Name: Moxuan Huang

Project: INF1340 Final Project

Date:12/12/22

Introduction

Since the 20th century, globalization has accelerated, and the number of immigrants and refugees has also been constantly changing, which is related to some historical factors as well as the current policies of different regions and countries. In this project, I will use the migration and refugee data set released by the United Nations for data cleaning. The excel file contains a general introduction, six tables, relevant data, an annex and a note. We mainly analyze six tables, which give the annual number of Chinese immigrants, total population, percentage of international migrant population in the total population, percentage of female migrants in the international migrant population, annual change rate of migrant population and estimated number of refugees in the middle of the year by gender and major regions, regions, countries or regions from 1990 to 2015. It is hoped that effective information can be extracted through the processing and analysis of these data, and the direction and scale of population flow can be measured through the statistical analysis of the data of migration volume, migration rate and refugee volume in different periods, so as to summarize the key data and trends of global migration, and focus on the regional level to explain the migration and their migration in different regions of the world in detail.

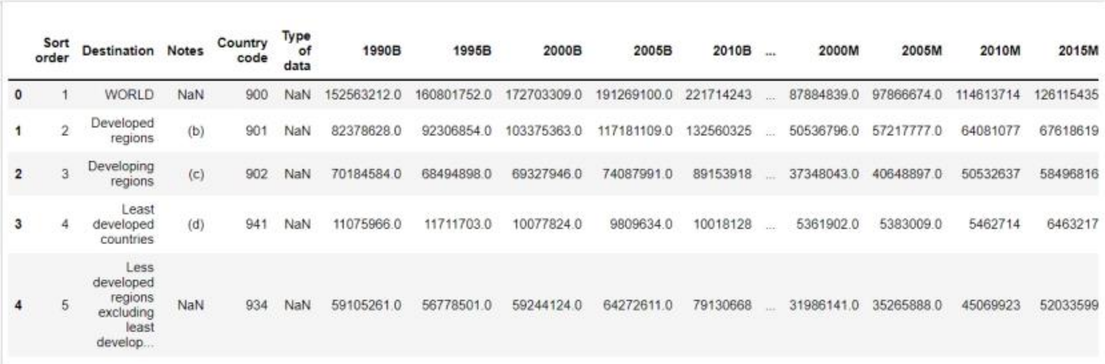
Part I. Re-tidying the UN dataset

Due to the poor tidying of data and write-up during the midterm, I re-tidied every table before my analysis. I use pandas to import excel file into Jupyter Notebook and convert it to DataFrame format. The data consists of 9 tables. CONTENTS is an explanatory document for the remaining 8 tables. Table 1 - Table 6 are the data to be analyzed next. ANNEX and NOTES supplement the field data in these six tables. In the following analysis, I mainly used the data of Table1 - Table6. ANNEX data supplement the above data. Since the 6 tables were relatively similar, I will explain how I tidied them more generally based on the 5 principles below.

The “5 Principles”

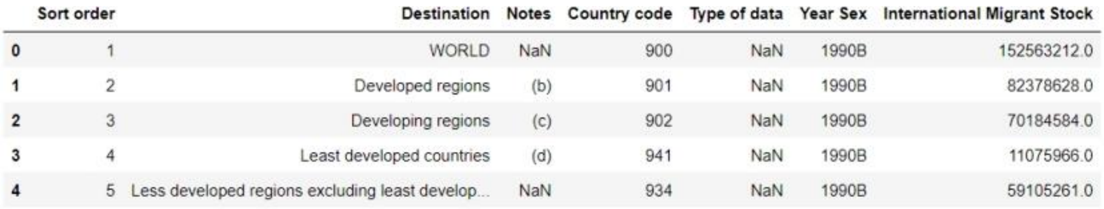
Principle 1: column names need to be informative, variable names and not values

One problem with most tables is that there are values in the table column names. They record gender or year in column names, making data processing and analysis difficult. I omit the original column names of the table, read the data in the correct format, and then use “melt” function in Pandas to transform the table in order to expressing gender or year in the data instead of the column names. The table 1 shows the raw data and table 2 shows the data after using the “melt” function below:



	Sort order	Destination	Notes	Country code	Type of data	1990B	1995B	2000B	2005B	2010B	...	2000M	2005M	2010M	2015M
0	1	WORLD	NaN	900	NaN	152563212.0	160801752.0	172703309.0	191269100.0	221714243	...	87884839.0	97866674.0	114613714	126115435
1	2	Developed regions	(b)	901	NaN	82378628.0	92306854.0	103375363.0	117181109.0	132560325	...	50536796.0	57217777.0	64081077	67618619
2	3	Developing regions	(c)	902	NaN	70184584.0	68494898.0	69327946.0	74087991.0	89153918	...	37348043.0	40648897.0	50532637	58496816
3	4	Least developed countries	(d)	941	NaN	11075966.0	11711703.0	10077824.0	9809634.0	10018128	...	5361902.0	5383009.0	5462714	6463217
4	5	Less developed regions excluding least develop...	NaN	934	NaN	59105261.0	56778501.0	59244124.0	64272611.0	79130668	...	31986141.0	35265888.0	45069923	52033599

Table 1. Raw data.

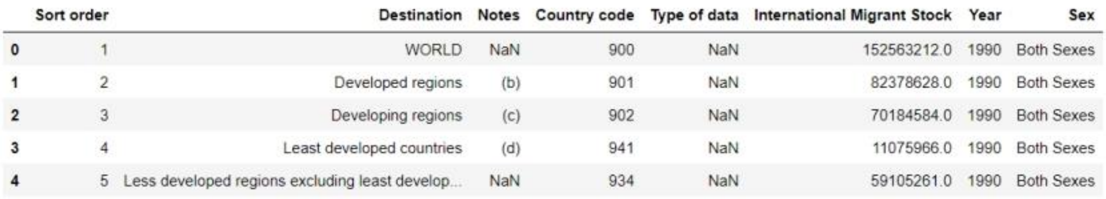


	Sort order	Destination	Notes	Country code	Type of data	Year Sex	International Migrant Stock
0	1	WORLD	NaN	900	NaN	1990B	152563212.0
1	2	Developed regions	(b)	901	NaN	1990B	82378628.0
2	3	Developing regions	(c)	902	NaN	1990B	70184584.0
3	4	Least developed countries	(d)	941	NaN	1990B	11075966.0
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990B	59105261.0

Table 2. Data after the “melt” function.

Principle 2: each column needs to consist of one and only one variable

After melting the year values, the column contains a combination of two variables, which is the population of international migrant stock and male and female. To distinguish them, I use the “apply” function in Pandas to split this mixed column into two separate columns. Table 3 presents the data after using the “apply” function.

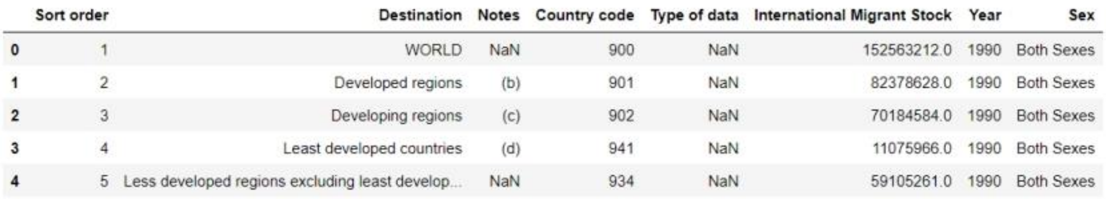


	Sort order	Destination	Notes	Country code	Type of data	International Migrant Stock	Year	Sex
0	1	WORLD	NaN	900	NaN	152563212.0	1990	Both Sexes
1	2	Developed regions	(b)	901	NaN	82378628.0	1990	Both Sexes
2	3	Developing regions	(c)	902	NaN	70184584.0	1990	Both Sexes
3	4	Least developed countries	(d)	941	NaN	11075966.0	1990	Both Sexes
4	5	Less developed regions excluding least develop...	NaN	934	NaN	59105261.0	1990	Both Sexes

Table 3. Data after the "apply" function.

Principle 3: variables need to be in cells, not rows and columns

According to the completed processing, the column names of data no longer contain value information. Through observation, there is no data information in the data row. Therefore, the Principle 3 has been reached. Variables are only in cells.

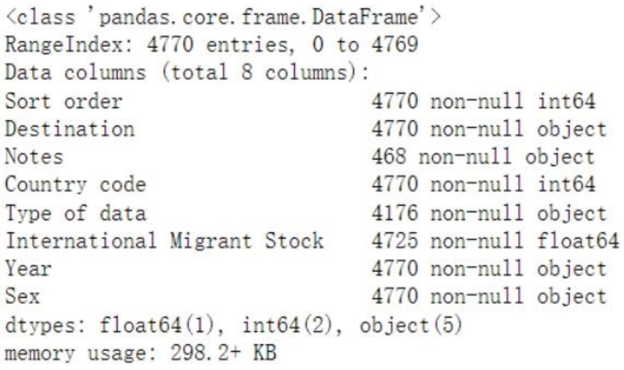


	Sort order	Destination	Notes	Country code	Type of data	International Migrant Stock	Year	Sex
0	1	WORLD	NaN	900	NaN	152563212.0	1990	Both Sexes
1	2	Developed regions	(b)	901	NaN	82378628.0	1990	Both Sexes
2	3	Developing regions	(c)	902	NaN	70184584.0	1990	Both Sexes
3	4	Least developed countries	(d)	941	NaN	11075966.0	1990	Both Sexes
4	5	Less developed regions excluding least develop...	NaN	934	NaN	59105261.0	1990	Both Sexes

Table 4. values in cells.

Principle 4: each table column needs to have a singular data type

The original data has missing values and abnormal values, such as the "International Migration Stock" in Table 1. Normally, this column should be data of float type. But the value '..' in the data is a string. Therfore, the data type of this column cannot be unified. I replace the value '..' with np.NAN and unify the format to float type.

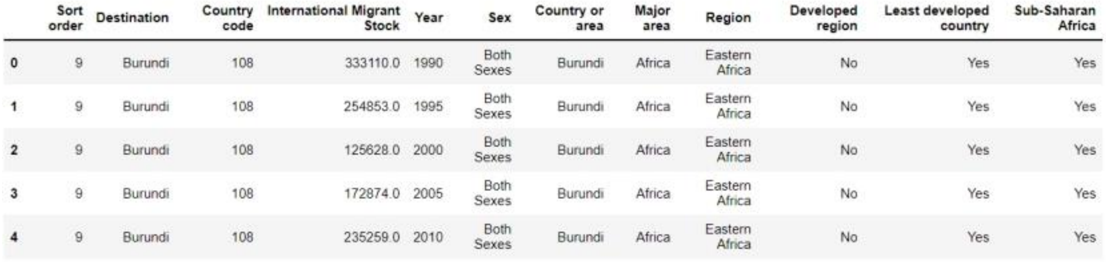


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4770 entries, 0 to 4769
Data columns (total 8 columns):
Sort order                    4770 non-null int64
Destination                   4770 non-null object
Notes                         468 non-null object
Country code                  4770 non-null int64
Type of data                  4176 non-null object
International Migrant Stock   4725 non-null float64
Year                          4770 non-null object
Sex                           4770 non-null object
dtypes: float64(1), int64(2), object(5)
memory usage: 298.2+ KB
```

Table 5. table after the "replace".

Principle 5: a single observational units must be in 1 table

The data in Table 1-Table 6 are in the same column when dividing regions and sub regions. They cannot be distinguished only by the data in the table. So I borrow the data in the ANNEX table. I use the merge function in Pandas to connect data in the "inner" way with the "Country code" as the primary key to obtain more information about the same data.



	Sort order	Destination	Country code	International Migrant Stock	Year	Sex	Country or area	Major area	Region	Developed region	Least developed country	Sub-Saharan Africa
0	9	Burundi	108	333110.0	1990	Both Sexes	Burundi	Africa	Eastern Africa	No	Yes	Yes
1	9	Burundi	108	254853.0	1995	Both Sexes	Burundi	Africa	Eastern Africa	No	Yes	Yes
2	9	Burundi	108	125628.0	2000	Both Sexes	Burundi	Africa	Eastern Africa	No	Yes	Yes
3	9	Burundi	108	172874.0	2005	Both Sexes	Burundi	Africa	Eastern Africa	No	Yes	Yes
4	9	Burundi	108	235259.0	2010	Both Sexes	Burundi	Africa	Eastern Africa	No	Yes	Yes

Table 6. table after the "merge" function.

Part II. Explanatory data analysis (EDA)

Methods

Tufte principles of data visualization will be referenced and followed for plotting the graphs and analyzing the data. The two packages seaborn and matplotlib.pyplot will be used to construct the statistical plots and demonstrate data visualization for the UN dataset.

The matplotlib.pyplot function, figure(figsize=()) determines the width and height of the figure in inches. Depending on the number of images, I have set them to either (12, 6) or (16, 6), choosing the former when there are two images, and latter when there is only one image. The other matplotlib.pyplot function "Subplot()" function takes three arguments that describes the layout of the figure, for example in the first two graphs (Figure 1 and Figure 2) representing "Changing trend of world migration population by gender" and "Changes in the number of world migrants by gender", plt.subplot(1, 2, 1) stands for 1 row, 2 columns, and first plot, and plt.subplot(1, 2, 2) stands for 1 row, 2 columns, and second plot.

Then there are the plots created using the seaborn package: sns.lineplot() for a line graph, sns.scatterplot() for scatterplot, sns.barplot() for barcharts, sns.boxplot() for constructing a boxplot, and sns.violinplot for a violin plot. The data on the plot is determined by what we want to measure, in this analysis, we mostly look at the data for the world, thus, setting data = df[df['Destination'] == 'WORLD'. In some cases, we would have to filter the DataFrame rows that contain a list of values, this is done through isin().

Furthermore, axis of the plots is all labelled by stating what x is and what y is. For all the line graphs, legend for the scatters are removed as we only require one legend for the lines, this is done through legend = False. Hue determines which column in the data frame should be used for colour encoding. Since figure names are mandatory for all the plots placed in this write-up document, we don't want to repeat them again in the title, therefore, I only labelled the titles on the first two plots to demonstrate the knowledge of knowning how to label titles, and this is done by the function plt.title(). At least, the plt.show() function displays the graphs.

There may be codes and images for pie charts in the ipynb file, that part is included just for a learning process and for my own data visualization, it is not included in the write-up

Result & Discussion

Table1 - International migrant stock at mid-year by sex and by major area, region, country or area, 1990-2015

First of all, let's take a look at the number and change trend of international migrants in the world in 1990, 1995, 2000, 2005, 2010 and 2015. According to Tufte, “Above all else, show the data”. I started plotting different type of graphs to present the data. Here, we illustrated the number and change trend of migrants by sex as well as the overall number and trend through a multiple bar graph and a multiple line graph. Both graphs have displayed an upwards trend.

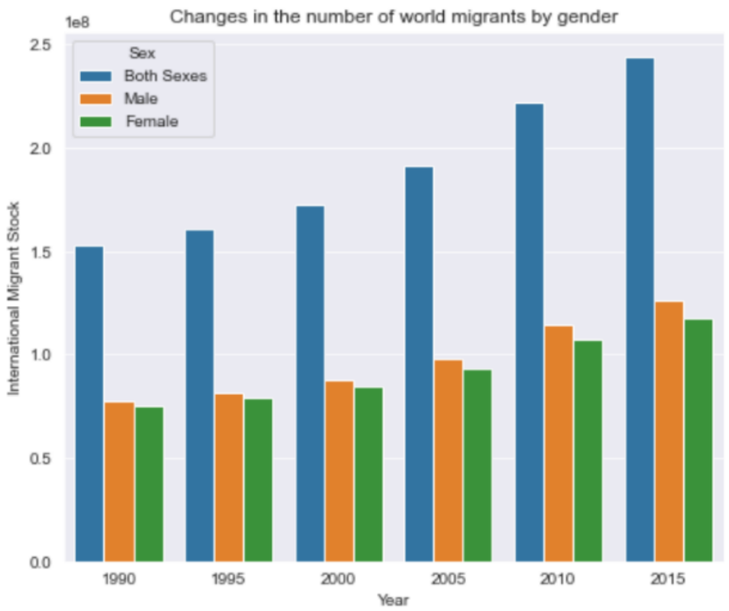


Figure 1. Changes in the number of world migrants by gender.

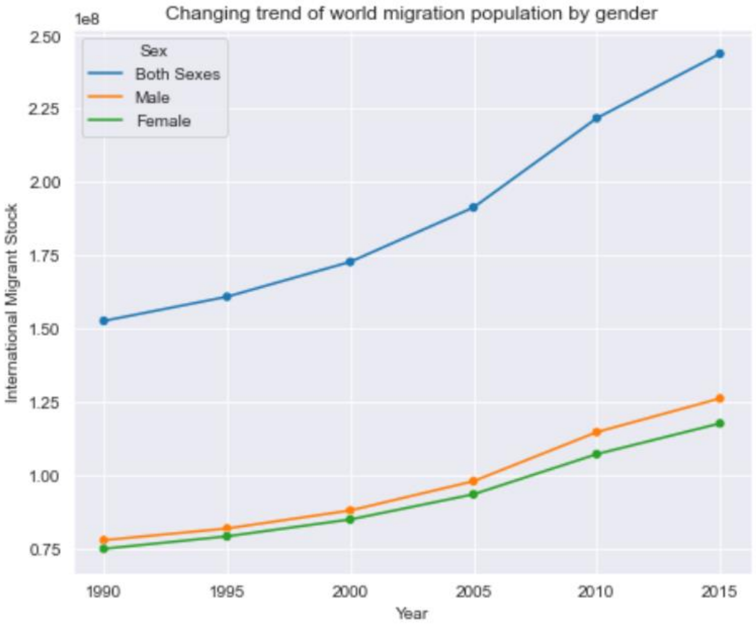


Figure 2. Changing trend of world migration population by gender.

Next, through plotting a side-by-side barchart, we clearly visualize the number of immigrants from different continents and the proportion of immigrants from different continents in 2015. Note how the number and letters on the x-axis “1e7” is expressed using scientific notation, meaning 1 multiplied by 10 to the 7th power, e representing exponent, in this case, Asia would have a migrant stock around 75,000,000.

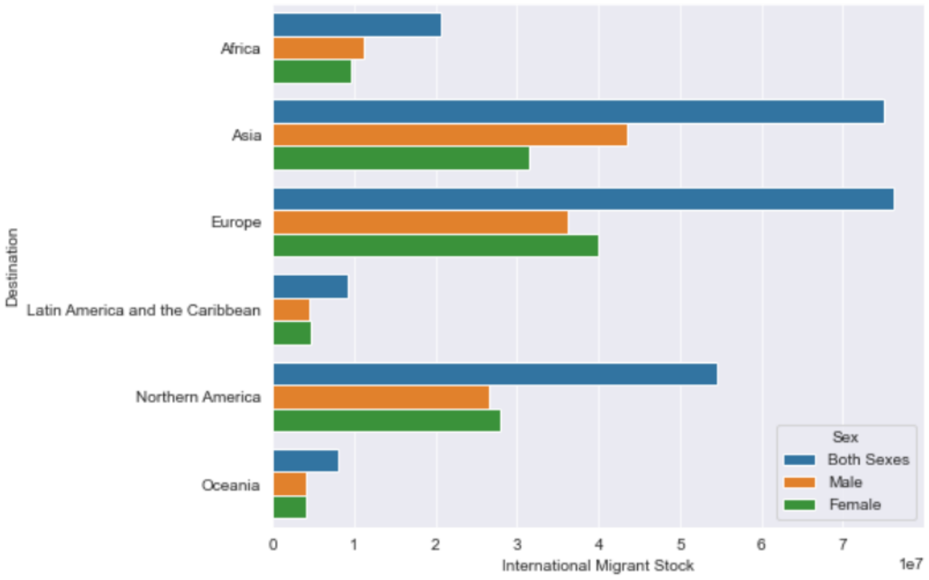


Figure 3. International migration statistics by continent in 2015.

Next, a line graph is used to see the changing trend of the number of immigrants from 1990 to 2015, keeping in mind that one of Tufte's ideas is "data-to-ink ratio, generally stating that when creating a visualization, you should try to maximize the amount of "ink" (physical or digital) used to represent the data, and minimize the amount used to represent everything else not important to understanding." The x-axis "International migrant stock" from figure 3 is now the y-axis for figure 4. Therefore, similar to figure 3, 1e7 would represent 1 multiplied by 10 to the 7th power.

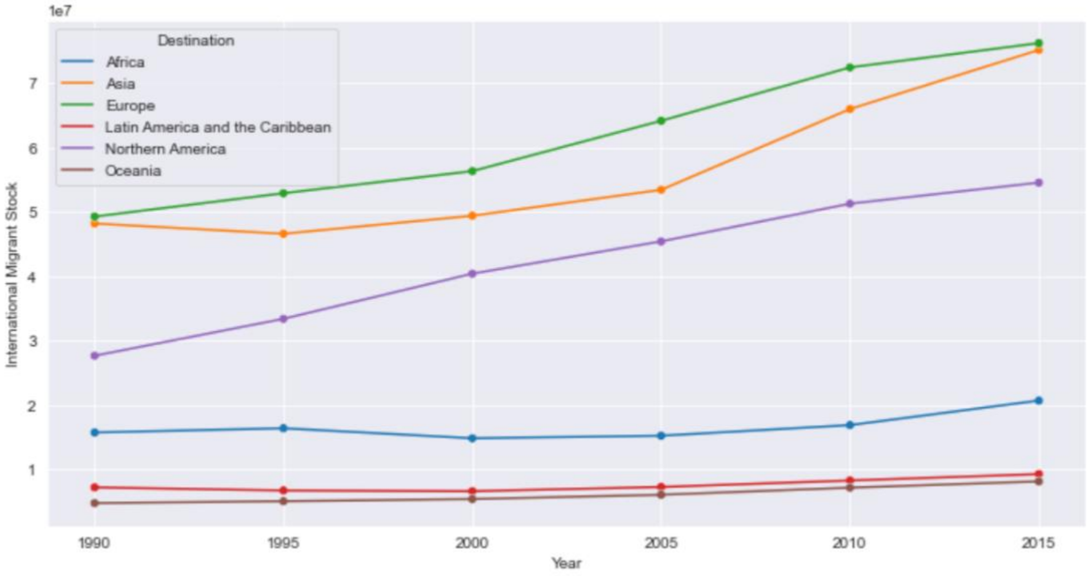


Figure 4. Change trend of the number of immigrants by continent

Here, I have also created a boxplot to look at the distribution of the number of immigrants in six continents. The upper whisker and lower whisker are extremely close for Africa, Oceania, and Latin America and the Caribbean, while the range for Asia, Europe, and Northern America is higher, which presents a more dispersed graph.

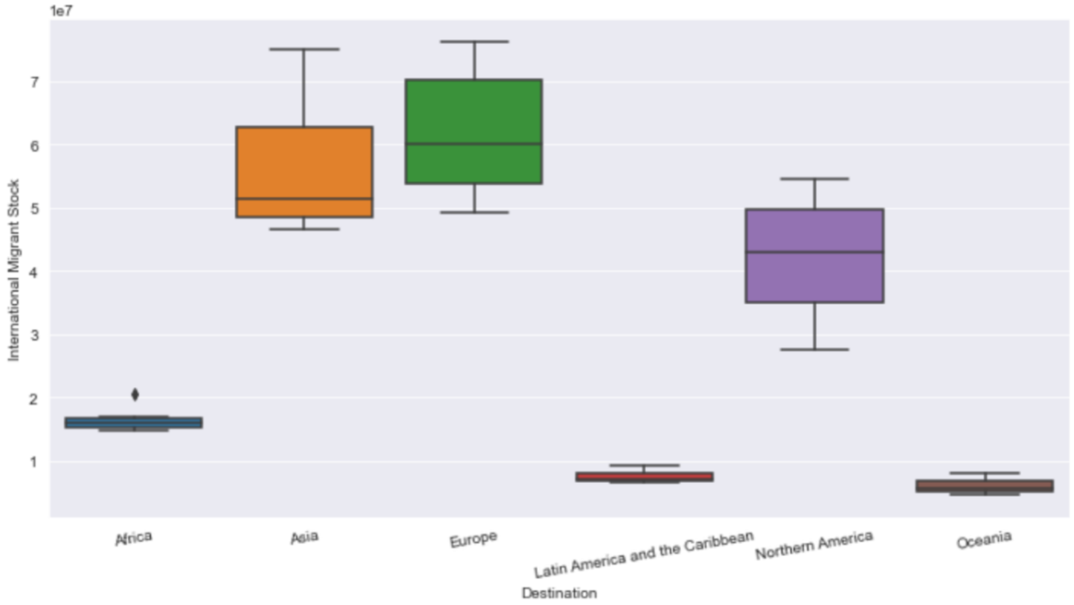


Figure 5. Distribution of migrants in different years on different continents.

Table 2 - Total population at mid-year by sex and by major area, region, country or area, 1990-2015

Same ideas from Tufte’s principles was kept in mind while doing the graphs for this section, I have attempted to emphasize cleanness and minimalism in the graphs I have created for table two. A bar chart and a line graph was plotted to visualize the total population, gender specific population and their changing trends from 1990 to 2015, with a 5-year increment. Both have displayed a linear growth.

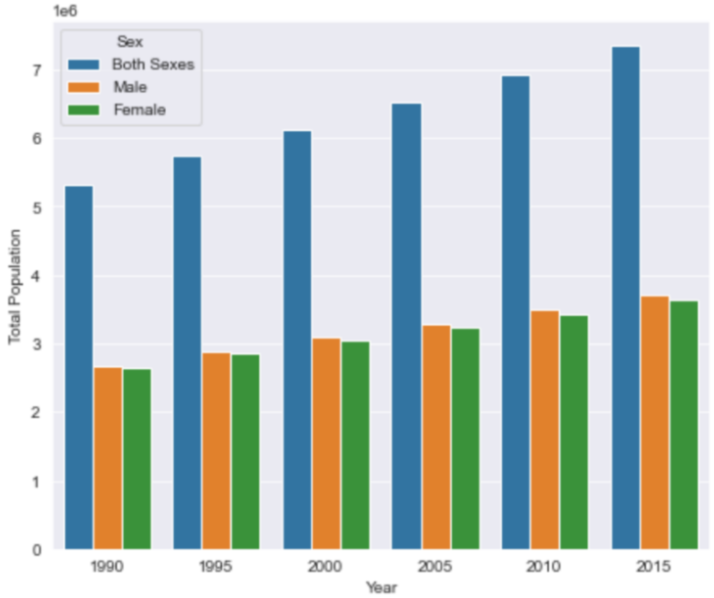


Figure 6. World Population Statistics by Gender.

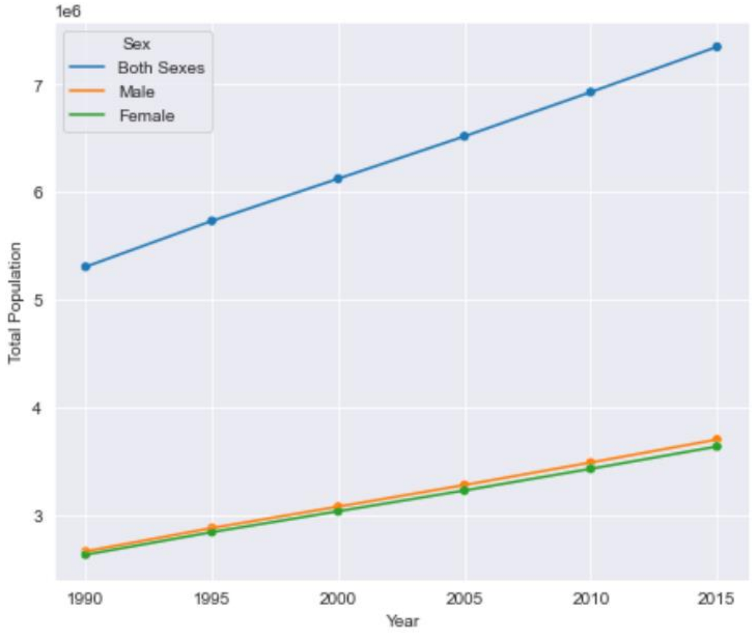


Figure 7. World population change trend by Gender.

The side-by-side bar chart below, figure 8 shows the specific population of each continent in 2015, Asia is showed to have the leading total population.

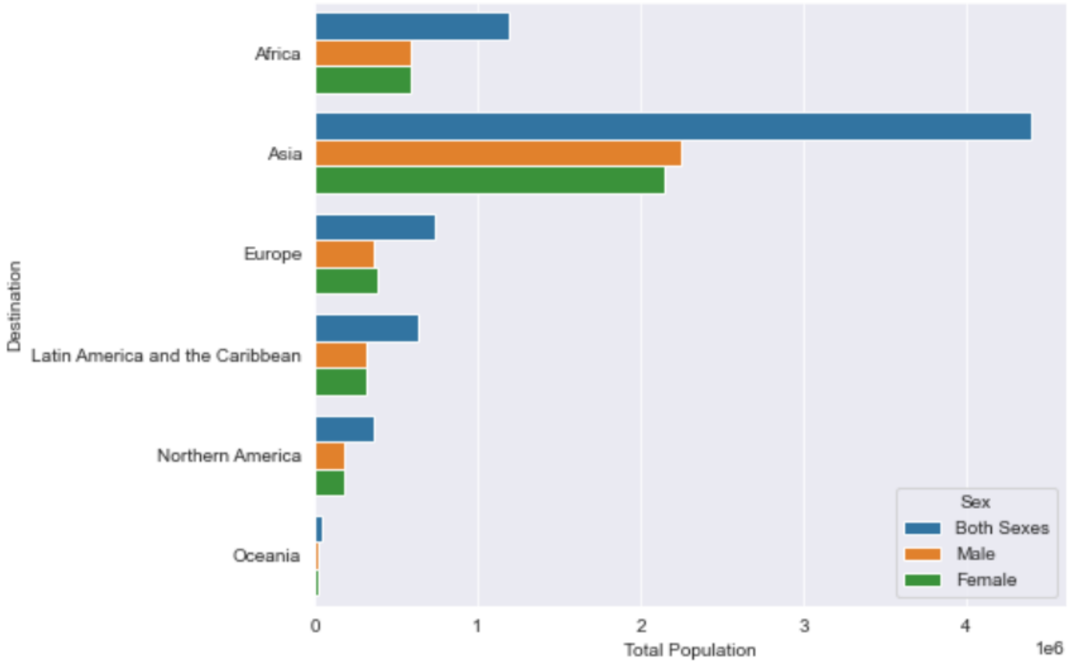


Figure 8. Population distribution of each continent in 2015.

Then, I analyzed the growth trend of the population of each continent, here we can see that Asia presented not only the greatest change, but also the greatest population to begin with in 1990.

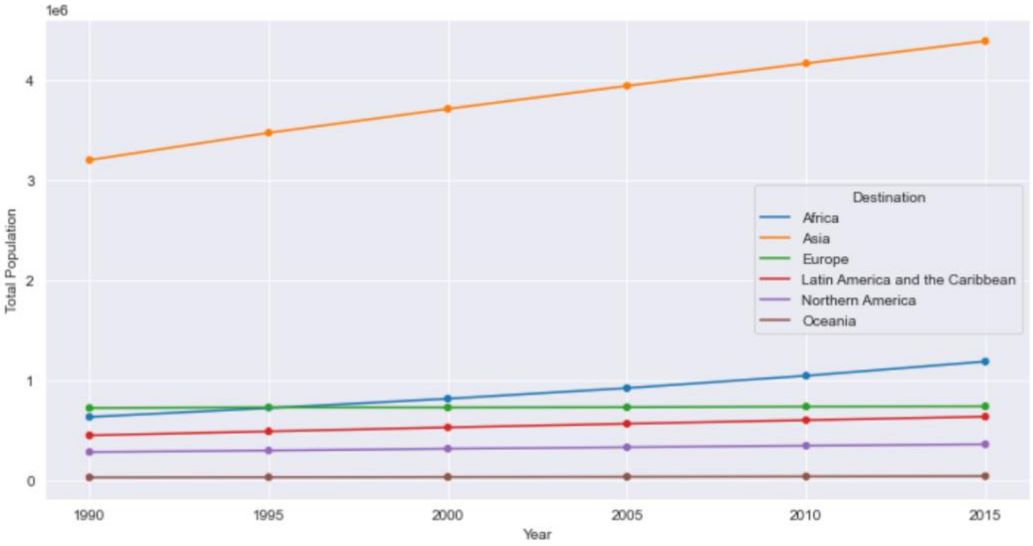


Figure 9. Population change trend by continent.

Furthermore, I thought it would be appropriate to construct a violin plot to visualize the distribution of numerical data. With each continent as the abscissa and the population of different countries in each continent as the ordinate, gender is distinguished by color, in which blue represents men and yellow represents women. We can draw the population distribution of different countries in the six continents by gender.

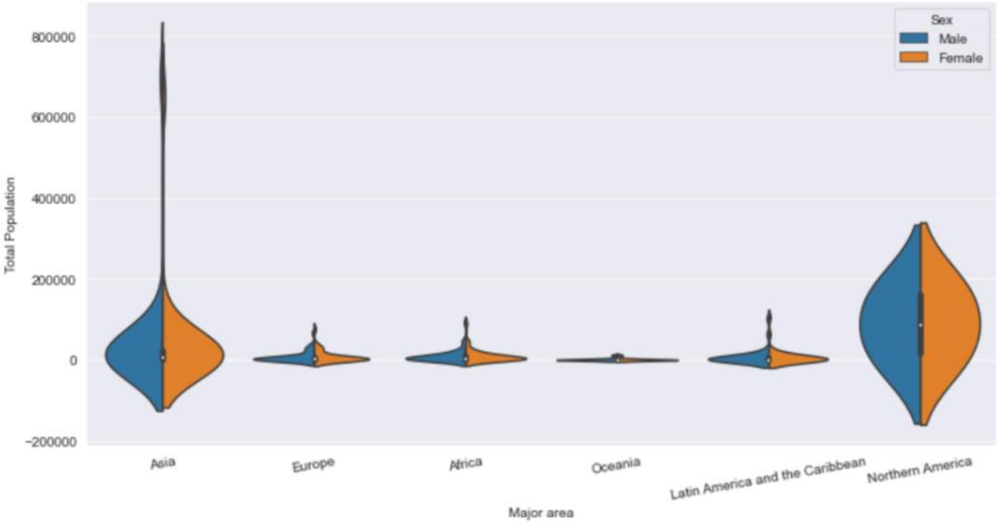


Figure 10. Population distribution of countries in different continents by gender.

Table 3 - International migrant stock as a percentage of the total population by sex and by major area, region, country or area, 1990-2015

According to the data provided in Table 3, we can analyze the percentage of international migrant population in the total population, and also look at whether there are differences between migrant populations of different genders.

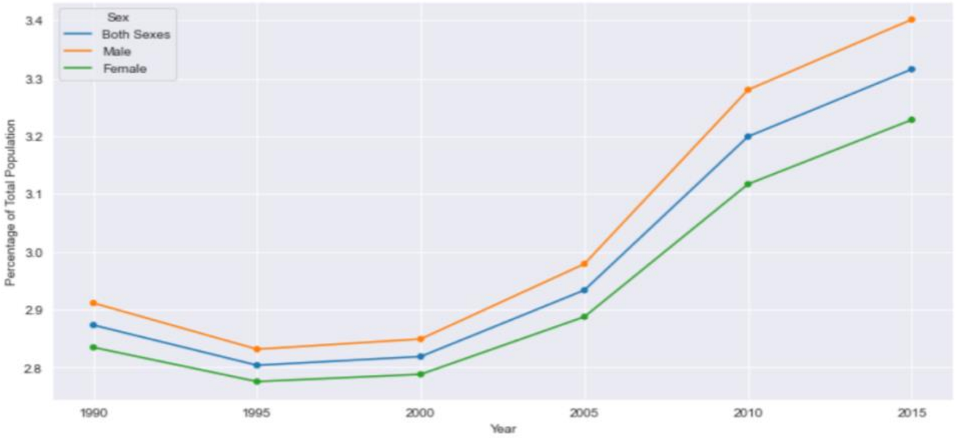


Figure 11. Trends in the percentage of international migrants in the total population.

I have constructed another violin plot here, taking each continent as the abscissa, and the percentage of international migration population in the total population in each continent as the ordinate. I distinguish men and women by color, in which blue is male and yellow is female. I have drawn a distribution map of the proportion of population migration in different countries in each continent.

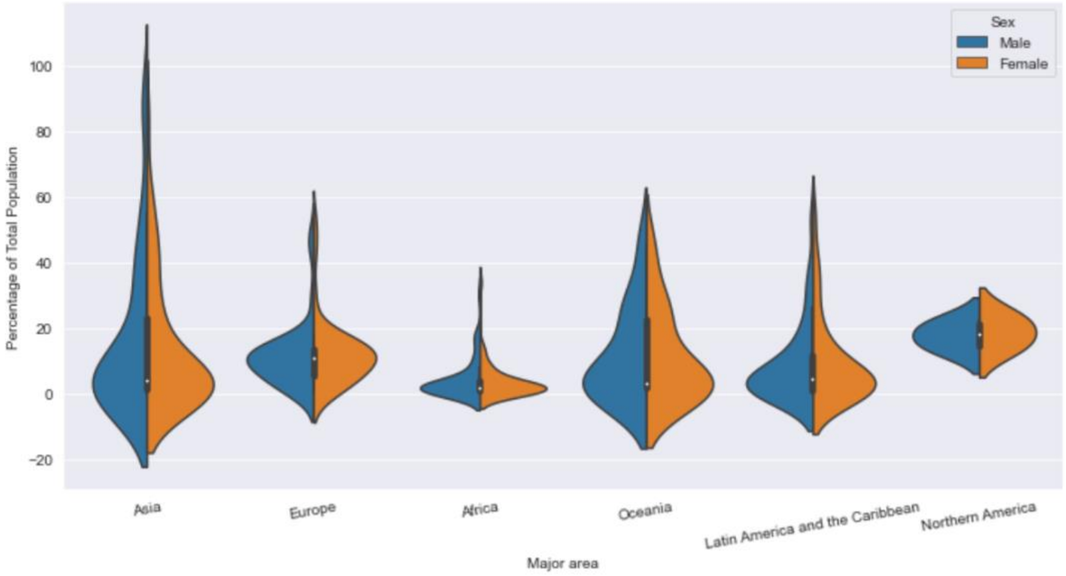


Figure 12. Distribution map of the proportion of population migration by continent.

Table 4 - Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015

In addition, we will analyze the percentage of female migrants in the international migrant population according to the data in Table 4. First, we use time as the abscissa to draw a broken line chart of the proportion of female migrants, so as to reflect the changing trend of the proportion of female migrants.

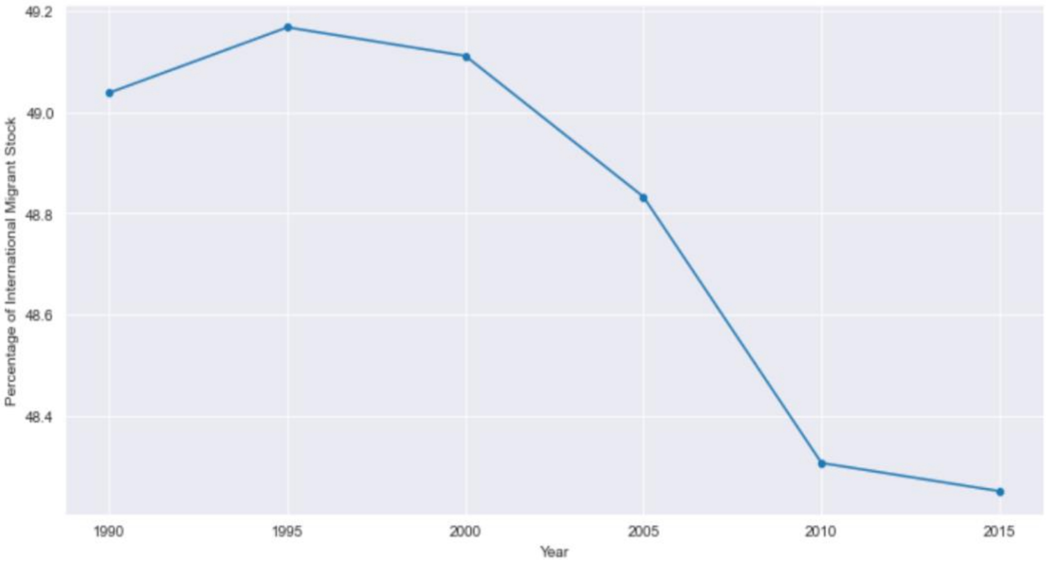


Figure 13. Trends in the percentage of female migrants.

Next, let's analyze the proportion of female migrants in the international migrant population and its changes in the six statistical years from 1990 to 2015 in different continents.

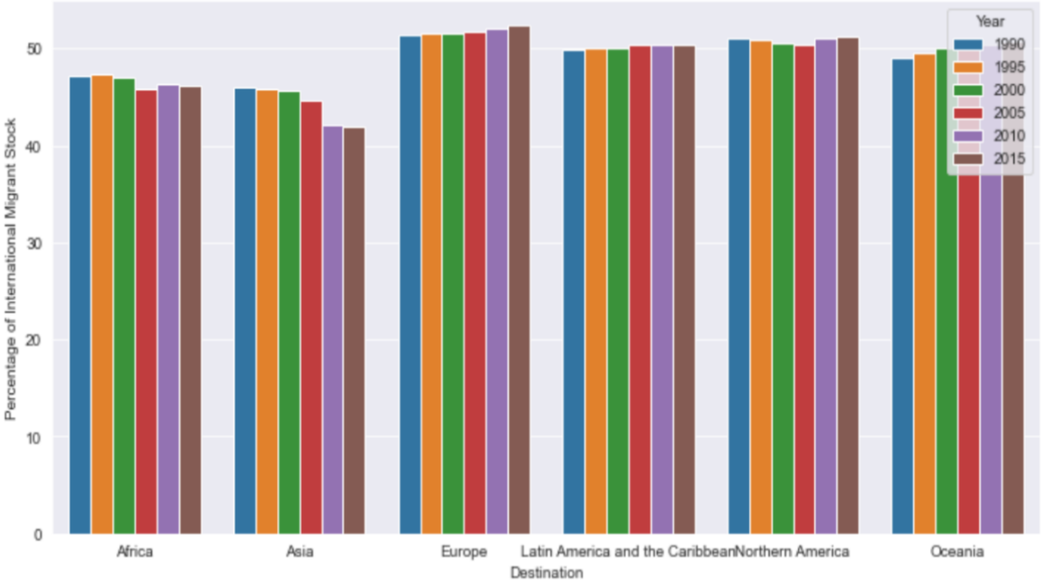


Figure 14. Changes in the proportion of female migrants by continent.

Table 5 - Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015

Table 5 is the data related to the annual change rate of population migration. The line graph below is plotted to present the change trend of migration rate, measured annually.

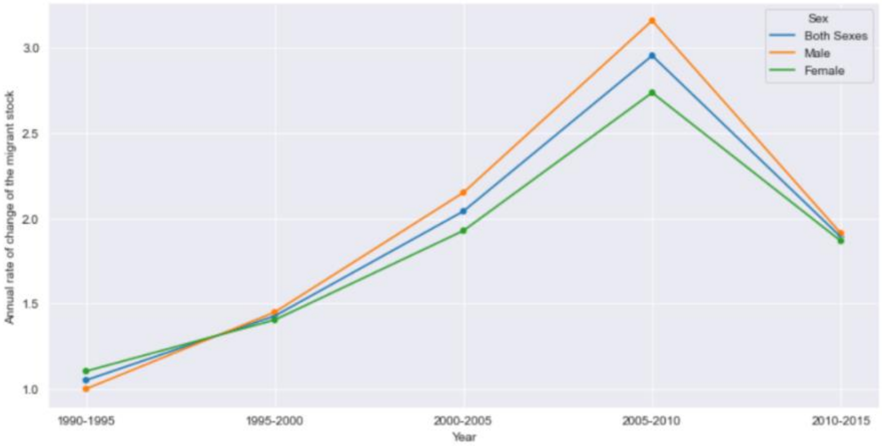


Figure 15. Change trend of migration rate.

Through the distribution of different continents, I can also see that there are differences between men and women in different continents.

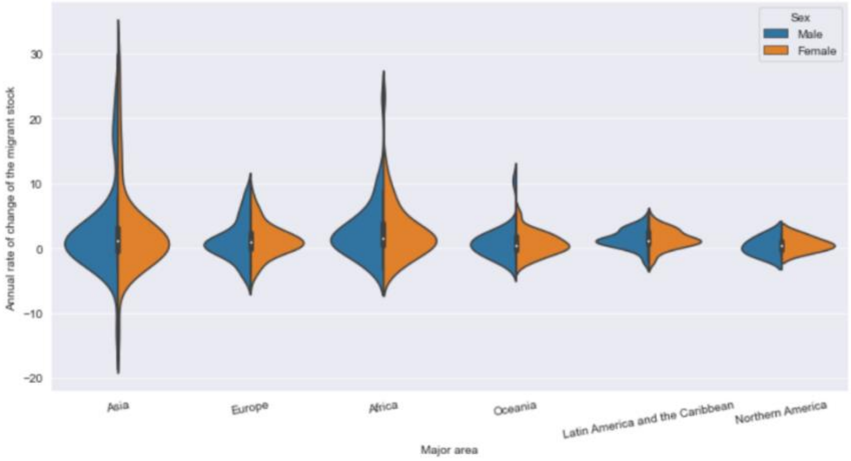


Figure 16. Change trend of migration rate by continent.

Table 6 - Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015

Table 6 is the data related to refugees in the migrant population. I analyze the estimated number of refugees from 1990 to 2015 and the proportion of refugees in the migrant population.

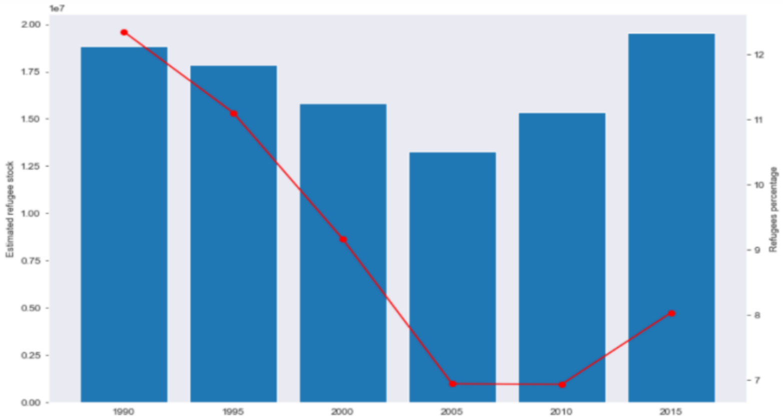


Figure 17. The number and trend of refugees among the migrant population.

For the last graph, I observe the changes in the stock of refugees on different continents, and tried applying the "small principle" from Tufte to create multiple barplot displayed all at once, in order to make allow the readers to compare the inter-frame differences immediately.

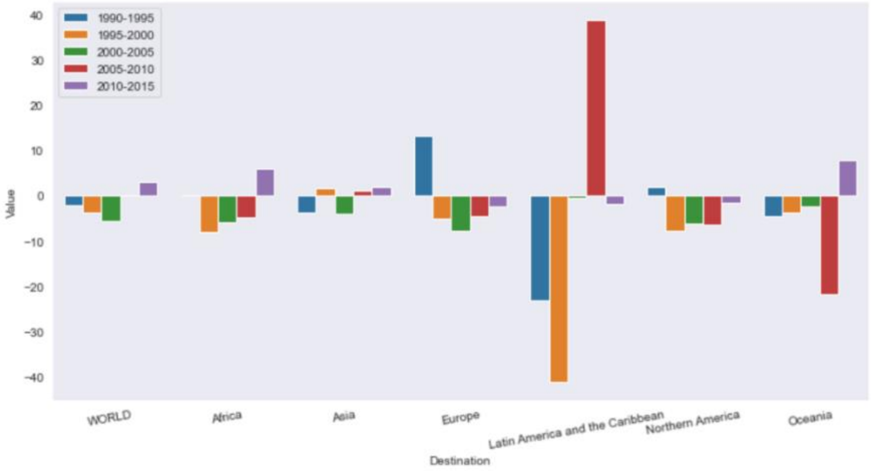


Figure 18. Annual rate of change of the refugee stock.

Discussion

There are numerous takeaways from the UN dataset. As shown in Figure 1 and Figure 2, the number of international migrants has increased year by year. In 1990, there were more than 150 million migrants. By 2015, the number of migrants had increased to nearly 250 million. The growth rate of migration from 1990 to 2005 was slower than that from 2005 to 2015. But overall, in the six years with data statistics, the number of male immigrants is higher than that of female immigrants, and the number of male immigrants is growing faster. This is reasonable as men migrate mainly to work and employment, they are responsible for their families.

As shown in Figure 3, Europe has the largest number of immigrants, followed by Asia, North America has the third largest number, and Oceania has the smallest number of immigrants. In terms of the number of male and female migrants, there are also some differences between different continents. It can be seen that Africa, Asia and Oceania all have more male migrants than female migrants, of which Asia has the largest difference, while Europe, North America and Latin America have more female migrants than male migrants, of which Europe has the largest difference. However, looking at the changing trend of the number of immigrants from 1990 to 2015 displayed in Figure 4, it can be seen that the number of immigrants in all continents is on the rise as a whole, but the relationship between the number of immigrants in six continents remains unchanged. In the six different years with statistics, the number of immigrants in Europe, Asia, North America, Africa, Latin America and Oceania is from more to less. In terms of change trend, Latin America and Oceania have a relatively small change, with a relatively gentle upward trend. The number of African migrants declined briefly from 1995 to 2000, and then rose again. The number of migrants in North America has been rising, but the growth rate has slowed down. Since 2005, the number of migrants in Asia has increased faster, and by 2015 it was close to the number of European migrants. Although the number of European migrants has increased year by year, it has slowed down.

Figure 5 was a boxplot indicating the distribution of migrants in different years on different continents. We can see that the number of immigrants in Africa, Latin America and Oceania has not changed much, and the data is relatively centralized. The number of immigrants in Asia, Europe and North America has changed greatly, and the data distribution is relatively scattered. The number of immigrants in Asia is between 50 million and 60 million. The number of immigrants in Europe is between 55 million and 70 million. The number of immigrants in North America is between 35 million and 50 million. The maximum number of immigrants in Asia is close to that in Europe. In general, Asia, Europe, and North America offers better work and economic opportunities, better study programs and school, which makes people want to seek a better life, thus, making migration rates rise.

In addition, figure 6 to 10 showed the data. for table 2. It can be seen from Figures 6 and 7 that the total population of the world is increasing year by year, and the number of men and women is also growing. However, the growth trend of the number of men is slightly higher than that of women, and the difference between the number of men and women is further widening. From figure 8, Asia has the largest population, far more than other continents, and about three times that of Africa, the second largest. Oceania has the smallest population. From the perspective of male female ratio, the number of males in Asia is obviously more than that of females. The male female ratio in Africa, South America, North America and Oceania is relatively balanced, close to 1:1. The number of females in Europe is more than that of males.

From Figure 9, we could see that the growth rate of the population of each continent is slightly different. The growth rate of Asia and Africa is fast as there is more demand for labor, and for Europe it is almost zero, and the growth rate of the population of Latin America, North America and Oceania is also very slow. As shown in Figure 10, it can be seen from the figure that the distribution of the number of men and women in the same continent is basically the same, but the population distribution gap between different continents is large. Among them, there are several countries in Asia with a large male population and a large female population, which are far more than most other countries in Asia.

Moving on to table 3, as shown in Figure 11, from 1990 to 2015, the percentage of international migrant population in the total population first showed a downward trend and then an upward trend. In the six years with statistics, the proportion of male migrant population was higher than that of female migrant population. Additionally, the figure that the proportion of population migration varies from continent to continent. The overall distribution of migration proportion in North America is relatively concentrated, and the proportion of migration is relatively high. In Asia, Europe, Africa, Oceania and Latin America, there are some countries whose proportion of population migration is far greater than that of most other countries in the same continent, and the distribution is long tail, especially in Asia, where the proportion of population migration is relatively scattered. In the same continent, the proportion of migration of different genders is basically the same, but there are also different places. For example, in Asia, the number of countries with a higher proportion of male migration is more than that of countries with a higher proportion of female migration; In North America, the proportion of female migration is more dispersed than that of male migration, and there are more countries with higher proportion of female migration.

The fourth table from the UN dataset outlines the female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015. It can be seen from Figure 13 that from 1990 to 2015, the proportion of female migrants never exceeded 50%, and the highest proportion was in 1995, accounting for nearly 49.2%. From the perspective of change trend, the proportion of female migrants showed an upward trend from 1990 to 1995, and a downward trend from 1995 onwards. From 2005 to 2010, the decline rate was the fastest, and from 2010 to 2015, the decline rate slowed down. It can also be seen from Figure 14 that the proportion of female migrants is the highest in Europe and the lowest in Asia. Over time, the proportion of female migrants in Africa and North America decreased first and then increased. Europe, Latin America and Oceania all showed a slight upward trend, while Asia showed a significant downward trend, with the largest decline from 2005 to 2010.

According to the broken line chart under the analysis for table 5, the migration trend of men and women remain the same, with an increasing migration rate from 1990 to 2010 and a slight decline from 2010 to 2015. However, the migration rate of men has been increasing, which was still lower than that of women in 1990, but higher than that of men after 2000. This also shows that men have a higher desire to migrate. In Africa and Europe, the distribution of migration rates for men and women is almost the same. In Asia and Latin America, the distribution of migration rates between men and women is only slightly different. In South America, the mobility variance of men is higher than that of women, which shows that men's life in South America is not stable. In Oceania, the migration change rate of women is higher than that of men. It can be seen that women have more desire to like Oceania more than men.

The last two figures relate to table 6 of the UN dataset, which estimates the refugee stock at mid-year by major area, region, country or area, 1990-2015. According to the histogram in figure 17, refugees decreased year by year from 1990 to 2005 and increased year by year from 2005 to 2015. There were more refugees in 2015 than in 1990. According to the line chart, the proportion of refugees decreased year by year from 1990 to 2005, remained stable in 2010, and increased in 2015. According to the information in the two charts, the number of refugees has gradually increased in recent years, with only a large number of refugees migrating around 2010. In the rest of the time, the proportion of refugees in the migrant population was relatively stable. For the last graph in figure 18, changes in most continents are consistent with the world. Only Latin America has a large number of refugee flows, which is the main reason for the increase of refugees in the world from 2005 to 2010.

Conclusion

With consideration of Tufte’s principles and ideas during explanatory data analysis on the six UN dataset tables, I found that the total population of the world is increasing year by year, and the number of international migrants is also increasing year by year. The proportion of refugees in the migrant population has also shown an upward trend in recent years. In addition to the impact of globalization, to a certain extent, the increase in the number of refugees is also an important factor causing the increase in the number of immigrants.

The number of migrants varies greatly in different regions. The number of migrants in Europe is the highest all the year round. Although Asia has the largest population, the number of migrants is not as large as that in Europe. When comparing data from different continents and regions, there are also some interesting conclusions, especially when distinguishing by gender. There are more women in Europe, more men in Asia and Africa, and the corresponding European female migrant population is more than men, while the male migrant population in Asia and Africa is more than women. There may be gender discrimination in some countries in Asia and Africa.

The changes of the international migrant population deserve our attention. We can use the analysis of these data to formulate coping strategies in advance, so as to better adapt to the opportunities and challenges brought by the