female Migrant Percentage Distribution in Each Continent

Question2: Are there continent differences in migrant?

To answer this question, we first plot the overall migrant stock data by continents, as Figure 12 below. It’s interesting that even though the total population in Asia is about 6 times of total population in Europe, the International migrant stock in Europe is higher than which of Asia.

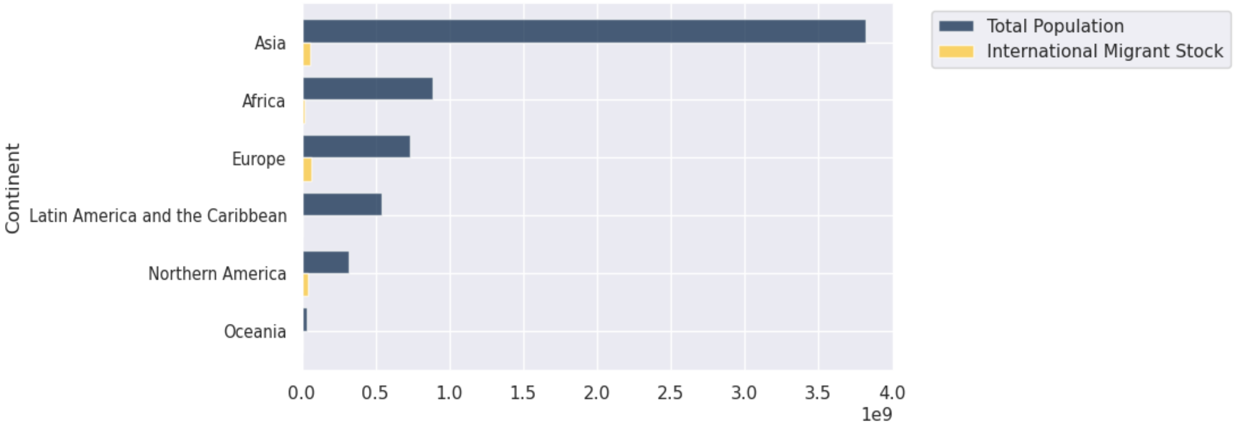


Figure 12 Bar plots for Total Population and International Migrant Stock by Continent (Both Sexes)

The boxplot below is not very clear in showing distribution details, but we can quickly observe that there is a significant difference between continents. Especially in Northern America, all the countries are migrant percentage significantly above 10 and concentrated. In order to make it easier for observation, we put the continents of similar scales in a group for comparison.

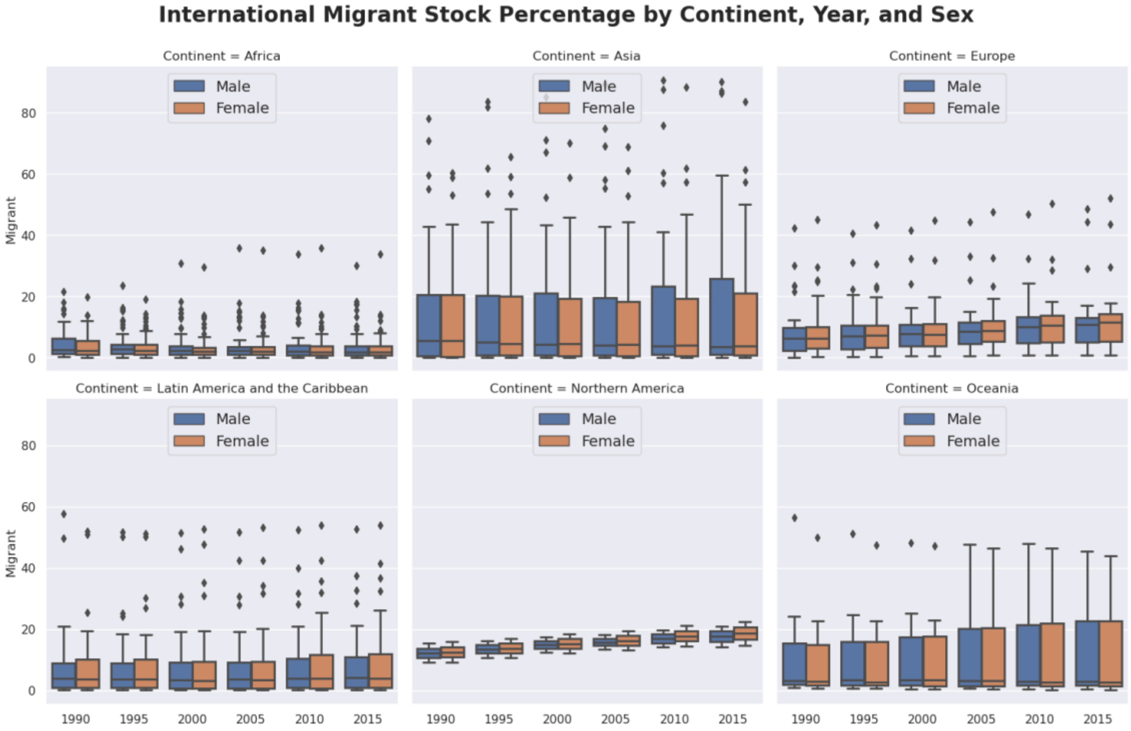


Figure 13 Migrant Stock by Continent, Year, and Sex

Let's look at North America separately. The boxplots in figure 14 and 15 shows an increasing trend in migrant stock percentage. Moreover, almost all countries in Northern America have similar and concentrated migrant stock values, but there exist one extreme value, United States of America, which may significantly pull the continent's overall value up.

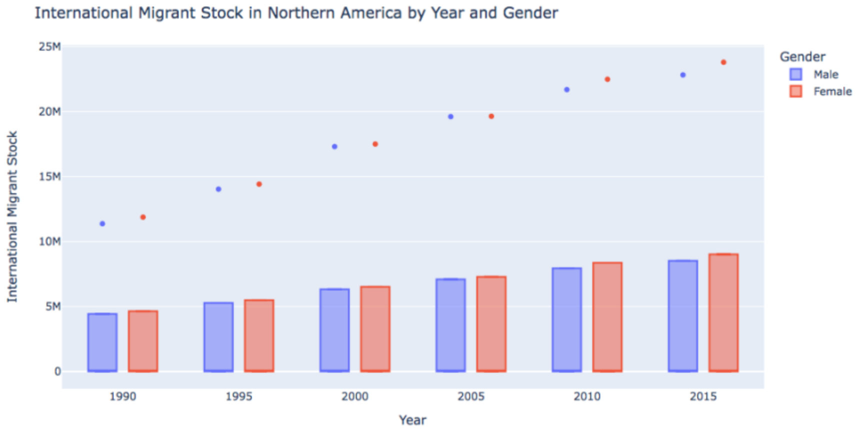


Figure 14 Boxplot for Migrant Stock in Northern America by Year and Gender

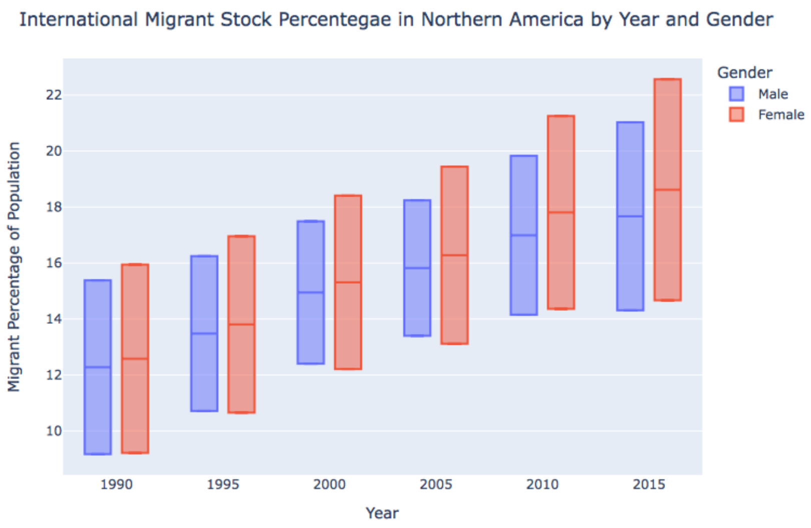


Figure 15 Migrant Stock Percentage in Northern America by Year and Gender

The figure below plots the distribution for Africa, Asia, Europe, and Latin America and the Caribbean. Africa and Latin America have similar distribution that data are more concentrated around 0 to 0.2, while Asia and Europe have similar positively skewed distribution that the overall data value is relatively large and exists some extreme values.

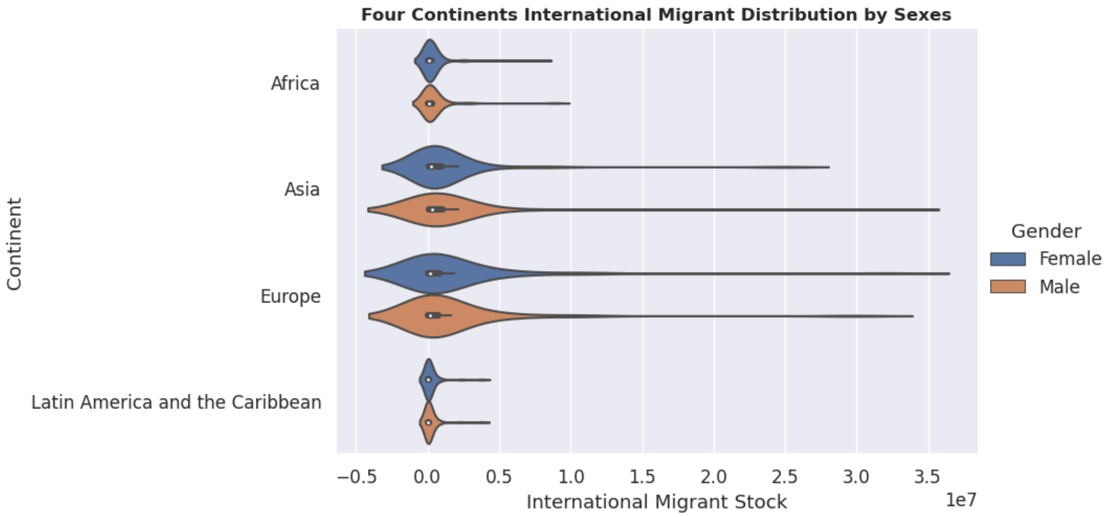


Figure 16 Violin Plot for Four Continents Migrant Distribution by Gender

As the boxplot for Asia, Europe, and Oceania showed, the scale and distribution of migrant percentage are very similar in Asia and Oceania. There exist more extreme values above the upper fence in Asia and Europe. The center of Europe migrant percentage is kind of far from 0 compared to the other two.

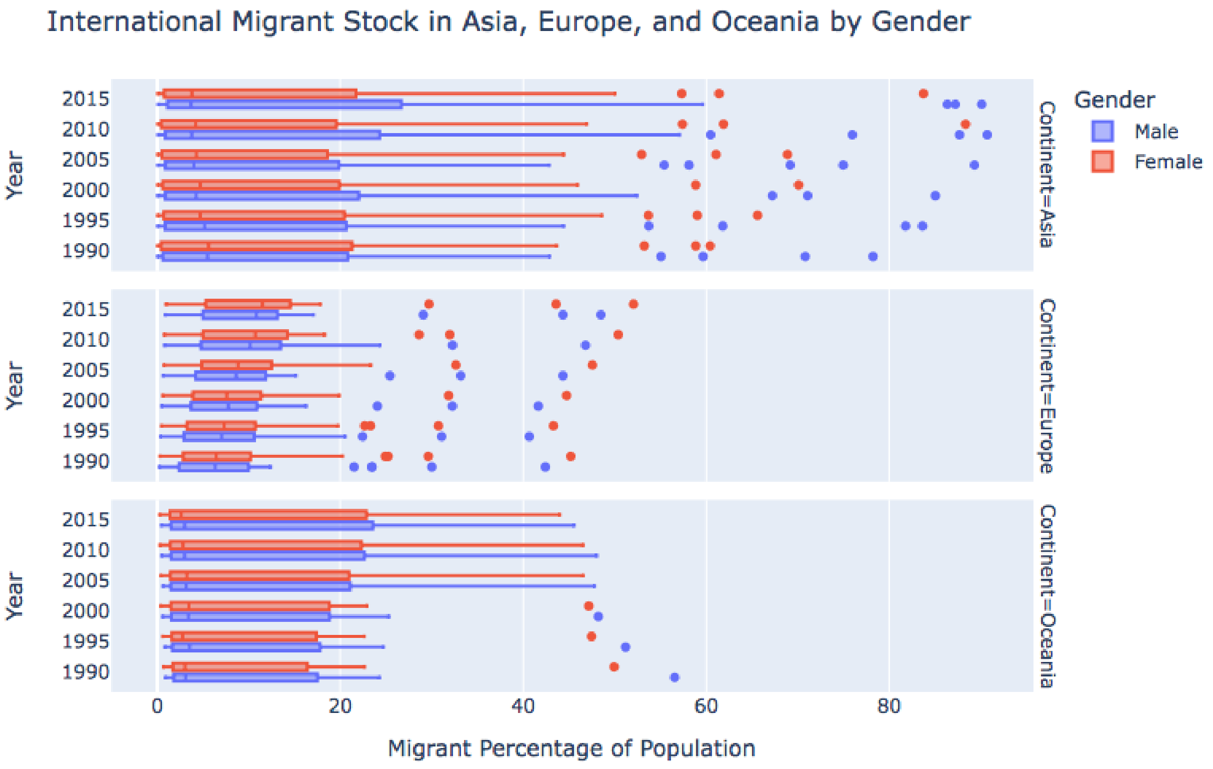


Figure 17 Boxplots for Migrant Stock in Asia, Europe, and Oceania by Year and Gender

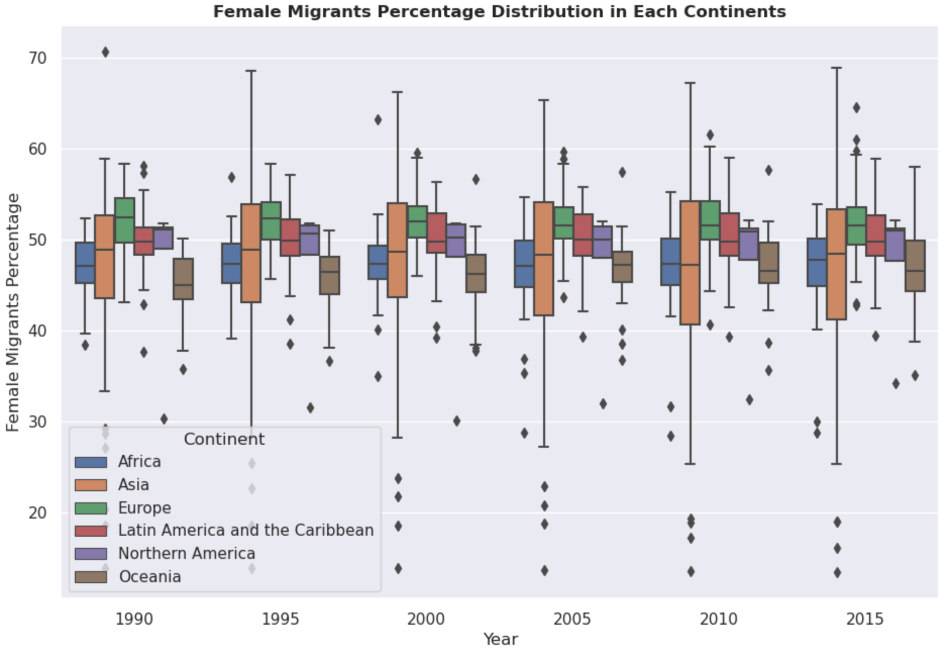
Conclusion and Discussion

For question 1, we found that the worldwide amount of female migrants and their proportion to the population is all significantly lower than male migrants. However, when we looked into the smaller groups and layers, the trends disappeared or even significantly reversed in some continents. This kind of phenomenon is also referred to the Simpson's Paradox that we have to carefully take into account all parameters and draw conclusions based on both macro and micro levels of the data, especially when there are many variables each with multiple levels. In that case, the use of small multiples is very important, since it can illustrate many features from different facets and aspects of the data, and also ensures synchronization mappings across facets.

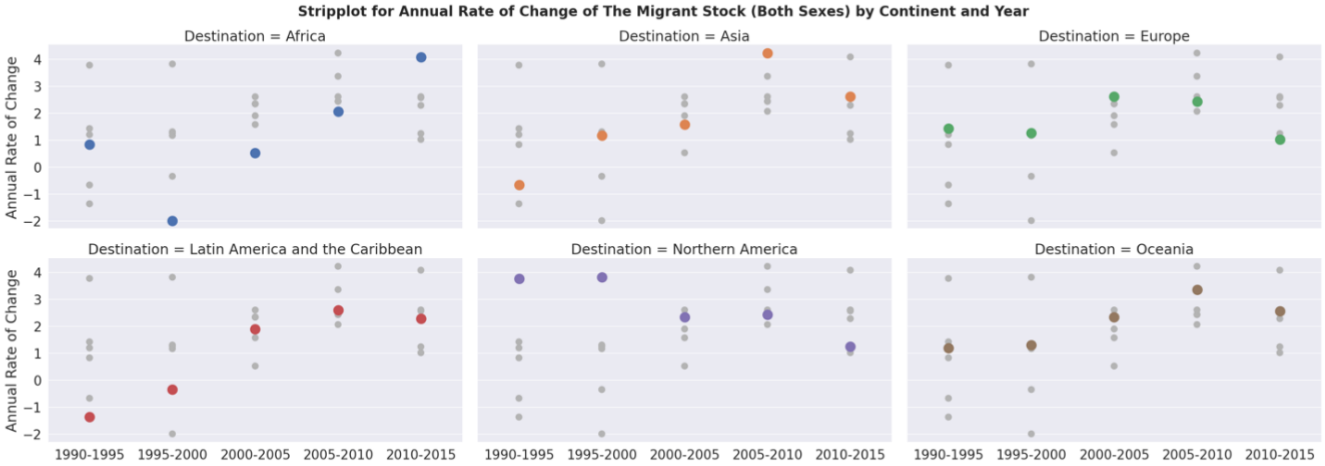
For question 2, we can conclude that there is significant difference among continents in migrants. It's interesting to find out that some continents with very different migrant numbers and populations share the same distribution in migrant percentage, and some continents with relatively lower population has very high percentage of migrant. Remarkably, there are many extreme values in our data set and the sample size for each subgroups are not uniform, which made it hard to compare variables with in the same scale. Also, some data summary statistics like mean and sum will sometimes overgeneralize the result, because they are very sensitive to the outliers. In this dataset, those extreme values are not outliers that can be removed easily, cause they all have meaningful values. In order to achieve easier comparison, it’s necessary to adopt a variable in the same scale/unit for everyone, such as the migrant stock percentage.

Appendix

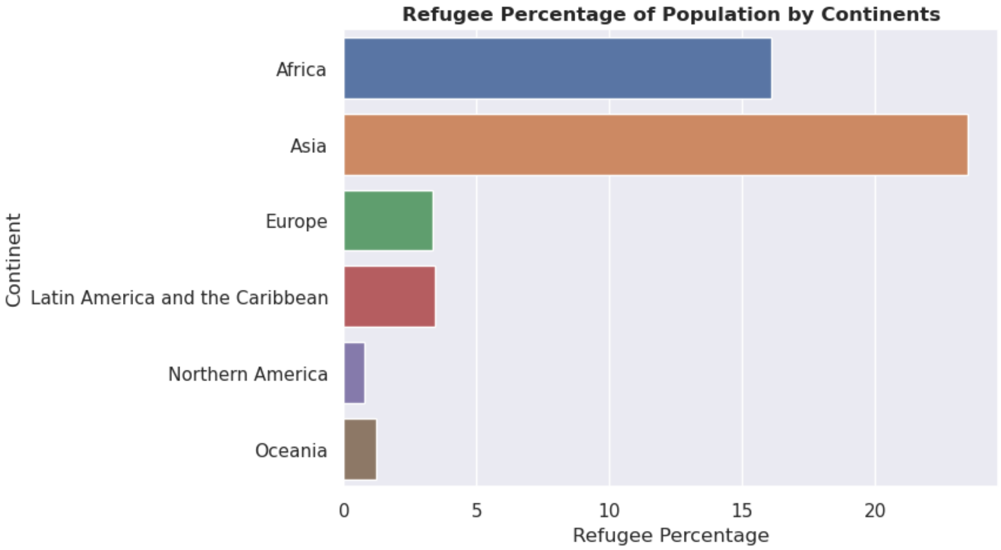
Here are some extra experiments and plots that don’t fit in the narrative of the main report.



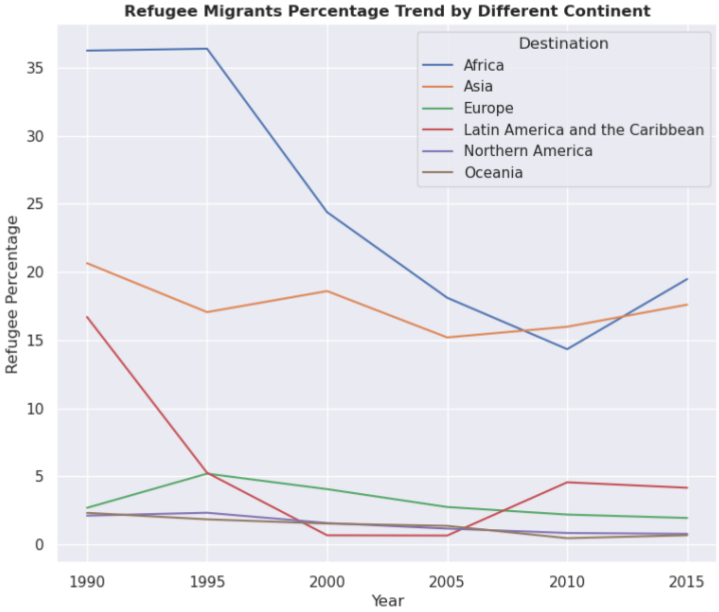
Appendix Figure 1 Female Migrant Percentage Distribution of the countries in each continent by year



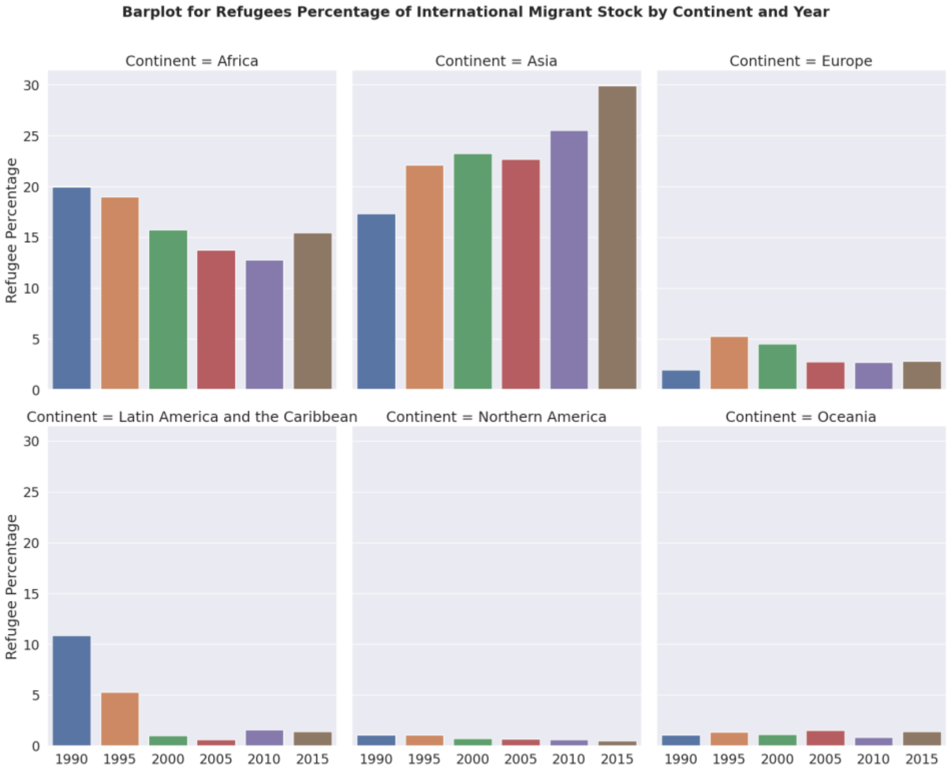
Appendix Figure 2 Scatterplot for Annual Rate of Change of the Migrant Stock (Both Sexes) by Continent and Year



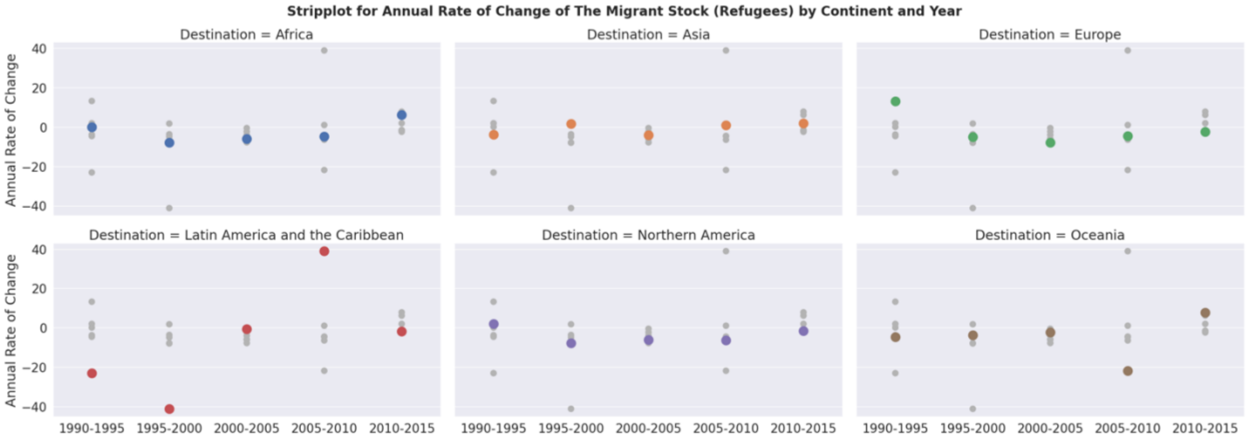
Appendix Figure 3 Bar Plot for Refugee percentage of total population by continent



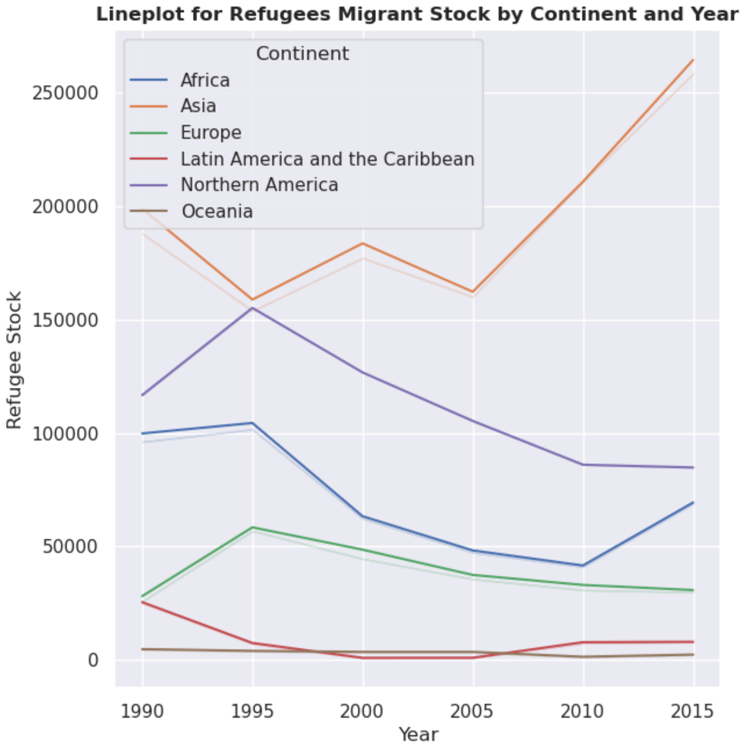
Appendix Figure 4 Line Plot for Refugee Migrant Percentage Trend by Continent



Appendix Figure 5 Small Multiple Bar Plot for Refugee Percentage of Migrant Stock by Continent and Year



Appendix Figure 6 Small Multiple Scatterplot for Annual Rate of Change of Refugee Migrant Stock by Continent



Appendix Figure 7 Line plot for Refugee Migrant Stock by Continent

Reference

Leinhardt, G., & Leinhardt, S. (1980). Exploratory Data Analysis: New Tools for the Analysis of Empirical Data. *Review of Research in Education*, *8*, 85–157. https://doi.org/10.2307/1167124

Tufte, Edward R. (2001) (1983). The Visual Display of Quantitative Information (2nd ed.), Cheshire, CT: Graphics Press, ISBN 0-9613921-4-2.