INF 1340 Final Project: Data Visualization

Huan He
Professor Shion Guha
15th December 2022

INTRODUCTION

From our textbook and the experience shared by Shion, we have known that human-centered data science is widely used as a communication tool that bridges differences in values. However, data science itself can be boring and abstract to people who do not have an understanding of what data science is and what it can help, but data visualization can help with presenting complex information and ideas with clarity, precision and efficiency. Therefore, in this final project, I was thinking about the relationship between different variables and trying to tell a whole story using my graphs. Meanwhile, some of Tufte's principles have been applied to help me make my visualization decisions.

METHODS AND RESULTS

PREPARATION

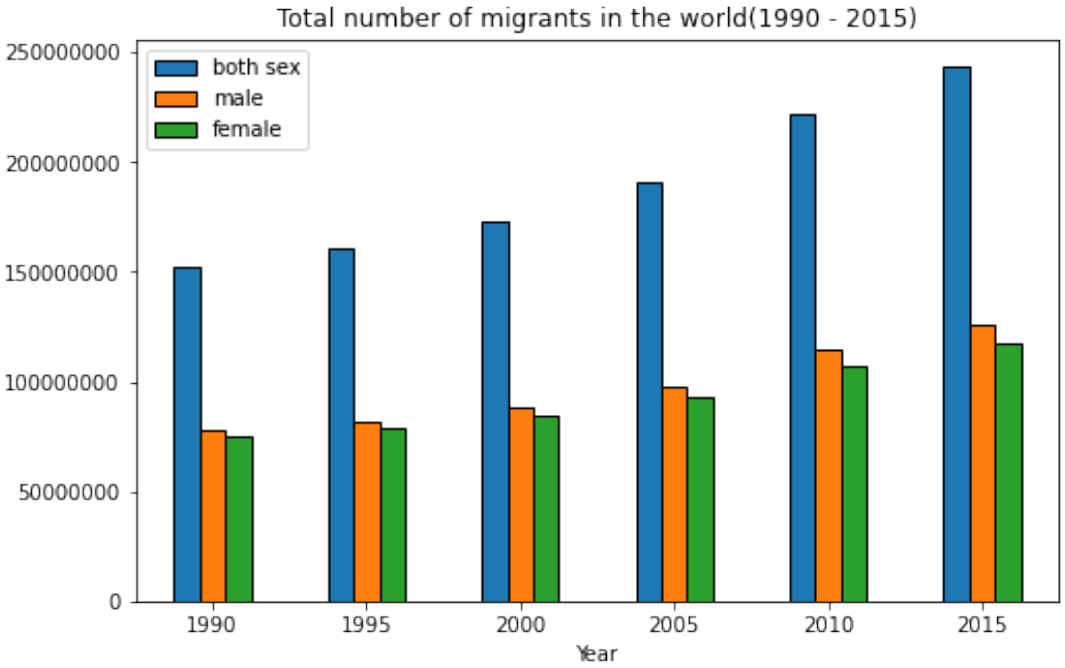
The data that was used for visualization for this project was from my prior data cleaning project about the UN Migration Stock dataset. Rather than merging a couple of my datasets to tell a better story, I did not make any further changes to my previous dataset after cleaning because it already met 80% of the tidy data principles. Also, my visualization project was conducted on Google Colaboratory based on my data cleaning project.

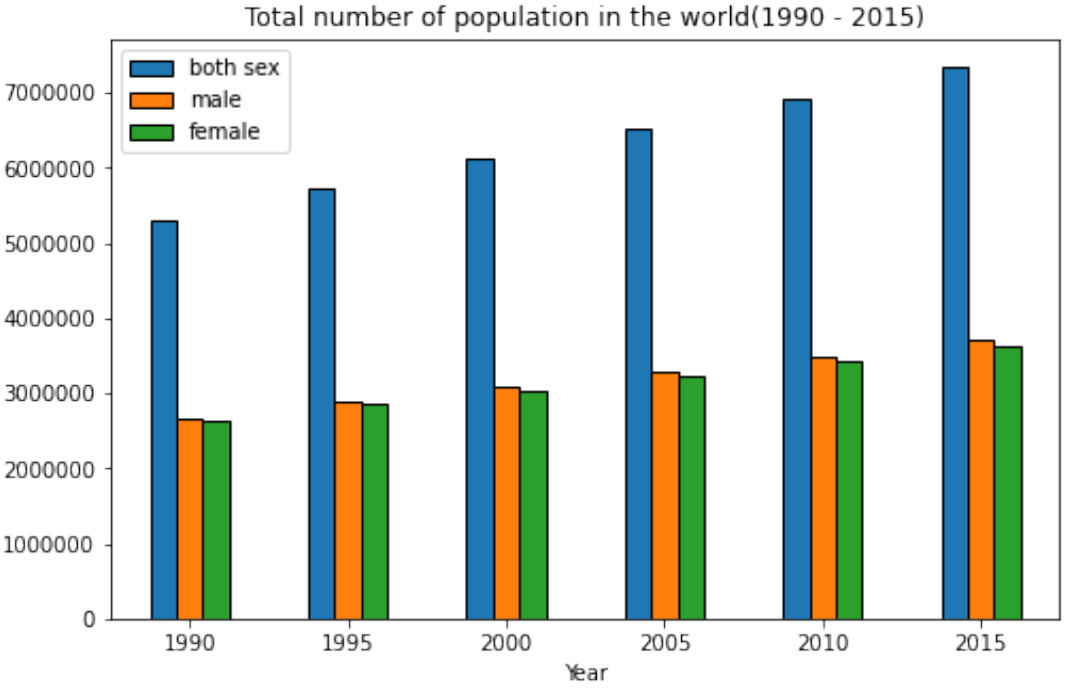
Due to the reason that the UN Migration Stock dataset is a large dataset with much complex information even after the cleaning, in this project, I decided to narrow down the destinations to some specific regions and countries. In this project, I selected two groups of regions and countries to conduct the analysis. The first group is composed of all the continents and the world which are *World, Africa, Asia, Europe, Latin America and the Caribbean, Northern America and Oceania*. The second group is composed of main countries I selected from each continent which are *Canada, Congo, Cook Islands, Cuba, India, Indonesia, Japan, Pakistan, Switzerland, Ukraine, the United Kingdom of Great Britain and Northern Ireland and the United States of America.*

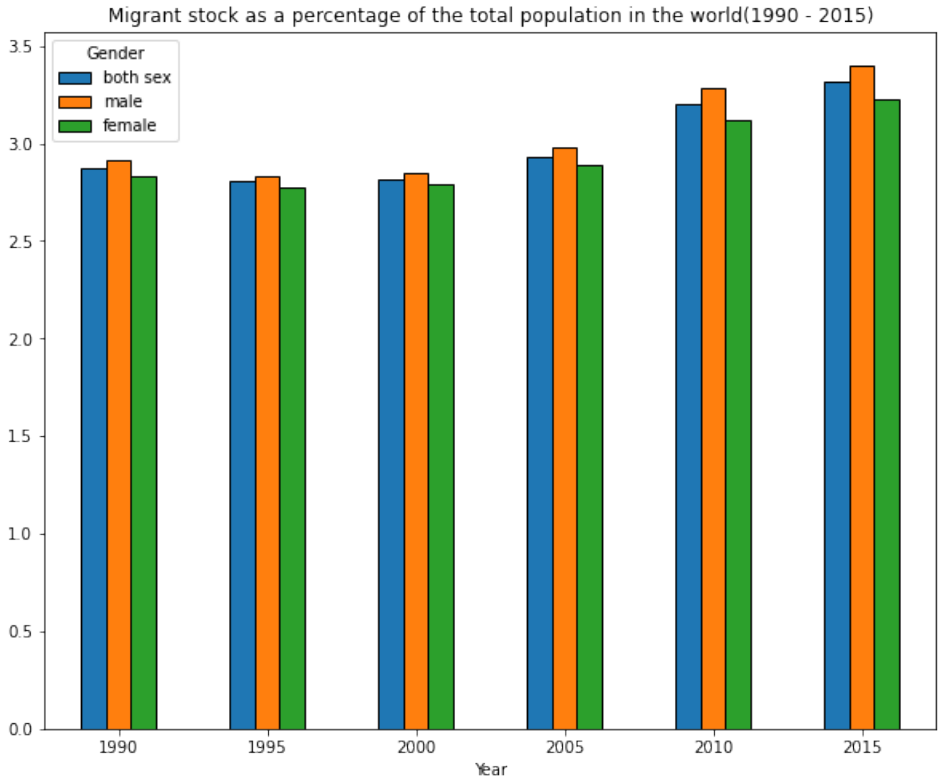
METHODS AND RESULTS

At the very beginning, I wanted to look at how the population and migrants changed and how was the general tendency of migrant stock as a percentage of the population like in the past years in the world. Therefore, I came up with question 1 to see what the tendency looks like. Refers to *Tufte's principle #2: Use clear, detailed, thorough labelling, and principle #3: Show data variation, not design variation*, to make sure my graphs are with less ambiguity and with more clarity to show the tendency, I decided to use **histograms** to visualize the data for this question. Finally, it came up with three histograms.

Question 1: What are the trends of population, migrants and migrant stock as a percentage of the population in the world?


Total number of migrants in the world(1990 - 2015)
both sex
male
female
250000000
200000000
150000000
100000000
50000000
0
1990
1995
2000
2005
2010
2015
Year


Total number of population in the world(1990 - 2015)
both sex
male
female
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
1990
1995
2000
2005
2010
2015
Year

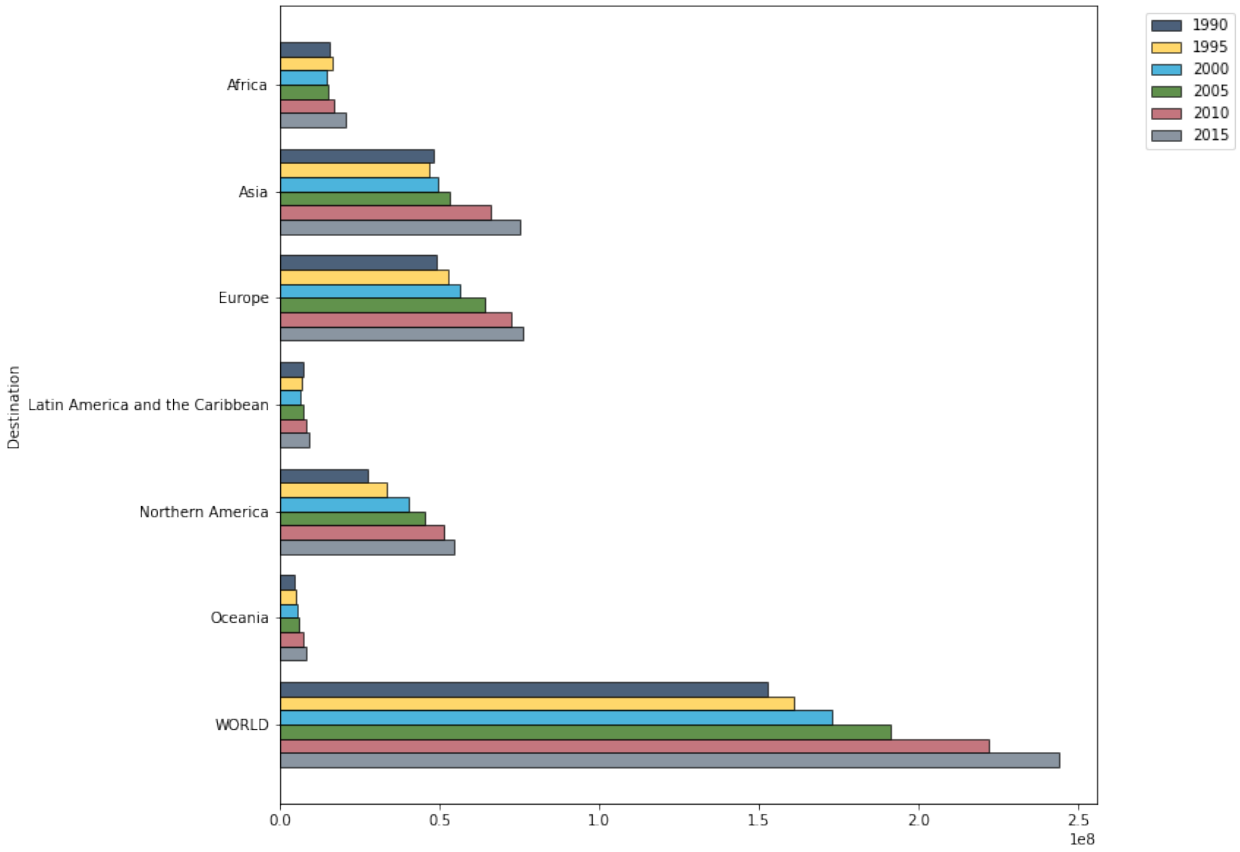


By using histograms, I am able to see the big picture of how the tendency looks like and compare the data tendency of male and female migrants at the same time.

Then, after looking at the world in the big picture, I decided to zoom in a little bit to see what happened on the continents, and came up with the second question

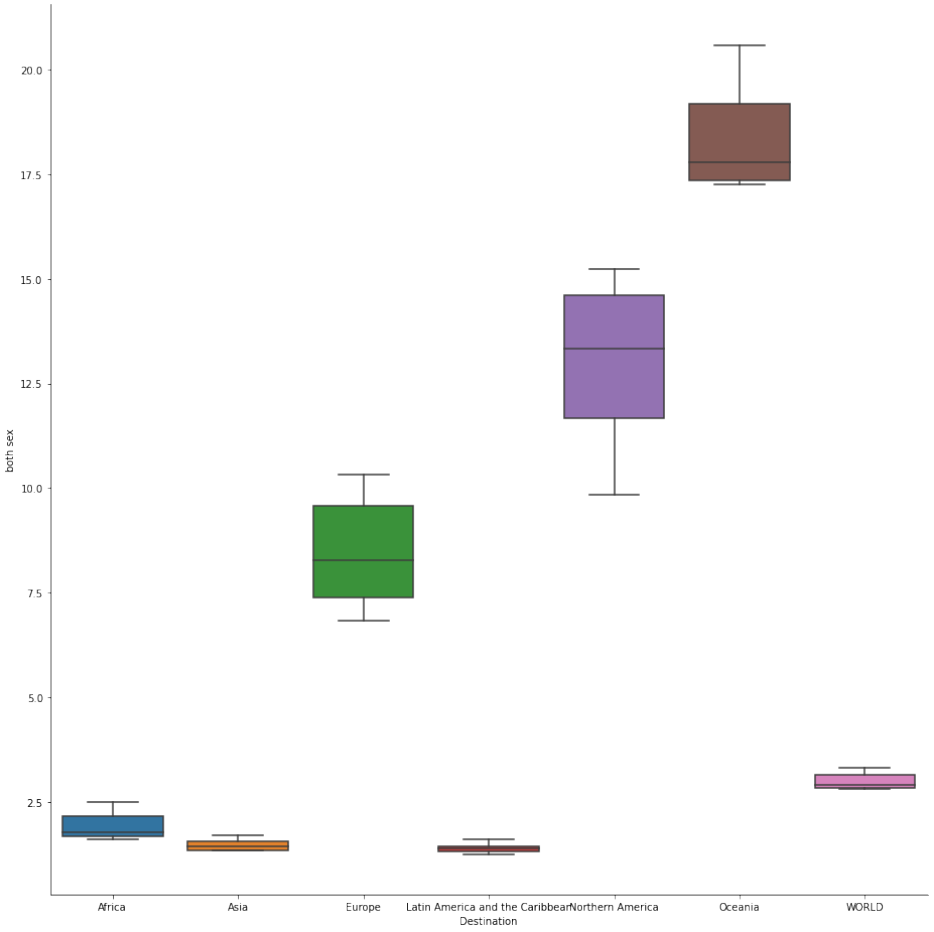
Question 2: What is the trend of migrant stock growing in continents?

In order to compare each continent in a more straightforward way, refer to *Tufte's principle #1: Make the representation of numbers proportional to quantities.* I decided to use **barh** to show the difference between continents.



After discussing the general trend of migrant stock in continents, I wanted to see how the percentage of migrant stock to the total population was distributed in these continents. Did the number have a significant fluctuation between these continents? That's how question 3 proposed. In order to have a better look at the distribution, refer to *Tufte's principle # 6: Quote data in full context*, I expected the difference of distribution in these continents would not be small, I chose to show the **boxplot** to demonstrate the data.

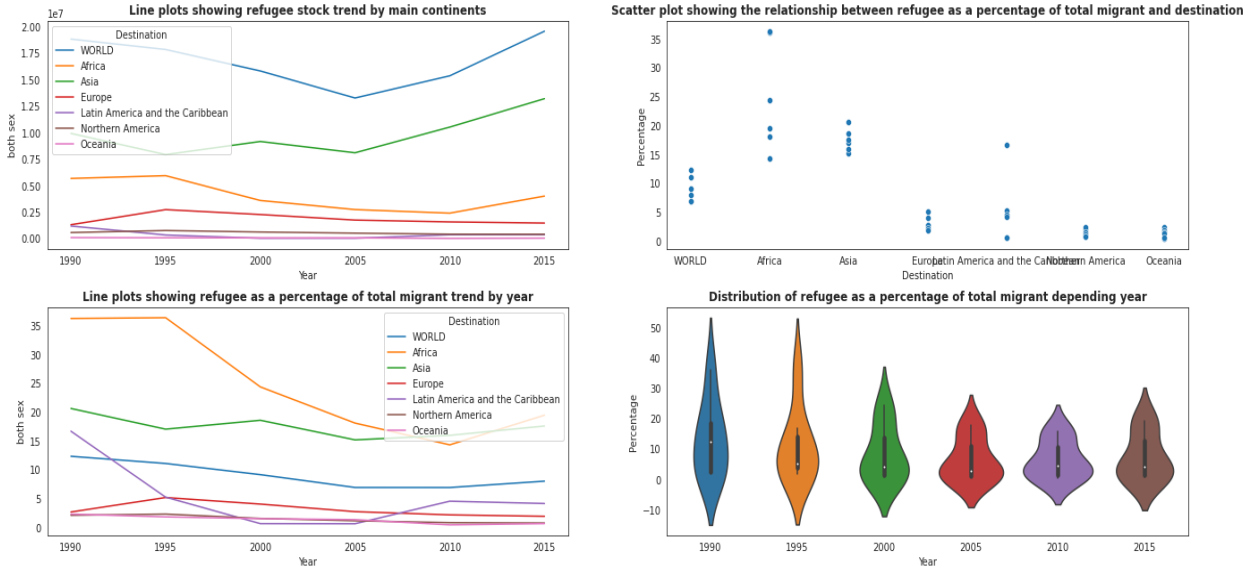
Question 3: What is the distribution of migrant stock as a percentage of the total population by continent?



Then, besides the changes in migrant stocks, I want to see the how the refugee stocks changed in these continents. Therefore, I made my 4th question to be

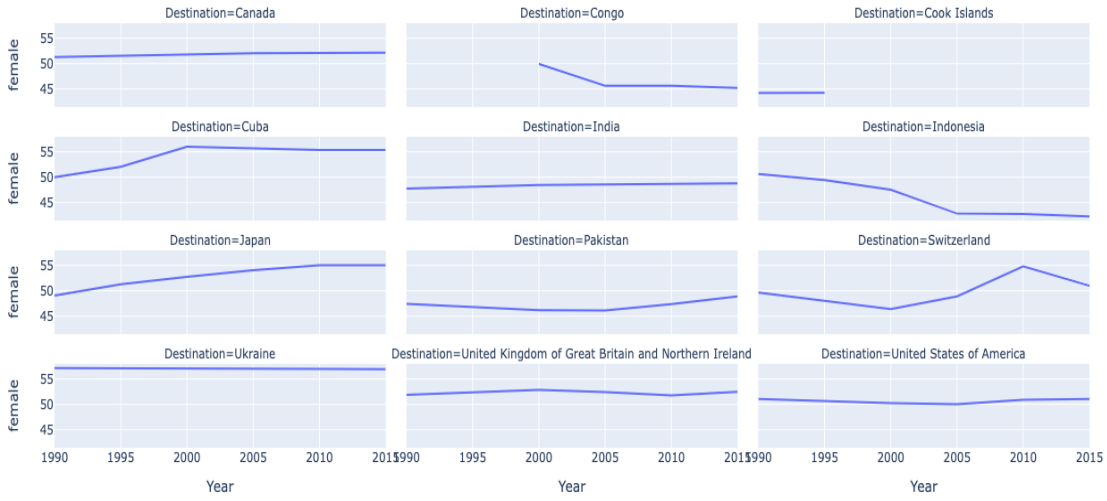
Question 4: What is the trend of refugee stock by year and refugee stock as a percentage of total migrant in counties, and what is the relationship between refugee stock as a percentage of total migrant and destination, and what is the distribution of refugee stock as a percentage of total migrant in these countries?

I hoped to use a simple way to see all the data about the refugee, so I chose to put all graphs in one canvas. Refer to *principle #3: Show data variation, not design variation.* I used line pots to demonstrate the tendency of refugee stock as a percentage of the total migrant go by years. I expected that the relationship between refugees as a percentage of total migrant and destinations would not be a linear relation, so I chose the scatter pots to show how the data is centred and discrete. Also, I used a violin pot to see how the distribution of refugees as a percentage of total migrants goes by year.



Finally, I was curious about what was the percentage of female migrants as of total migrants in countries, so I chose 12 different countries from different continents and used the **line pots** to see and compare their general tendency.

Question 5: What is the trend of the female migrant as a percentage of the international migrant in some main counties?



DISCUSSIONS

There are some interesting findings from all the graphs. Question #1, it is showing that the population and the migrant stock were increasing in the world. Meanwhile, the migrant stock as a percentage of the total population is increasing too, but the growth rate is lower than the overall increase of population and migrant stock. Also, the total male population and male migrant stock is always higher than females. Question #2, we can see that the overall migrant stocks are increasing in each continent, and Asia, Europe and North America are the three continents that witnessed the most significant increase of migrants. Question #3, 23 can see that

the distribution of migrant stock is stable in the world from year to year, and Oceania, North America and Europe are three continents that had the most significant changes of migrant stock from 1990 to 2015. Then, question #4, it is showing that there is not a significant of increase of refugee number in the world. In some continents, for example Africa and Latin America, the percentage of refugee as the total migrant population was decreasing. Also, it is showing in the graph that the percentage of refugee of total migrant tended to be stable after 2000, and Africa witnessed the biggest discrete. Last but not least, question #5, a fun fact addressed is that the tendency of the percentage of female migrant was stable in Canada and India, but in Canada, the female migrant was consistently over 50% while in India, the female migrant was always lower than 50%.

CONCLUSION

In conclusion, the objective of this project is to do the visualization data analysis based on the data that was cleaned. During this process, it is essential to come up with the right question and think about the proper graphs to demonstrate the data. Visualization helps the reader to get a close look at data science and make the meaning of data clearer, In the future, I hope I can understand more about the logic in the data, and think about the questions before I master more code about visualization.