INF1340: Final assignment write-up

Ruolan Zhang
1004711010

Introduction

As I was browsing around the cleaned dataset, all tables are interrelated and show some features about international migrant's stock from 1990 to 2015. A data scientist needs to understand the definition of every proper noun inside the dataset. Therefore, to clarify, ‘international migrants’ are defined as the group of people who change their country of usual residence, and ‘stock’ refers to the number of international migrants within a country at a specific time point. As international migrants become a worldwide consideration in policy-making, collecting related statistics and monitoring the change of stock have growing importance. Basically, this dataset introduces the changing tendency of international migrant stock across areas, regions, and countries. Gender is another variable I would like to prioritize to visualize preference between male and female international migrants.

As an extension of the assignment to clean up the dataset ‘UN_MigrantStockTotal’, this assignment will focus on using Python functions to perform Exploratory Data Analysis(EDA). Instead of naming it simply as data visualization, EDA is a more complete and insightful process that uses visual tools to perform the classification and summarize characteristics of the dataset. To better present the interaction of data as well as comparing the developing trend, EDA enables stakeholders with different backgrounds to get involved in understanding statistical results without having any basis for data modeling.

Method

Throughout this project, bar charts are used most of the time as the main graphical method, since this method is able to summarize large amounts of data into easy-understanding visualizations. According to Tufte's visualization Principles, visualizations we created need to show comparison, causality and multivariate data. Following those guidelines, we should display as many variables as possible in the EDA process to show a comprehensive picture of the dataset.

Therefore, I choose to use a multi-series bar chart where each category on the x-axis will be further stratified by a secondary category level indicated by different colors. In this dataset, statistics are recorded with stratified levels of locations, such as major areas, regions, and countries. For example, when I want to display the international migrant's stock from 1990 to 2015 by major areas, the final bar chart will be stacked by areas. In this way, stakeholders are able to compare the specific stock for each major area at any time point in one single visualization.

To be noticed, since each table contains too many statistics, it would be super confusing to display data for every country. As a result, in some of the bar charts, I only picked the top 10 countries with the most international stocks. However, it does not mean that the rest of the countries have no analysis significance, I hope to focus more on countries with representativeness. Also, the reason why both horizontal and vertical bar charts are presented in this assignment is that some category-level names overlap with each other in the vertical version of bar charts.

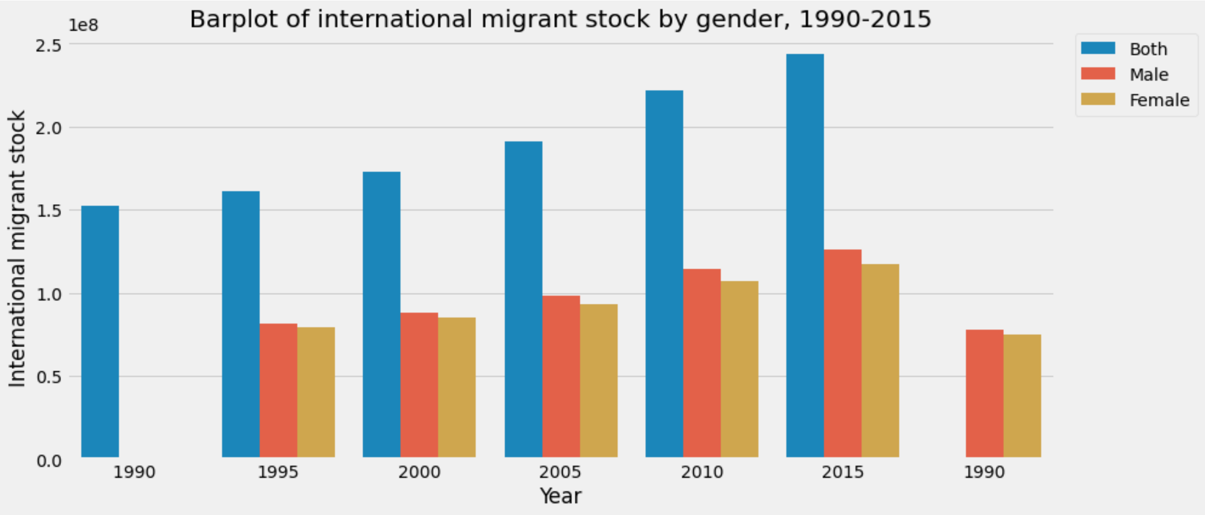
Another graphical method I used in this EDA process is the line chart when I prioritized the developing trend of dependent variables within a time period over the specific values at a time point. For example, to visualize the annual change rate of international migrants by year, a line chart would be a better choice rather than drawing a bar chart. Boxplot was used when comparing the international stock of developing countries and developed countries, to display statistical features such as median, upper/lower quartiles, and outliers. The advantage is it can graphically demonstrate the value and spread of a variable at a glance.

In addition, talking about aesthetics and clarity, each plot will be carefully accompanied by proper labels, titles, and legends with indicating colors, which are mentioned by Tufte's visualization principle. Scales will be adjusted according to plot characteristics at the very last step to facilitate comparison between each category. Deleting unnecessary elements to maximize data-ink ratio is also another guideline we should follow when we perform EDA. As I stated in my midterm project, for clarification, I split data in each table by stratified districts, such as 'World', 'Major area', 'Region', and 'Country', to facilitate the data extraction in the EDA process. For each chart, I filter rows and columns that I would like to be included in my visualizations and apply the different charts functions under the 'Seaborn' package to draw every chart. Adding table titles and modifying the scale of the visualization is the final step of the visualization process.

Result

EDA for table1

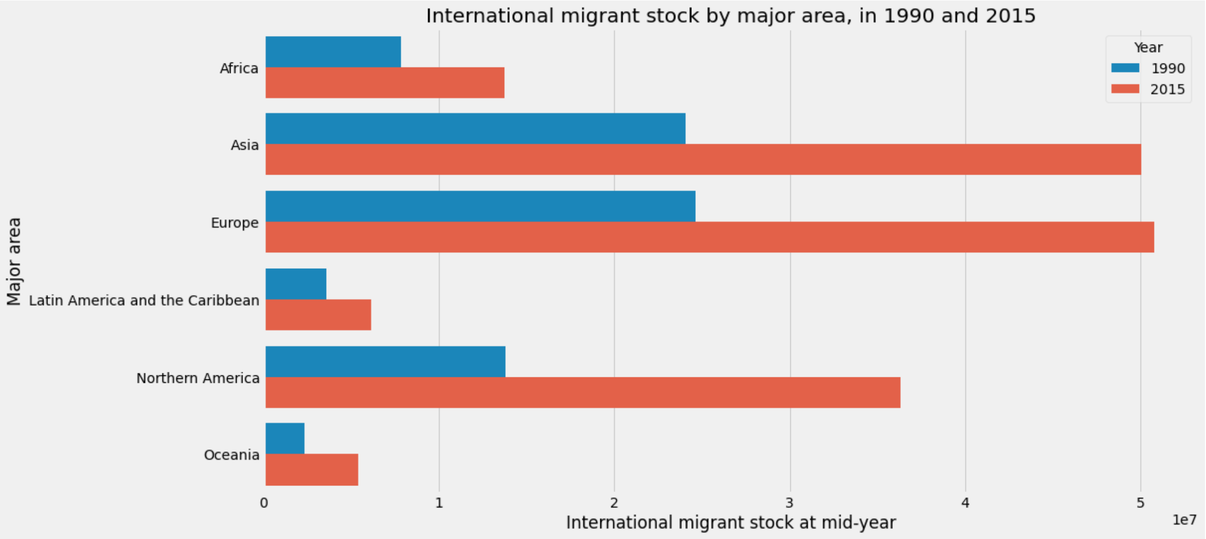
Figure1.1:



Firstly, I want to apologize for the weird format of bars showing the stock in 1990. Bars standing for males and females are separated from the bar of "Both". I found no error in my code and cannot figure out the problem here. I will definitely bring this question to the professor afterward.

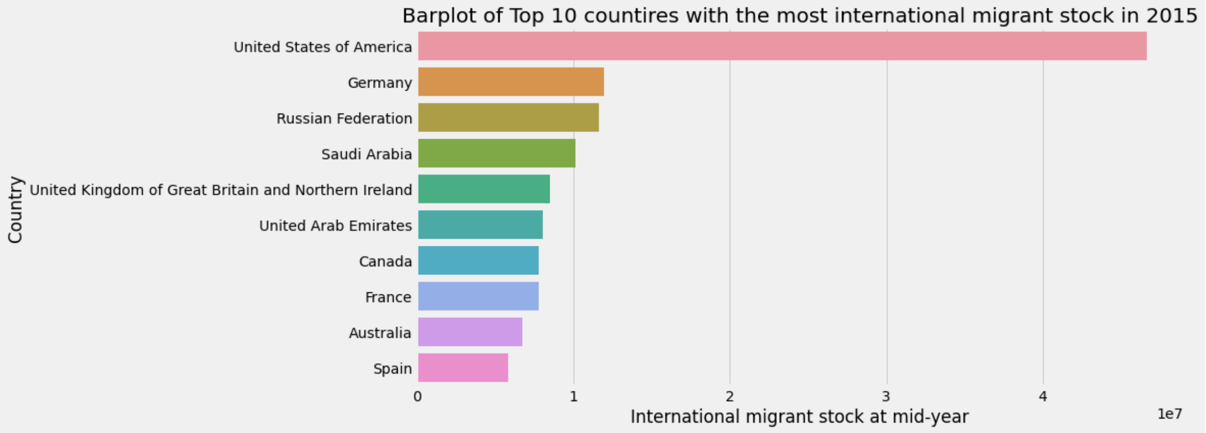
Figure 1.1 is a multi-series bar chart that displays the international migrant stock by gender between 1990 to 2015. I filtered out every row indicating ‘World’, and stacked the bar by the gender variable. There is a clear trend showing the increasing number of international migrant stock for both males and females from 1990 to 2015. The number of male migrants is always slightly higher than that of females.

Figure1.2:



After getting an overview of the total number of migrants for each year, let’s narrow the scope and focus on the stock in each major area. Figure 1.2 is a multi-series bar chart with this information. I am also curious about the change in stock between 1990 and 2015, so I grouped statistics from 1990 and 2015 together and plotted them in the same graph. As can be seen from the graph, Europe and Asia have been in the leading place all the way since 1990. The international migrant stock in Northern America more than doubled in 2015 compared to 1990, which has the greatest growth rate than any other major area in the world.

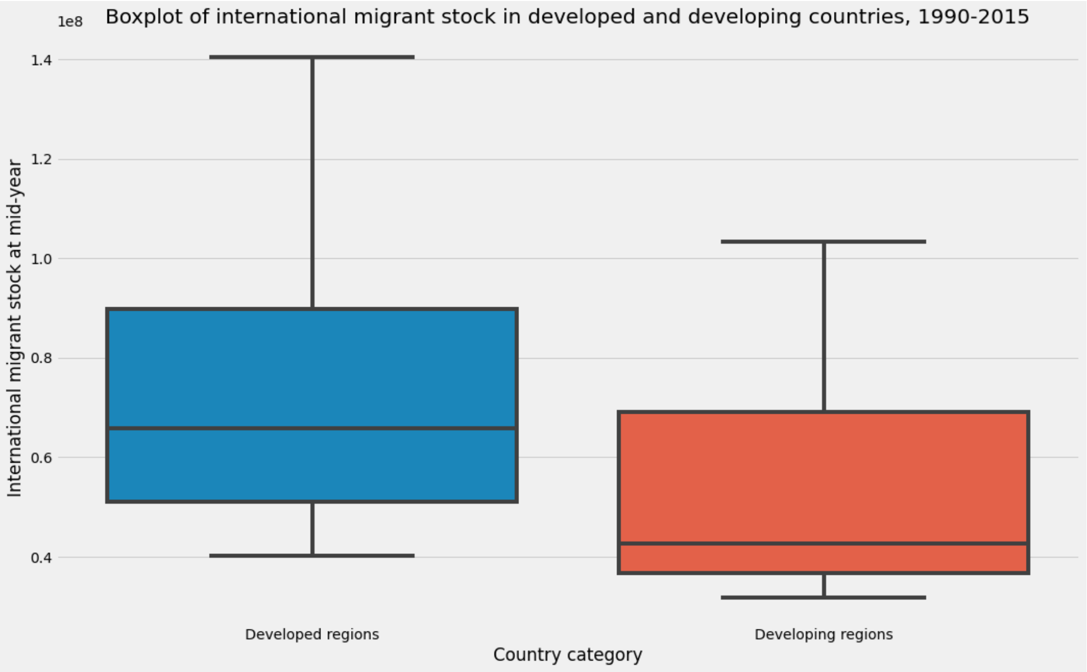
Figure1.3:



Keeping on zooming in on the scope, we arrived at the country level. After filtering rows and columns I would like to be included in , I sorted countries by their international migrant stock. Figure 1.3 is a basic bar chart to display the number of migrants in each country. It is impractical and unnecessary to mention all countries listed in the table, so I picked the top

10 countries with the largest number of international migrants from the most current information, 2015. Not surprisingly, international migrant stock in the United States of America is significantly larger than the rest of the countries listed in this bar chart.

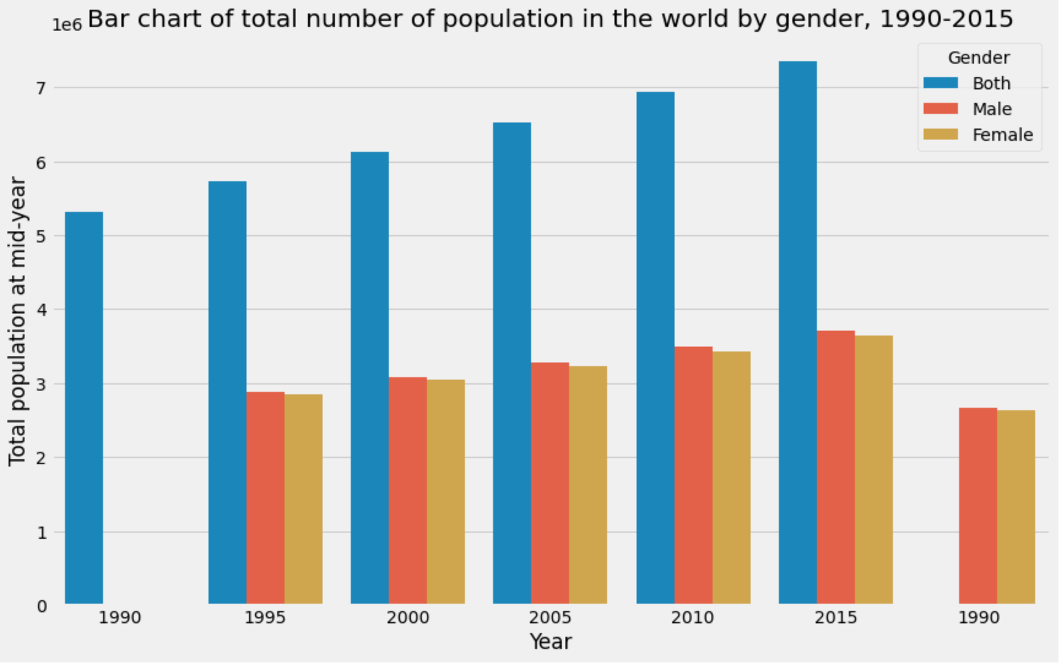
Figure 1.4:



I filtered out rows indicating developed and developing regions and columns I would like to visualize in this box chart. Figure 1.4 is the boxplot to compare the difference in the international migrant stock in developed and developing countries, data collected once per 5 years from 1990 to 2015. It facilitates us to compare the mean international stock from 2 categories over 25 years. As can be seen from the graph, the stock in developed countries is dramatically higher than that of the developing countries.

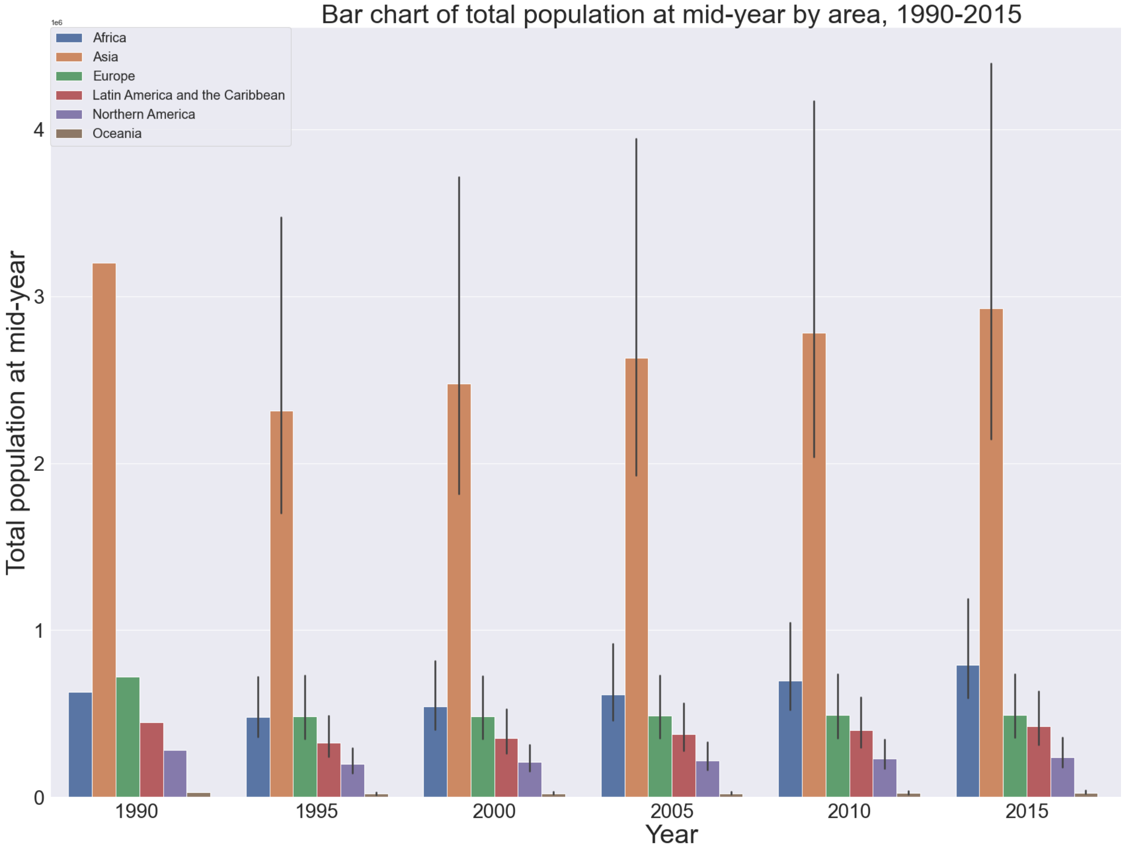
EDA for table2

Figure2.1:



Have the same problem for bar 1990 as I stated in Figure 1.1.
Figure 2.1 is a multi-series bar plot to display the population in the world by gender between 1990 to 2015. There is also a clear trend showing the increasing population in the world for both males and females in this 25 years interval. The number of males population is always slightly higher than that of females.

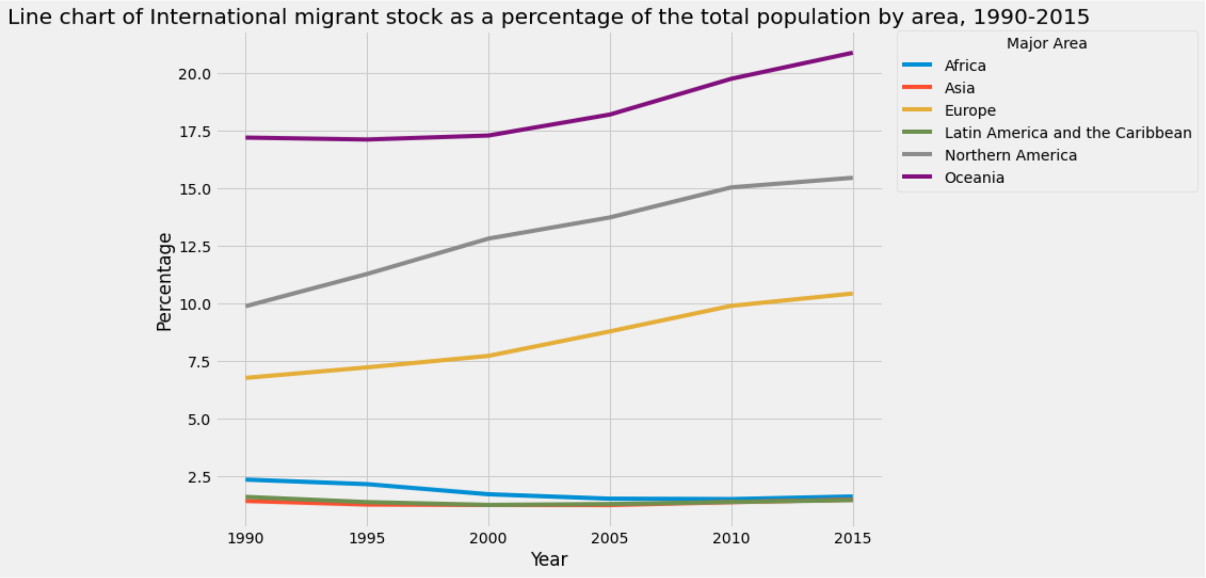
Figure2.2:



Zooming in to the major area level of the total population to see the population change in each major area. Extracting data from the major area dataset I cleaned for this table, I stacked the bar by their major areas. Figure 2.2 is also the multi-series bar chart for recording population by year. Asia held the undisputable greatest population during these 25 years. Surprisingly, all major areas experienced a population drop between 1990 to 1995 but kept growing after 1995. To be noticed, the population in Africa ranked third at first but quickly caught up with Europe in 1995. Then easily surpassed to the second place after 2000.

EDA for table3

Figure3.1



Extracting data from the major area dataset in table 3, figure 3.1 is a line chart showing the trend of international migrant stock as a share of the total population in different major areas between 1990 and 2015. As can be seen from the chart, Oceanic and Northern America have a relatively larger share of international migrant stock in their total population, followed by Europe. At the same time, their percentages have kept growing in the 25 years interval. In comparison, In the rest of the area, due to unknown reasons, international migrant stock only occupied a very small share of their total population.

EDA for table4

Figure4.1

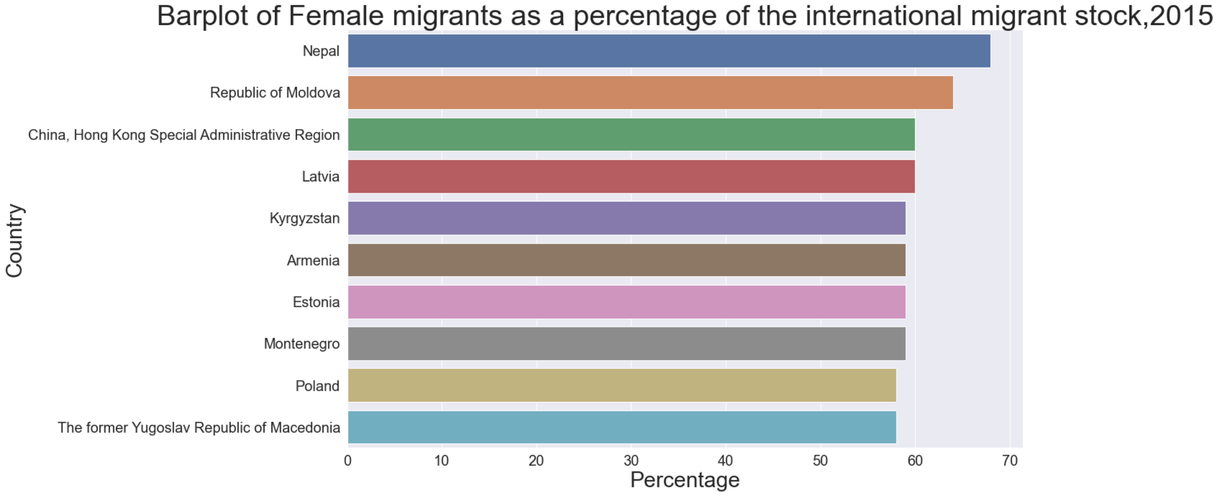
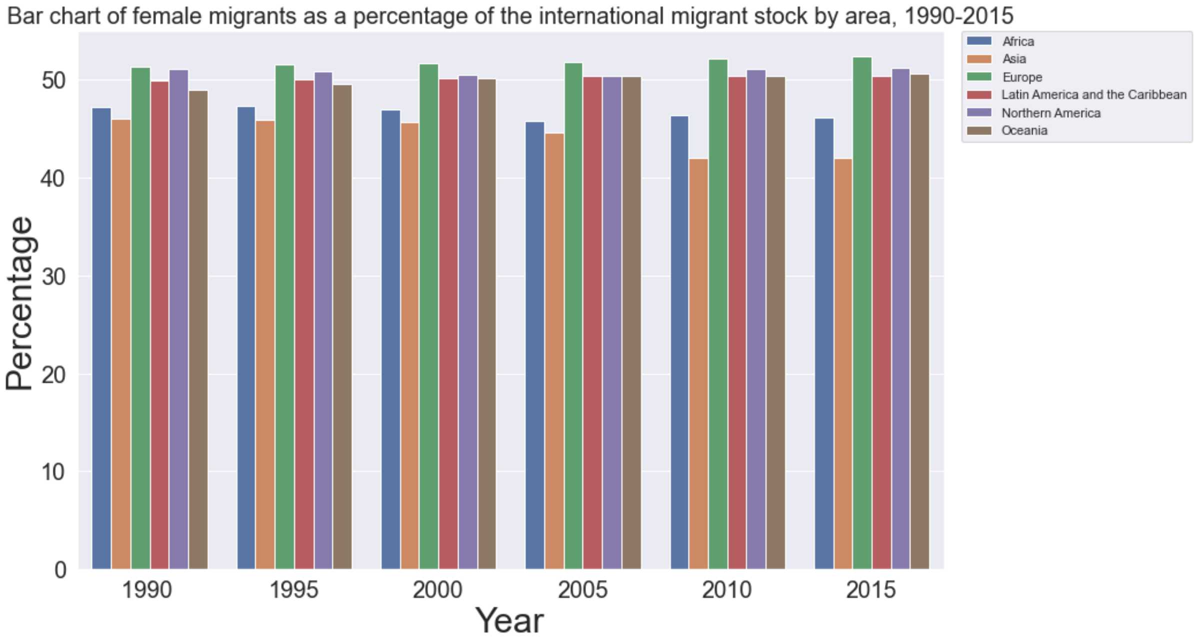


Table 4 collects the percentage of female migrants in the total international migrants stock. In Figure 4.1, I picked 10 countries with the largest share of female migrants to figure out females' preferences in immigration in 2015. Ten countries listed in the above bar chart have relatively similar percentages of female migrants in 2015, between 58% to 67%.

Figure4.2



Moving to the major area level, Figure 4.2 is a multi-series bar chart showing the percentage of female migrants in each area from 1990 to 2015. As can be seen from the graph, Asia had a relatively lower percentage of female international migrants compared to the rest of the areas, and the percentage kept dropping after 1990. In contrast, female migrants in the Europe area occupied a slightly larger percentage than that of males within 25 years.

EDA for table5

Figure5.1

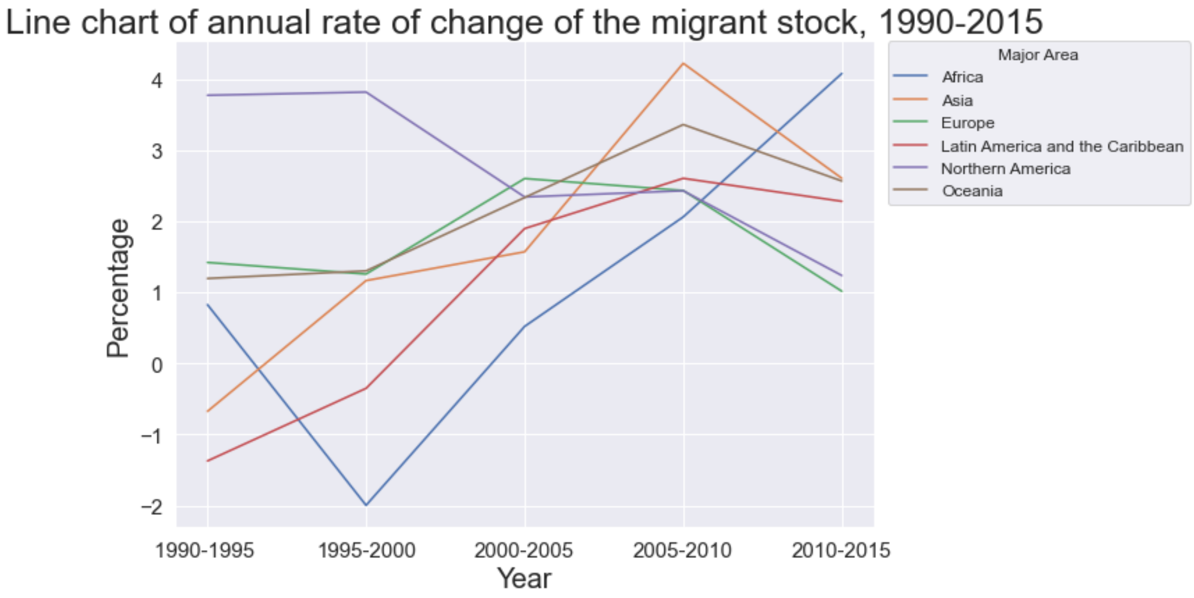


Figure 5.1 is a line chart showing the annual rate of change of international migrant stock in different major areas from 1990 to 2015. As can be seen from the graph, three areas displayed an upward trend during 25 years except for Northern America, Africa, and Europe. Africa experienced a significant drop between 1995 and 2000 and then kept rising to first place in the time interval of 2010 to 2015. In comparison, Northern America showed a general downward trend over the 25 years.

Figure5.2

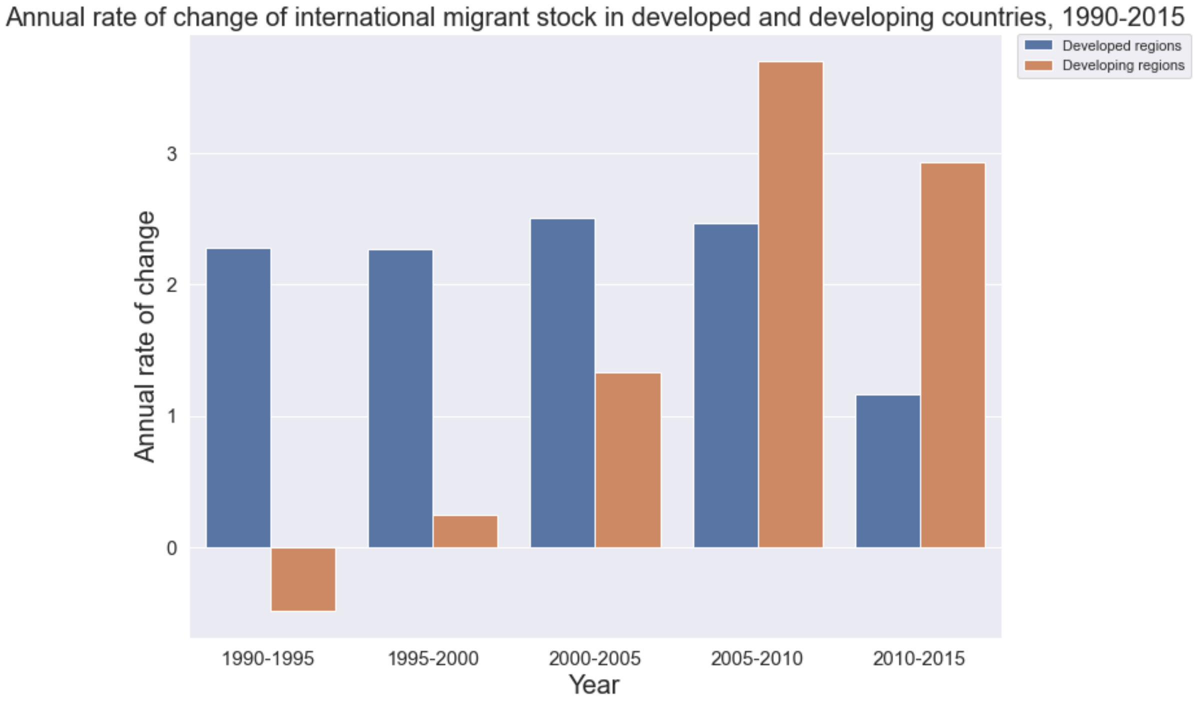


Figure 5.2 is a multi-series bar chart showing the annual rate of change of international migrant stock in both developed and developing countries between 1990 to 2015. In the very first 15 years, the annual rate of changes for developing countries was significantly lower than that of developed countries.

Between 1990 to 1995, the annual rate of change for developing countries was a negative value but started to grow afterward. Its ascent reached its peak between 2005 to 2010 and remained the leading place between 2005 to 2015.

EDA for table6

Figure6.1

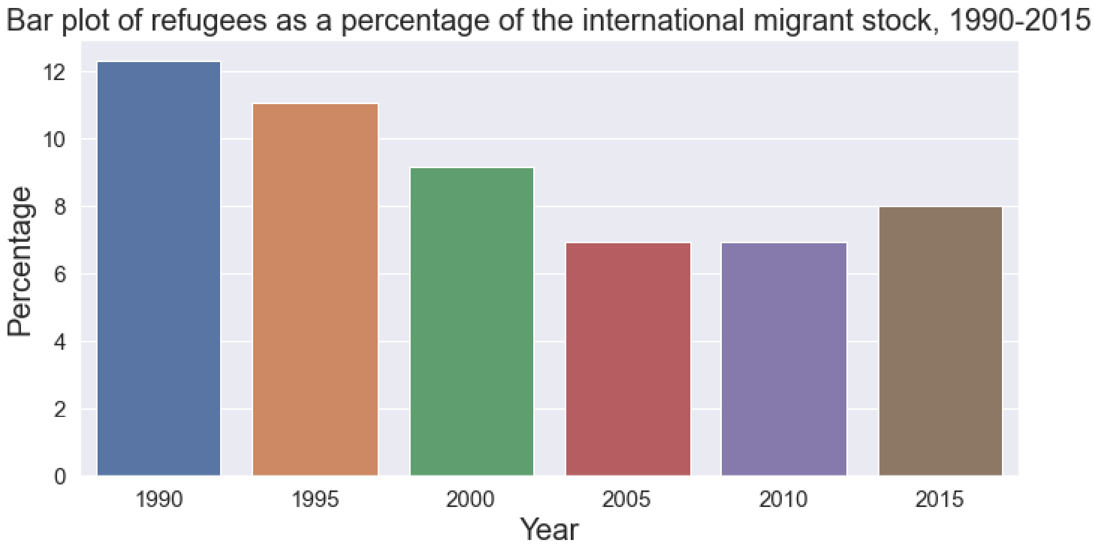


Figure 6.1 is a basic bar chart showing the change in refugee as a percentage of international migrant stock from 1990 to 2015. The percentage of refugees reached around 12% in the first 10 years but kept declining to 6% 15 years after.

Figure6.2

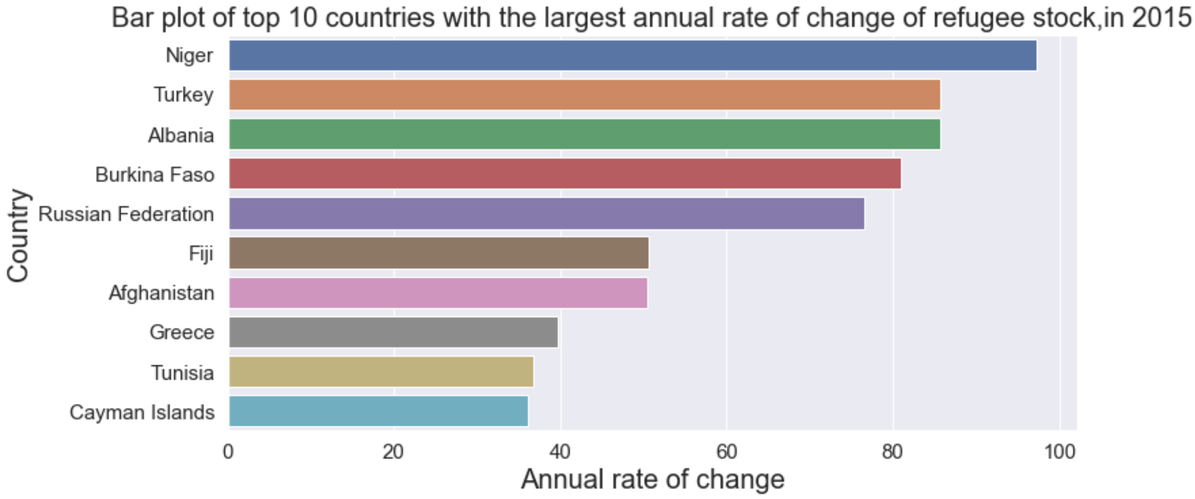


Figure 6.2 displays the top 10 countries with the highest annual rate of change of refugee stock in 2015. The country ranked at the first is Niger, located in Africa, where its annual rate of change of refugees rose to around 95% in 2015. It means that Niger almost doubled its refugee population in 2015.

Discussion

As international migrants become a predominant consideration in both political and social affairs, worldwide governments and organizations keep track of the flow of those migrants each year in stratified districts. Considering both the economical and cultural advantages, encouraging international migrants is able to flourish diversity and cultural exchange. However, at the same time, migrant stocks exceeding the limit boundary could become a burden for local authorities. This dataset gathered information from diverse aspects of statistics about international migrant stocks from 1990 to 2015. As a part of the EDA process, I finished plotting visualizations by different graphical methods. In this discussion part, I don't have any specific research questions about the international migrant stock but will present some of my considerations and summaries about this dataset. To be noticed, we only have data from 1990 to 2015 but lack statistics for the recent few years. Therefore, my analysis will focus on the pattern between that 25 years interval, accompanied by some predictions for future years.

Firstly, talking about the overall trend I observed from the bar chart. The number of international migrants all over the world kept increasing for 25 years after 1990 and reached almost 250 million in 2015(Figure 1.1) when the total number of populations kept rising in the world(Figure 2.1). In general, the number of male migrants was slightly higher than that of females all the way from 1990 to 2015. Compared with the developing regions, obviously more people are willing to migrate to developed regions, which perfectly corresponds to the normal sense that countries with mature industrialization and economics are more welcomed by citizens. The mean international migrant stocks of developed regions over 25 years are around 65 million, which is about 20 million excessed that of the developing regions(Figure 1.4). However, despite the annual rate of change of international migrants in developing regions being negative between 1990 to 1995, it experienced significant growth after 2000 and even exceeded that of the developed regions(Figure 5.2). The prediction is, there will be an increasing number of international migrants favoring developing regions

Zooming into the major area level, most of the international migrants, about two-thirds of the total migrant stock, preferred to live in Asia or Europe since 1990. To be noticed, being an international migrant in Northern America has become a popular trend since the migrant stock in this area in 2015 was more than doubled from 1990(Figure 1.2). For major areas such as Oceanic, Northern America, and Europe which have relatively smaller population bases than the rest areas, international migrants occupy larger shares of their total population although their annual changing rate is not as large as that of Asia(Figure 3.1). This is the reason why people would consider Northern America and Europe as the realm of diversity. So we are able to get in touch with people of diverse races and cultural backgrounds. Specifically, female international migrants favored areas such as Northern America and Europe, where percentages of female stocks stayed at or exceeded 50% of total stocks(Figure 4.2).

Quickly talked about the information I extracted from the country level. The United States of America was always the country having the largest number of international migrant stocks over the 25 years although the annual rate of change in Northern America displayed a slow downward trend. Combining this information, it is not hard to guess that international migrants living in Northern America would possibly set The United State of America as their target destination(Figure 1.3).

Last but not the least, refugee statistics and related policies are other important features to assess a country. The world is having a decreasing number of refugees(Figure 6.1) as reported and we can easily figure out a country's willingness to accept refugees from the statistics. Increasing numbers of refugees are migrating to countries such as Niger, Turkey, and Albania showing the acceptance of refugees in developing countries(Figure 6.2). Refugees are usually not calculated in the censuses, however, their rights as a human should always be properly protected and respected.

Conclusion

However, this dataset has some intrinsic limitations. There exist missing values in several tables and instead, data collectors put two dots inside boxes. They did not specify whether it means that the annual rate of change of international migrants is 0 or they did not receive any data about that specific country. This would significantly influence the result if we want to continue building models or applying other non-graphical analysis methods. Another point is, this dataset did not record where those international migrants came from, since statistics of inter-continent migrants and cross-continent migrants should be manipulated according to different social contexts and economic backgrounds.

What I understand as EDA is a step-by-step process to prepare, clean, visualize and brainstorm a dataset. It enables stakeholders and clients with diverse academic and professional backgrounds to interact with messy statistical results. The process of Visualizing data with graphical methods is repetitive but extremely essential for presenting overall trends, patterns, and outliers. Not limited to designing skills but instead, it is a way of storytelling and attracting stakeholders' attention in a way they can understand. In this assignment, we are acting to present our idea to governmental organizations to better understand the flow and pattern of international migrants, aligned with Tufte's Visualization principles. I do hope my EDA process can provide some supplementary support for assessing features of international migrant stock in the past years.

Reference

"International Migrant Stocks." *Migration Data Portal*, https://www.migrationdataportal.org/themes/international-migrant-stocks.

CN, Prajwal. "EDA - Exploratory Data Analysis: Using Python Functions." *DigitalOcean*, DigitalOcean, 3 Aug. 2022, https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python#exploratory-data-analysis-eda.

Brush, Kate, and Ed Burns. "What Is Data Visualization and Why Is It Important?" *Business Analytics*, TechTarget, 8 Dec. 2022, https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization.