UNIVERSITY OF TORONTO
FACULTY OF INFORMATION
The iSchool

INF 1340H F LEC0101 20229
Programming for Data Science

Final Project

Descriptive Statistics + Data Visualization

Student Name
Miaomiao Yang

Dec 15th 2022

Table of Contents

Table of Figures

1. Introduction

The midterm project focuses on wrangling messy datasets into a tidy format, and the final project is a continuation of the midterm project. The final project focuses on data visualization based on the tidy tables that were cleaned in the midterm project. Data visualization can translate information into a visual context, such as a map or graph, to make it simpler for the brain to interpret and draw conclusions (Kate Brush, 2022). The objective of the final project is to visualize the six UN datasets to make it easier to identify trends, patterns, and outliers in those large datasets. It also provides an opportunity to learn and compare various plot methods. All six tables will be visualized by **Google Colab**.

2. Data Description

By looking at six tables, all table involves data from 1990 to 2015; meanwhile, it's found that all tables have the column of 'Major area, regions, country or area of destination,' which involves the *world*, *developed and developing regions*, *major continents* as well as all *countries* around the world. The world is divided into developed and developing regions. From the Notes of Trends in International Migrant Stock, the developed regions comprise Europe, northern America, Australia/New Zealand, and Japan. In contrast, the developing countries include all areas of Africa, Asia (except Japan), Latin America and the Caribbean, plus Melanesia, Micronesia, and Polynesia. The major continents comprise Sub-Saharan Africa, Africa, Asia, Europe, Latin America and the Caribbean, and Oceania. This information is critical when doing visualization to target different audiences by regions or by purposes.

The next column to note is 'Type of data', it is not difficult to find that, except *Table 2*, all the other five tables have this column. The different letter indicates different types of data. There are four types of data in total.
- Type B refers to the estimated foreign-born population,
- Type C refers to foreign citizens,
- Type R refers to the number of refugees plus the estimate of international migrants,
- Type I refers to countries or areas that have no international migrants.

Different Table has their focus point. Some tables focus on the migrant population, and some look at the percentage. For example, *Table 1* focuses on the international migrant stock by gender; *Table 2* focuses on the total population by gender, while *Table 3* looks at the international migrant stock as a percentage of the total population by gender. *Table 6* differs from the other five tables because it concerns both

population and percentage change. Understanding and identifying each table's characteristics allow us to visualize the data better.

3. Methods

3.1 Tufte's Principles

Edward R. Tufte defined six principles implemented in the project (Tufte, 1983). These principles are laid in his book "The Visual Display of Quantitative Information," which are very beneficial for data visualization.

- Principle 1: The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.
- Principle 2: Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
- Principle 3: Show data variation, not design variation.
- Principle 4: In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
- Principle 5: The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
- Principle 6: Graphics must not quote data out of context.

The six principles were kept in mind. Misrepresenting data may lead to misunderstanding, which can have severe repercussions. The best visualization should be comprehended with few words, but it must be clear. A dashboard should be aesthetic and informative. Choosing the right color palette will make the dashboard clean and sophisticated. Chart design is tricky compared to picking colors. Simple chart designs could be more efficient and are readily comprehended by a large audience. Make sure to use units; any visualization with units is meaningful and clear. Don't make the visualization complicated; simply don't use 3-D for 2-D objects. Never quoted out-of-context, thereby changing the meaning of the information being conveyed.

3.2 Further cleaning and different packages used in Tables

Each table has been recleaned or further cleaned in varying degrees according to different purposes based on the original midterm cleaning. In the midterm project, all the tables were cleaned out the rows with missing 'Type of data'. Those rows involve *the world, developing and developed regions,* and *continents*; however, some of the tables were visualized based on major areas; thus, recovering these rows is necessary.

Before beginning to process data visualization, the initial stage in the project is to load in the *visualization* packages, including the *datascience package*, and *Plotly.express* provides good-looking graphs with few code lines and is interactive, while *Seaborn* offers a high-level interface for drawing attractive and informative statistical graphics. Visualizing Tables 1, 2 and 3 using *plotly.express*, while Table 4,5,6 uses *seaborn*.

3.3 Methods used in each table

Different visuals would be applied in each table to serve different decisions and convey various messages. Lineplot were applied in most of the tables. As per requirement, the method of boxplot and boxenplot were adopted as well. The following table organized the selected areas, sexes, parameters, and chart types applied in data visualization of each UN table.

Table 1. The details and methods used in data visualization of six UN tables.

Table	Figure	Areas	Sexes	Parameters	Chart
1	1_1	Countries	Both sexes	Internal migrant stock	Histogram
	1_2	Developed & developing regions	Both sexes		Line
	1_3	Continents	Both sexes		Line
2	2_1	Developed & developing regions	-	Total population (thousands)	Bar
	2_2_1 2_2_2	Developed & developing regions Continents	-		Scatter
	2_3	Developed & developing regions Continents	-		Bar
3	3_1	Countries	Both sexes	International migrant stock as a percentage of the total population	Line
	3_2	Countries	Male & female		Line
4	4_1_1	Countries	Female	Female migrants as a percentage of the international migrant stock	Box
	4_2	Developed & developing regions			Hist
	4_3_1	Continents			Point
	4_3_2	Continents			Point & replot
5	5_1	countries	Male & female	Annual rate of change of the migrant stock	Line
	5_2	countries	Male & female		bar
6	6_2	Countries	-	Refugees as a percentage of the international migrant stock	Line
	6_4	Countries	-	Annual rate of change of the refugee stock	Boxen

3.3.1 Methods used in Table 1

Table 1 contains internal migrant stock as mid-year for all observations. One histogram and two line charts were created to visualize which type of data was most collected and the overall trends of international migrant stock in regions and continents by using *plotly.express*. All three plots focus on both sexes; thus, the data of male-only and female-only were dropped. Also, the three figures target different area categories, including all countries, developed and developing regions, and continents. Remember Turfte's Principle 2: Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. An appropriate title and give x and y value's label should be added in each plot to avoid ambiguity.

3.3.2 Methods used in Table 2

Table 2 contains the total population at mid-year by gender. Two bar charts and two scatter chart was created. Each chart has plot of both sexes, male-only, and female-only. One bar chart focuses on the comparison between developed and developing regions; another bar chart involves regions and continents to see which continents contribute more to the corresponding regions. The scatter plot of the total population in developed and developing regions and continents compare the increasing tendency between males and females. Each marker was connected by a line to show the tendency. While creating scatterplot, setting columns for regions and continents lead to ambiguous observation. In order to making observation of changes and comparisons clearer, creating another scatter plot by setting columns for both sexes, male, and female.

3.3.3 Methods used in Table 3

Visualization of all countries was determined for table 3, which allowed us to analyze and observe the International migrant stock as a percentage of the total population of each country. Table 3 was split into data of 'both sexes' and 'male' and ''female'. The data of 'both sexes' for overviewing migrant stock percentage changes in each country, while the data of 'male' and 'female' for observing the changes and comparison between male and female in each country. Both diagrams created by lineplot.

3.3.4 Methods used in Table 4

Four diagrams were plotted using library *seaborn* and presented in boxplot, histoplot, and pointplot for visualizing Table 4. Boxplot enables us to study the distributional characteristics of Female migrants as a

percentage of the international migrant stock corresponding to each year. Boxplots are used in descriptive statistics to explain data. Box plots usually show data quartiles and averages to highlight data distribution and skewness (Mcleod, 2019). The histogram, actually is a bar plot but condensed under series into an easily interpreted visual which indicated the interval that the data female migrants as a percentage of the international migrant stock lie in. While plotting the pointplot, it was unable to show error bar with all continents, then recreating a new pointplot by *seaborn.replot.* Error bars based on **confidence intervals (C.I)** only tell reliability of a measurement.

3.3.5 Methods used in Table 5

Visualizing table 5 using two different plots, which are lineplot and barplot. Both methods can plot aggregated values with 95% confidence intervals. The lineplot shown error band, while the barplot shown error bars. Maintaining data of male and female in all countries for comparison of the annual rate of change of the migrant stock.

3.3.6 Methods used in Table 6

The data of table 6 were splitted into three groups of data, Estimated refugee stock at mid-year (both sexes), refugees as a percentage of the international migrant stock, and the annual rate of change of the refugee stock. In this project, only refugees as a percentage of the international migrant stock and the annual rate of change of the refugee stock will be plotted. Two figures will be created using lineplot and boxenplot to indicate the trend of refugees as a percentage of the international migrant stock and the data distribution. Boxenplot is similar to a box plot. By plotting more quantiles, it provides more information about the shape of the distribution, particularly in the tails.

4. Results & Discussion

In this section, all tables' plots would be listed and discussed one by one.

4.1 Table 1 - International Migrant Stock at mid-year by sex and by major area, region, country, or area, 1990-2015

Figure 1_1 looks at the distribution of typfe of data*. Before any interpretation of *figure 1_1*, two columns of type of BR were found. The column repetition may be because the data has not been cleaned up thoroughly. In fact, in 'Notes', there were only four types of data in the dataset: B, C, I, and R. Although the incomplete cleaned data affected the visualization, the figure still can be interpreted. Among the four

data types, Type B has the most amount of data and increased by years. The amount of collected data Type I was similar from 1990 to 2015 and is very small; the amount of collected data Type B, C, and R varied by degrees from 1990 to 2015.
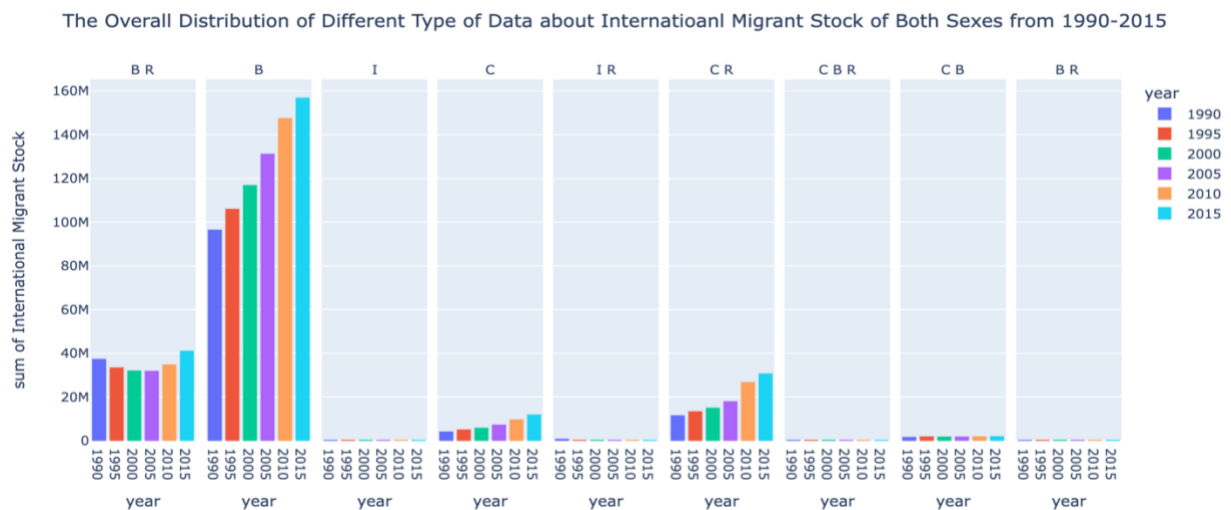

Figure1_ 1 The Overall Distribution of Different Type of Data from 1990-2015

Obviously, *figure 1_2* and *figure 1_3* use line graphs, which best show how international migrant stock changes and compare changes between developed and developing regions or continents over time. *Figure 1_2* indicates the international migrant stock of developed regions and developing regions both increased. From 1990 to 2015, the international migrant stock of developed regions was always larger than developing regions. The difference kept growing and reached a peak from 2005 to 2010.
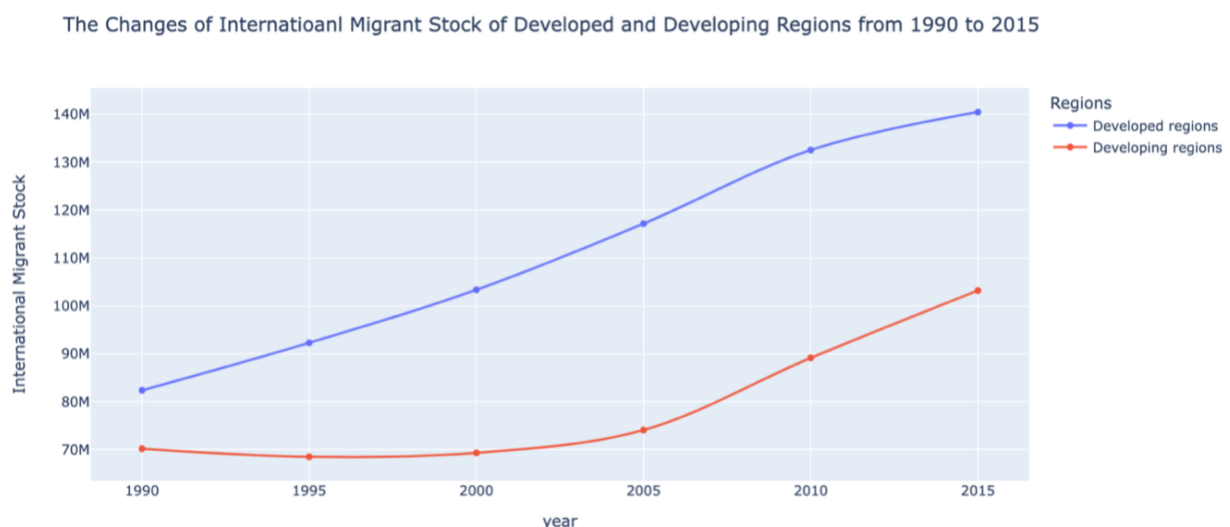

Figure1_ 2 The Changes of International Migrant Stock of Developed and Developing Regions from 1990 to 2015

The Changes of Internatioanl Migrant Stock of Different Continents from 1990 to 2015
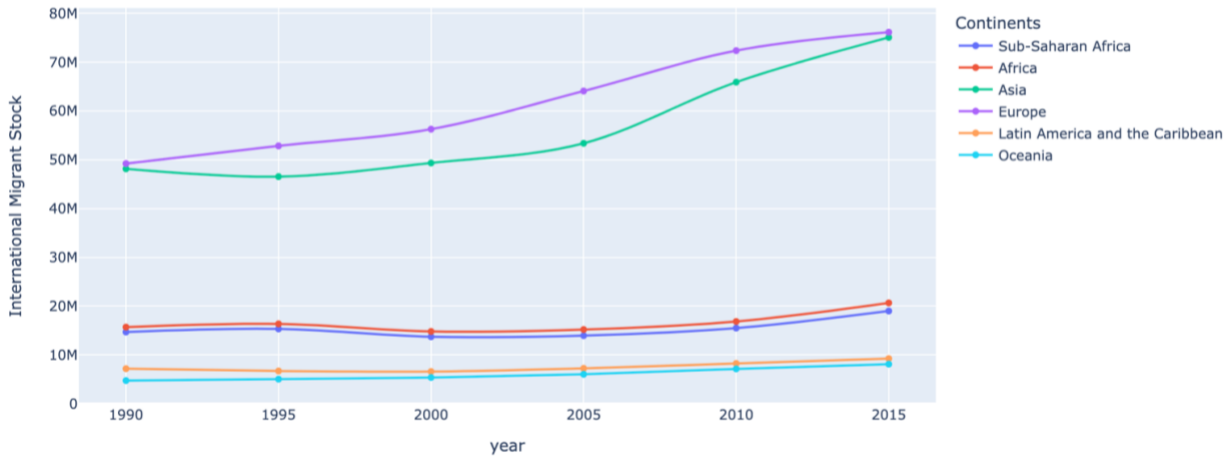
Figure1_ 3 The Changes of International Migrant Stock of Continents from 1990 to 2015

Figure 1_3 indicated the changes of international migrant stock of different continents over time. Clearly, the overall migrant stock increased in all continents from 1990 to 2015. However, the migrant stock in Asia was decreased from 1990 to 1995; while the migrant stock in Africa and sub-Saharan Africa was also decreased but was in 2000. Europe and Asia have most migrant stocks compare to Africa, sub-Saharan Africa, Latin America and the Caribbean, and Oceania, and the increasing rate of those two continents is great higher than other continents.

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

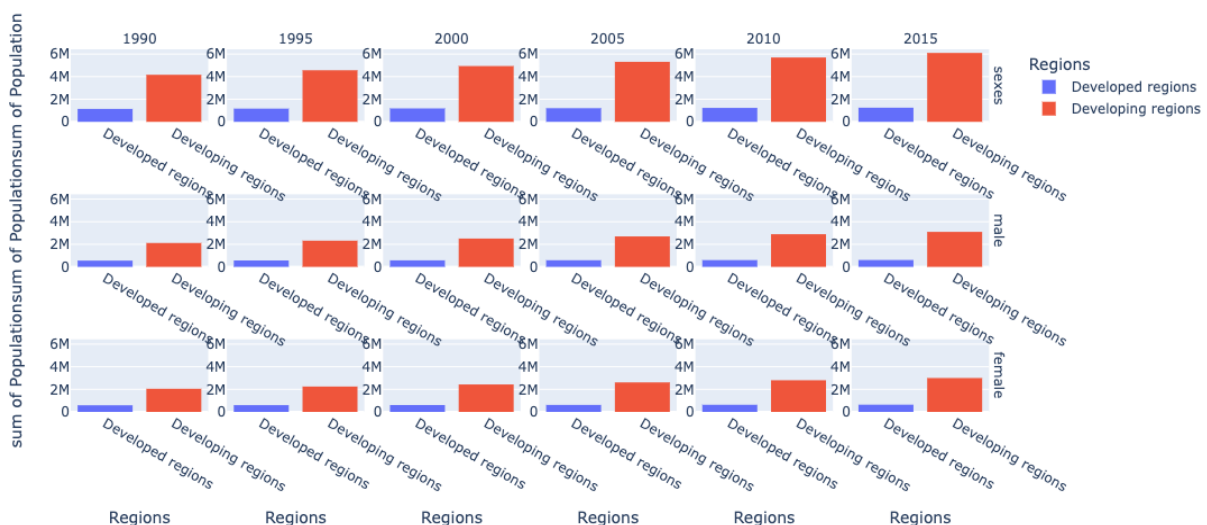Comparision of Total Population of Developed and Developing Region from 1990-2015

Figure2_ 1 The Changes of International Migrant Stock of Developed and Developing Regions from 1990 to 2015

From *figure 2_1,* the total population of developing regions is significantly greater than developed regions no matter both sexes, male, or female. The population of the developing regions kept growing from 1990 to 2015. It was a bit unnecessary to plot male-only and female-only together with both sexes, as the plots of both sexes can say it all. From the *figure 2_2, the* rising tendency of population in developed regions is very slow compared to population in developing regions. The female population in developed region is a bit higher than male population and the raising tendency of female population and male population is similar; in contrast, the male population in developing region is larger than female population and the rising tendency of the male population is higher than female population.
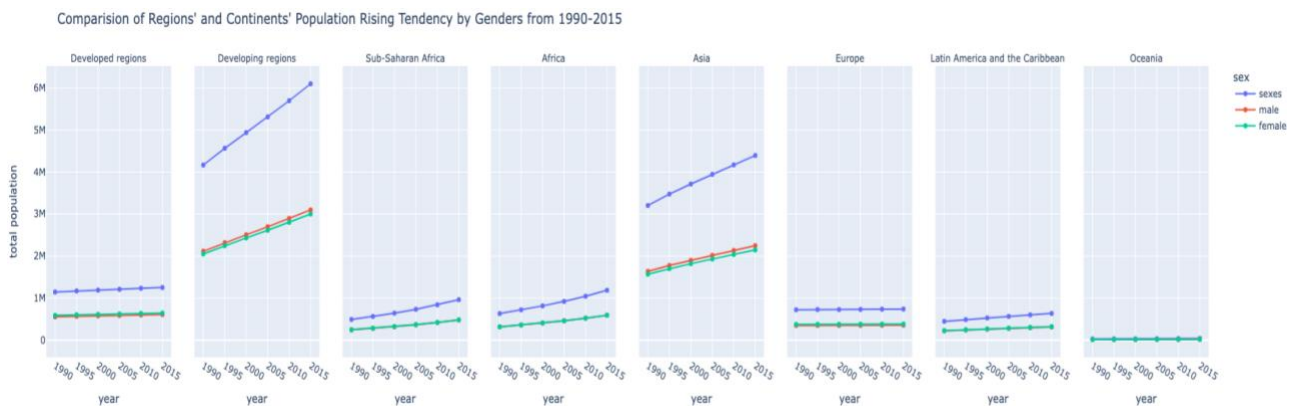

Figure2_ 2_1 Comparison of Regions' and Continents' Population Rising Tendency by Areas from 1990-2015

It is worth mentioning that except for Asia, the points representing male and female populations in other continents overlapped, which means the population of female and male increased synchronously; furthermore, the points of both sexes overlapped with the female in Oceania, which shows the female population dominated Oceania. There was no apparent change in population in Europe from 1990 to 2015.
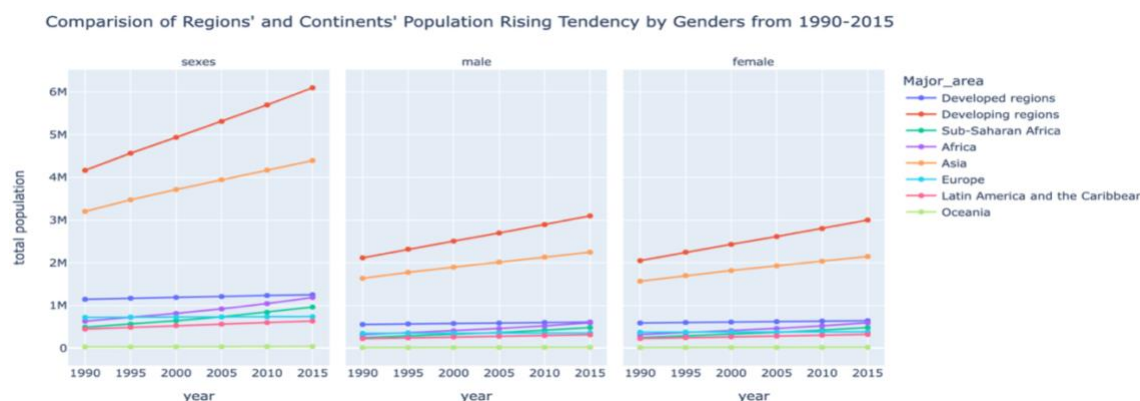

Figure2_ 3_2 Comparison of Regions' and Continents' Population Rising Tendency by Genders from 1990-2015
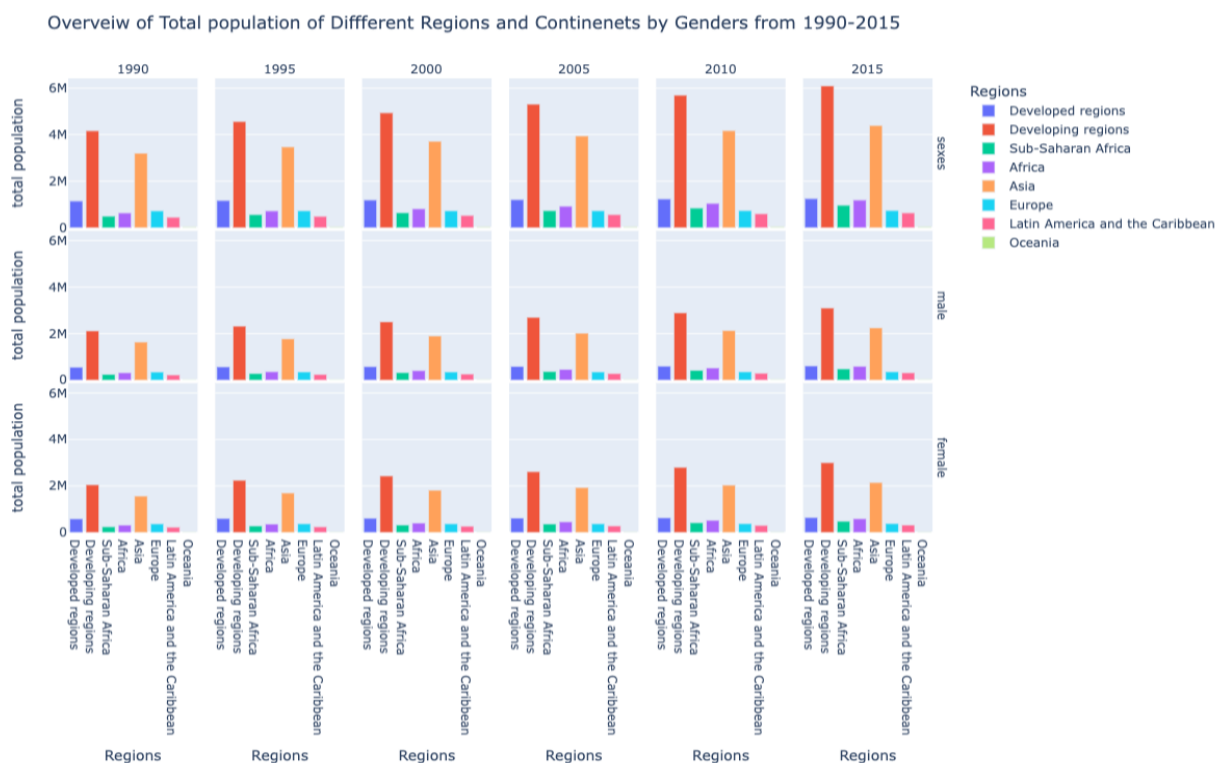
Overveiw of Total population of Diffferent Regions and Continenets by Genders from 1990-2015

Figure2_ 4 Overview of Total population of Different Regions and Continents by Genders from 1990-2015

No matter in which year, Asia contributed the most population in developing regions or the world (*figure2_3*). The next continents with comparable higher population are Africa and sub-Saharan Africa. Oceania had the least population.

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

Due to the limited space in the report, only several countries were attached for showing. In *figure 1_3*, taking Comoros and Mayotte as an instance, the international migrant stock as a percentage of the total population in Comoros is nearly 0 and there have been almost no changes within 25 years. However, there has been a significant change in Mayotte which kept increasing from 1990 to 2005, and gradually decreased after 2005.

From *figure3_2*, it was still taking Comoros and Mayotte as examples. The male line overlapped the female line in Comoros, which are horizontal. This indicated no difference between male' and female' international migrant stock as a percentage of the total population from 1990 to 2015; also, the
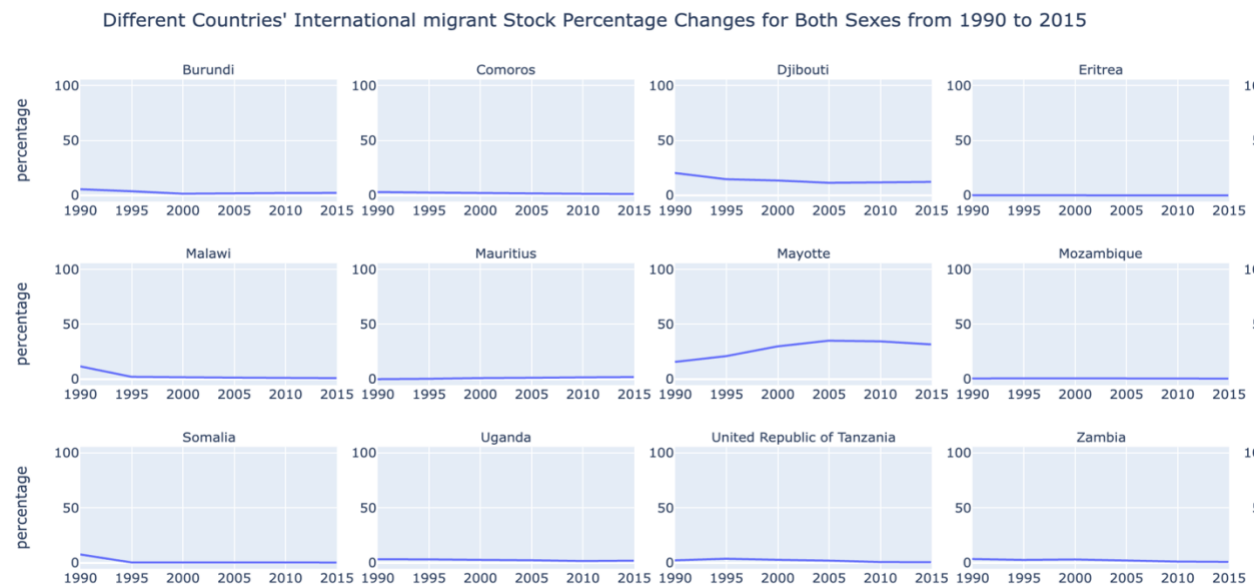
Different Countries' International migrant Stock Percentage Changes for Both Sexes from 1990 to 2015

Figure 3_ 1 Different Countries' International Migrant Stock Percentage Changes for Both Sexes from 1990 to 2015

percentage of both male and female was kept constant, with no increasing or decreasing. The international migrant stock as a percentage of the total male population was higher than that of female in 1990; however, this percentage started to reverse in 2005.
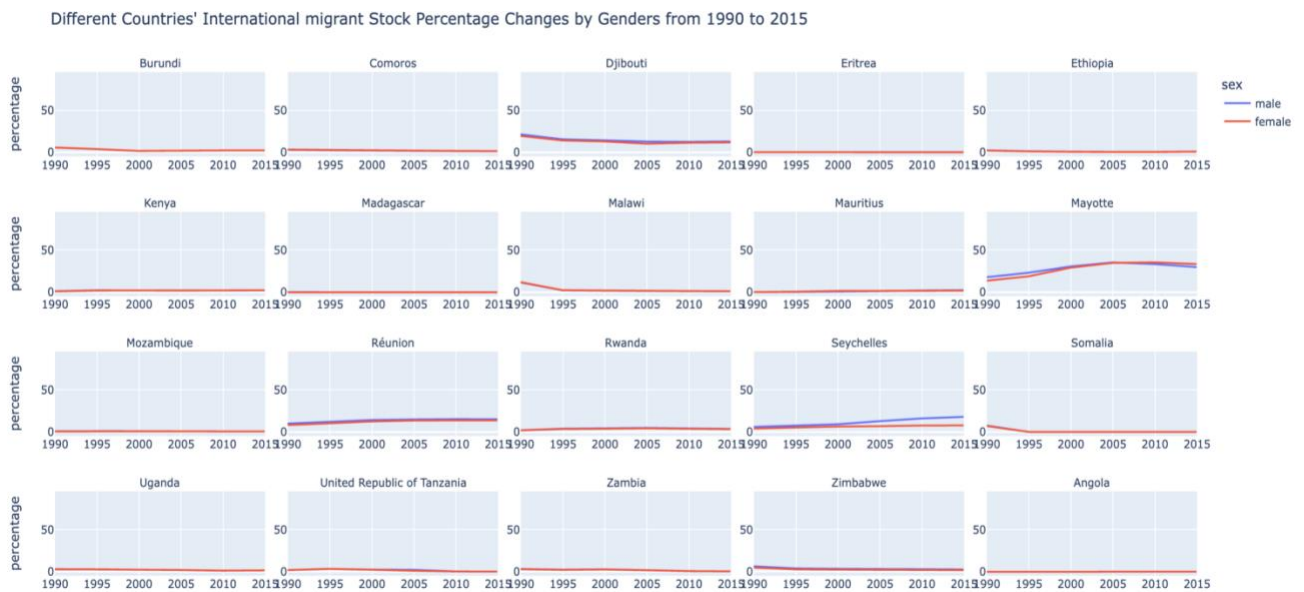
Different Countries' International migrant Stock Percentage Changes by Genders from 1990 to 2015

Figure 3_ 2 Different Countries' International Migrant Stock Percentage Changes by Genders from 1990 to 2015

During the visualization of table 3, the operation time increased, and the computer crashed several times. The webpage ***google colab*** may only be suitable for running a small number of plots simultaneously. This required us to plot the specified diagrams based on requirements.

From *figure 4_1_1*, as time moved on, the Female migrants as a percentage of the international migrant stock increased gradually from about 47% in 1990 to 49% in 2015. There also appears to be a slight decrease in 2005. Points outside whiskers show an extremely low and high female percentage. From the box of six years, it seems the year 2015 has the longest box, which means that year has more dispersed data while the year 1995 has less dispersed data. Relatively speaking, the year 2000 to 2015 had larger ranges of box, which indicated wider distribution, that is, more scattered data.

The boxplot was also created by *plotly.express* and attached in **Appendix**. By doing so, to compare the difference between boxplots *(figure 4_2 & figure 4_3)* by two packages. There is no discernible difference between the two plots. However, *plotly.express* provide same good looking graphs with few code lines.
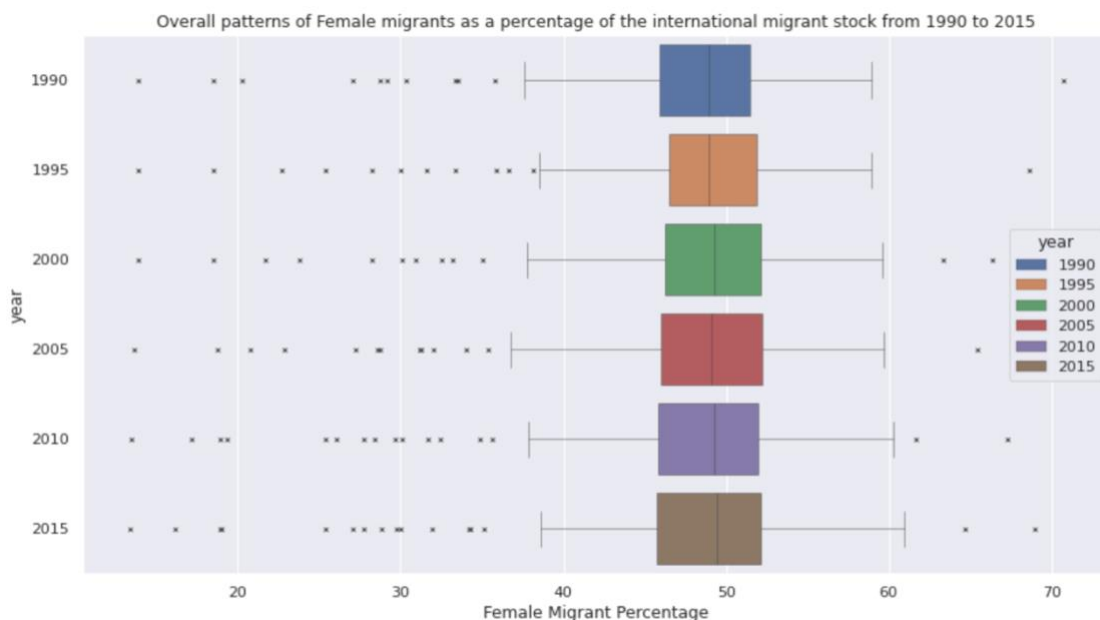

Figure4_1_ 1 The Overall patterns of Female Migrants as a Percentage of the International Migrant Stock from 1990 to 2015

From 1990 to 2005 in *figure 4_2,* the female migrants as a percentage of the international migrant stock in developing countries kept between 45% and 47%. After 2005 it decreased to a range of 43% to 45%, while it kept constant between 50% and 52% in developed countries.
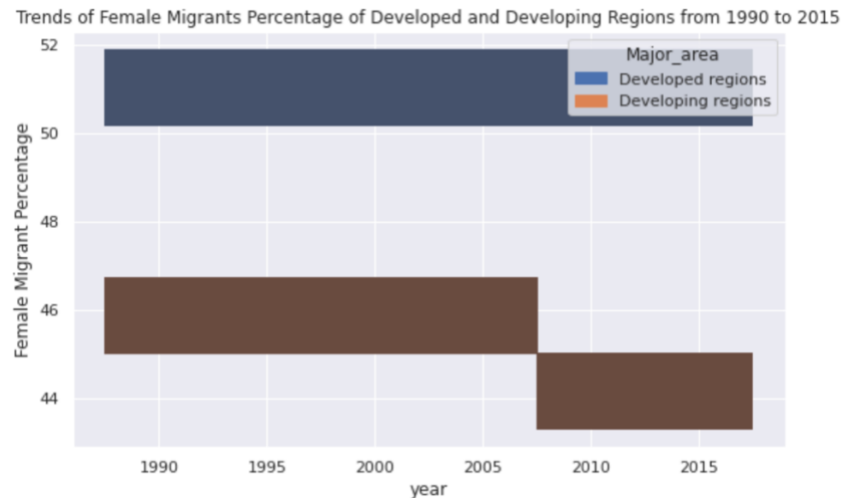
Figure4_ 1 Trends of Female Migrants Percentage of Developed and Developing Regions from 1990 to 2015

In big picture, the overall trends of female migrants as a percentage of the international migrant stock of Europe, Latin America and the Caribbean, and Oceania was increased. Also, from the *figure 4_3_1,* the female migrant percentage of Latin America and the Caribbean is almost same as that of Oceania from 2000 to 2015. The overall trends of female migrants as a percentage of the international migrant stock of Asia, Africa, and Sub-Saharan Africa were decreased.
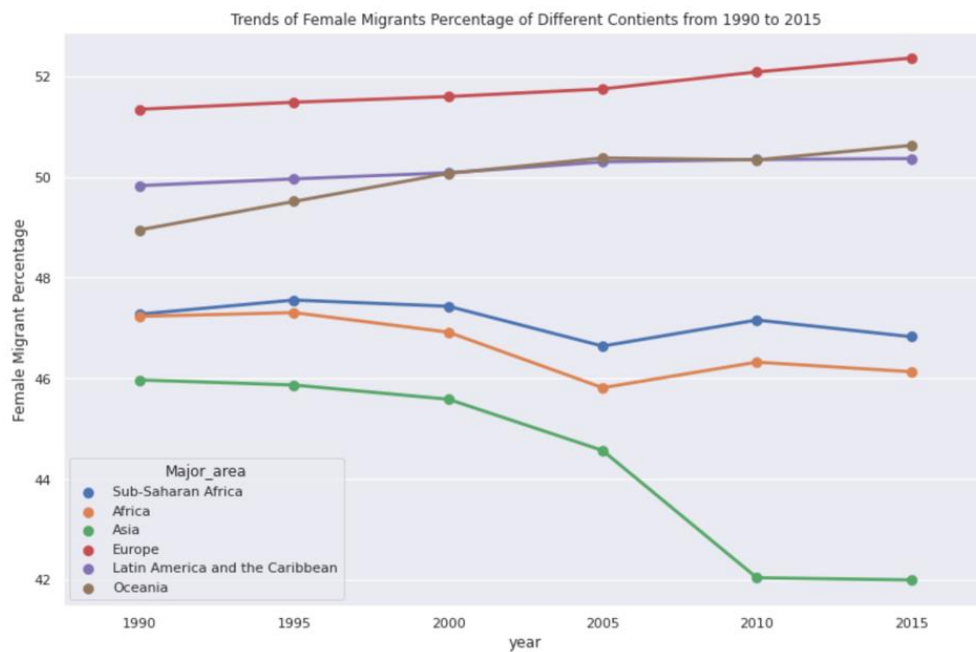

Figure 4_3_ 1 Trends of Female Migrants Percentage of Different Continents from 1990 to 2015

The index in Asia has fallen the most and shown no sign of increasing. As mentioned in the **Method section,** while creating the pointplot, it was unable to show error bars with all continents. This is because each data point is separated by labeling color, so there are no error bars because no data is being combined. Instead of removing "hue," the pointplot can be mapped onto the *seaborn.replot* by not specifying 'hue',
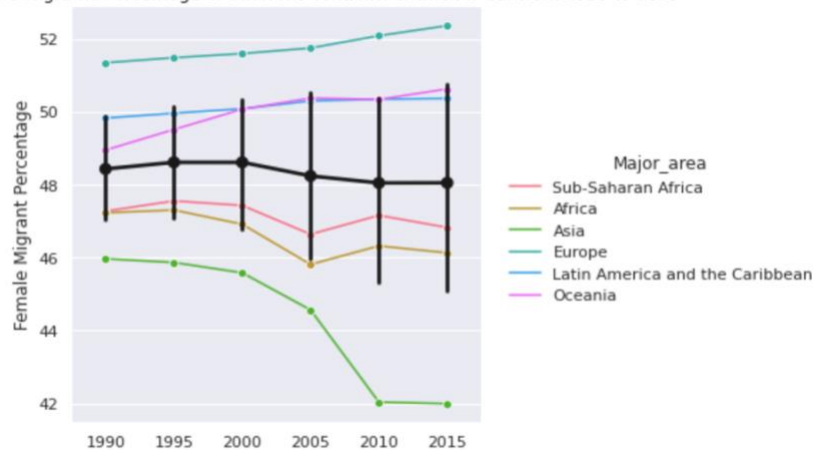

Figure 4_3_ 2 Trends of Female Migrants Percentage of Different Continents with Error Bar from 1990 to 2015

the API can create the error bars. From *figure 4_3_2*, the 95% CI bars are more generous and can overlap by as much as 60% and still indicate a significant difference. The length of the error bar is indicative of the degree of uncertainty. The error bar will be long if the number of collected data is small which reflecting the results based on a small number of cases are more uncertain and might not apply to the whole population (Confidence limits and confidence intervals (error bars) , n.d.) Among the six years, the year 2015 has the longest error bars, which expressed the data collected in 2015 reflecting more uncertain.

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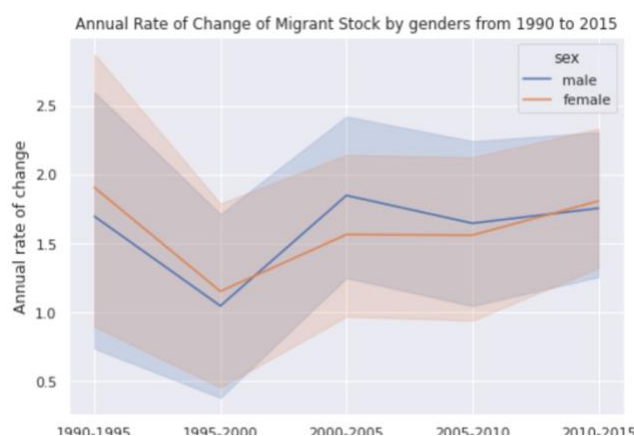

Figure 5_ 2 Annual Rate of Change of Migrant Stock by Genders from 1990 to 2015 (lineplot)
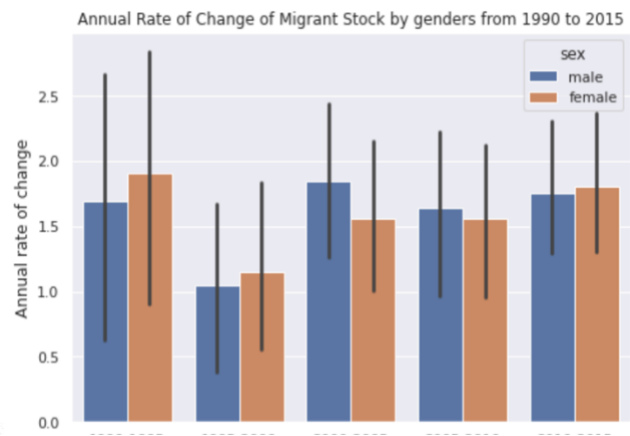

Figure 5_ 1 Annual Rate of Change of Migrant Stock by Genders from 1990 to 2015 (barplot)

Without regard to error bands and error bars, the overall trend of the annual rate of change of the migrant stock was decreasing, although it bounced in the middle time. The annual rate of change of the migrant stock of female was higher than male in the years from 1990 to 2000 and from 2010 to 2015. When considering errors, taking male for instance, the highest mean of annual rate of change of the migrant was in the period between 2000 and 2005 which is around 1.85% and the lowest mean of that was in the period between 1995 to 2000 which is around 1.05% which can be observed from both diagrams. *Figure 5_1*, lineplot was more likely to indicate the rising and decreasing tendency of annual rate change and the degree of reliability of data collected in each period. The data collected from 2010 to 2015 is more reliable than the data collected from 1990 to 1995. *Figure5_2*, the barplot shows the mean value of the data lies in the middle of the dataset clearer.

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

From *figure6_1*, the refugees as a percentage of the international migrant stock show a trend of downward first, then upward. The mean value in 1990 is highest at around 22.5%, and the mean
 value in 2005 is the lowest at around 17%. The error band is largest in the years 2010 and 2015, which means the data recorded in these two years lacks reliability. However, the reliability of the whole data of refugees as a percentage of the international migrant stock is not high.
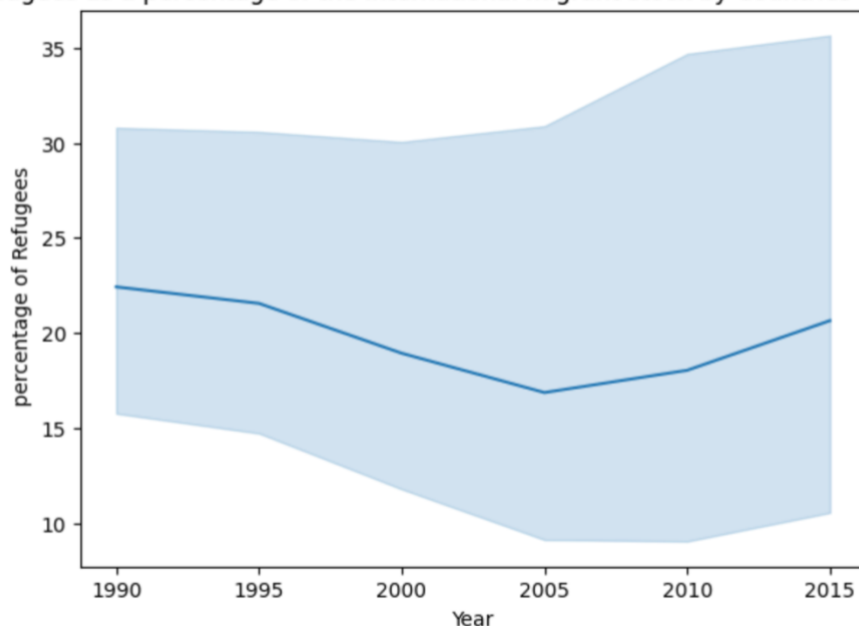

Figure 6_ 1 Trend of Refugees as a Percentage of the International Migrant Stock from 1990 to 2015

In *figure 6_2*, the median of 1990-1995, 1995-2000, and 2010-2015 is almost 0%. The median of 1995-2000 and 2000-2005 is lower than 0%. Thus, the trend of the annual rate of change of the refugee stock is similar to the refugees as a percentage of the international migrant stock which decreased and then increased. The figure below also indicated that the majority of data (25 percentile and 75 percentile) of each period was in the range of -10% to 10%.

The boxen plot provided more information. For example, the figure can tell the amount of data lies in different quantiles. Taking the period of 1995-2000 as an instance, 12.5% of data (first quantile) lie between -48% and -25% annual rate of change, and 12.5% (second quantile) lie between -25% and -10%. The figure6_2 can also show that each tail's outliers are equal, which is 7.
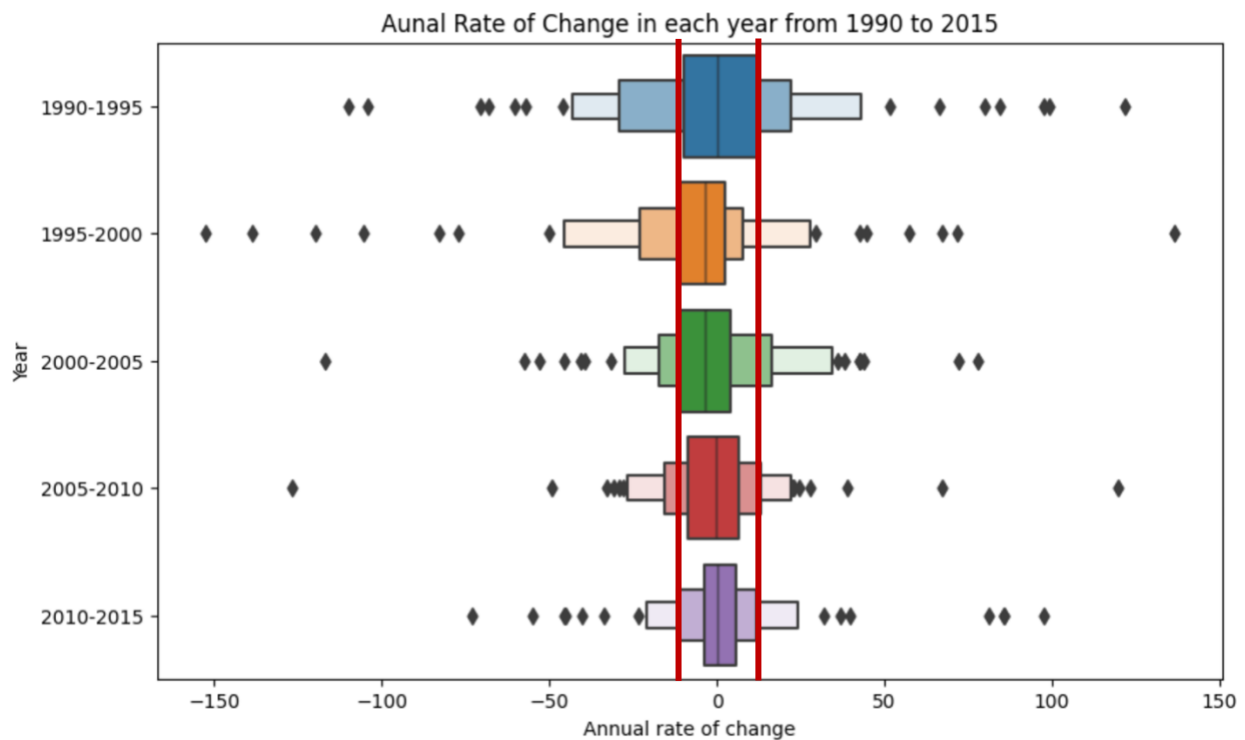

Figure 6_ 2 Annual Rate of Change of Refugees Stock from 1990 to 2015

5   Conclusion

This project made me realize that if characteristics of each table can be better identified during the midterm project, it would allow the visualization to be more efficient and successful; on the other hand, the data visualization can also check the data cleaning. Besides, data visualization helps explore data structure, detect outliers and unusual groups, identify trends and clusters, spot local patterns, evaluate modeling output, and present results. It is vital for exploratory data analysis and mining to verify data quality and familiarise analysts with the structure and characteristics of the data at hand. (Unwin, 2020).

Charts can disclose features of data that statistics and models may overlook, such as unusual data distribution, patterns, clustering, gaps, and missing values. Charts can spark research and ideas by raising questions. Interpreting diagrams requires knowledge and experience. In this project, charts can be plotted by software and by code successfully, but due to the lack of experience, most graphs were not interpreted appropriately, or even explained wrongly. In addition, this project has much weakness. Due to the categorical variable didn't convert to numeric variables; unfortunately, not all types of plots can be created.

6   Reference

Confidence limits and confidence intervals (error bars) . (n.d.). Retrieved from Radiation Effects Research Foundation: https://www.rerf.or.jp/en/glossary/confiden-en/#:~:text=They%20are%20usually%20displayed%20as,points%20plotted%20in%20a%20graph.

Jotform. (2021, Feb 11). Retrieved from How to do data visualization in 7 steps: https://www.jotform.com/blog/how-to-do-data-visualization/

K.EMERY, A. (2014, May 1st). *The Data Visualization Design Process: A Step-by-Step Guide for Beginners*. Retrieved from depict data studio: https://depictdatastudio.com/data-visualization-design-process-step-by-step-guide-for-beginners/

Kate Brush, E. N. (2022, Dec). *TechTarget*. Retrieved from data visualization: https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

Mcleod, S. (2019, July 19). *What does a box plot tell you?* Retrieved from SimplyPsychology: https://www.simplypsychology.org/boxplots.html

Tufte, E. R. (1983, August 28). *The visual Display of Quantitative Inforation.* Cheshire: Graphics Press. Retrieved from Tufte's Principles: https://thedoublethink.com/tuftes-principles-for-visualizing-quantitative-information/

Unwin, A. (2020, Jan 31). *Why is Data Visualization Important? What is Important in Data Visualization?* Retrieved from HDSR : https://hdsr.mitpress.mit.edu/pub/zok97i7p/release/4

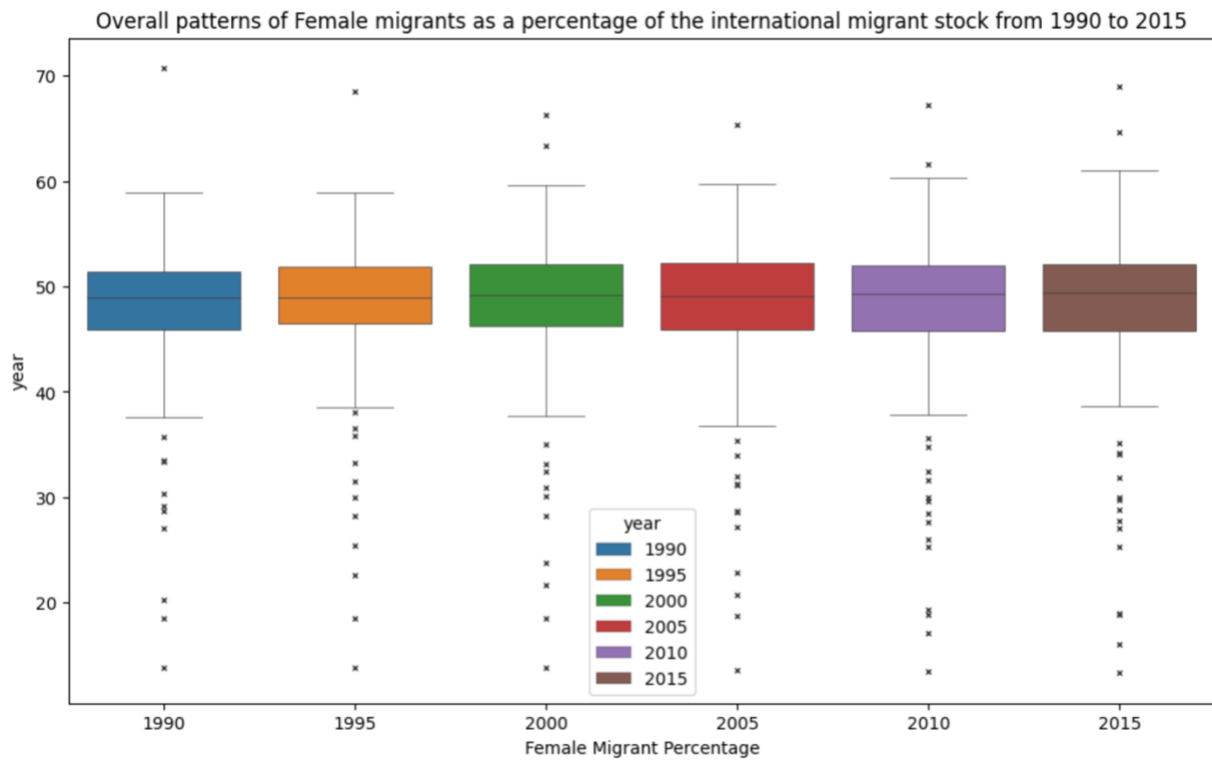Wickham, H. (2014). Tidy Data. *The Journal of Statistical Software*.

7 Appendix

Figure4_1_ 2 The Overall patterns of Female Migrants as a Percentage of the International Migrant Stock from 1990 to 2015 by Seabon
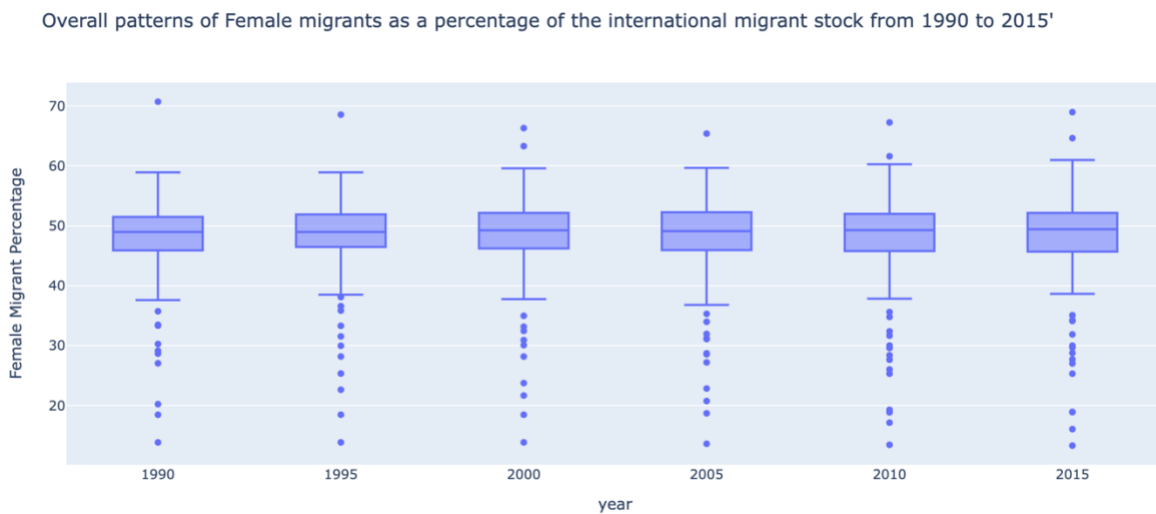

Figure4_1_ 3 The Overall patterns of Female Migrants as a Percentage of the International Migrant Stock from 1990 to 2015 by Plotly.express