INF1340 Final Project

Introduction

The dataset studied in this project consists of six tables about trends in international migrant stock. The background of this dataset is to categorize the number of international migrants from 1990 to 2015 based on origin, gender, and refugee stock. As some variables that are irrelevant to the dataset can negatively affect the results, I have cleaned up the six tables in the previous project according to the five principles of tidy data (Wickham 2014). In this way, all the tables have been organized to become more logical and observable. More importantly, it makes creating charts and graphs much more comfortable, making it easier to analyze and understand visually. On this occasion, I will use Tufte's visualization principles to perform Exploratory Data Analysis (EDA) on the cleaned-up tables. In statistics, EDA is a method that uses graphical techniques and specific data visualization. It is the process of extracting essential data and variables for analysis and generalization. This approach can show the trends and characteristics of the data itself in a very straightforward way. It also helps in the analysis and investigation of the project. Moreover, EDA can be used to check whether the variables between data sets are correlated, providing applicability and accuracy for the statistical techniques being used.

Method

First of all, before conducting Exploratory Data Analysis, the table was rechecked, and found several places where there were 0 or N/A data in the dataset. This is not ideal for graphing or displaying trends. For this reason, I removed all data from the table that had nonsense. Then I used Tufte's visualization principles to create the graphs and analysis.

Principle 1- Captions Are Not Optional.

For a graph, a caption is an excellent way to explain what the graph represents. The audience needs to get effective information when they are viewing, so the captions need to be visually obvious. This will help the audience understand the meaning of the paragraph's content.

Principle 2 - Not Trust the Defaults.

The defaults of a graph do not apply to all kinds of datasets. This is because all datasets are different in their characteristics and trends. If this is ignored, the final graph might show up in a way that does not make sense or even be hard to interpret. Therefore, in order to be able to ensure that the information is well communicated to the general audience, any graph needs to be manually adjusted on this setting. For instance, some dimensional parameters and interval sizes. This will give a smooth visual experience.

Principle 3 - Use Color Effectively.

Sometimes, charts with color can look very stunning and eye-catching. However, if color is not used correctly, it can have a devastating effect on the chart. For example, colors that are so bright that audiences can't navigate them or too many colors in a chart can make the chart look very messy and affect the understanding of its meaning. Thus, when choosing the color, try to choose a clear and sensible match based on the content of the data. It is not true that the more colors you have, the better you can convey the message.

Result

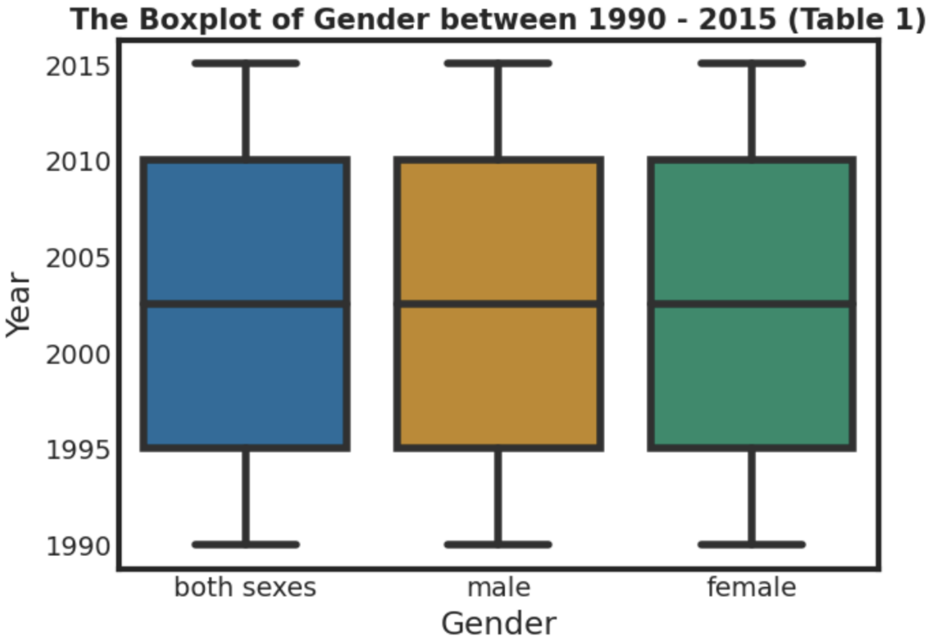

Figure 1: The Boxplot of Gender Between 1990 - 2015 From Table 1

Figure 1 is about the statistics of the number of international migrant stock by gender between 1990 and 2015. The information is obtained from Table 1. From this boxplot, we can see that three colors categorize the gender. From the data, the median of all genders is between 2000 and 2005. In the number of international migrant stock, the quantity of these three genders is very equally distributed. There is no outlier generated in them. It means that the gender of international migrant stock is balanced between 1990 and 2015.

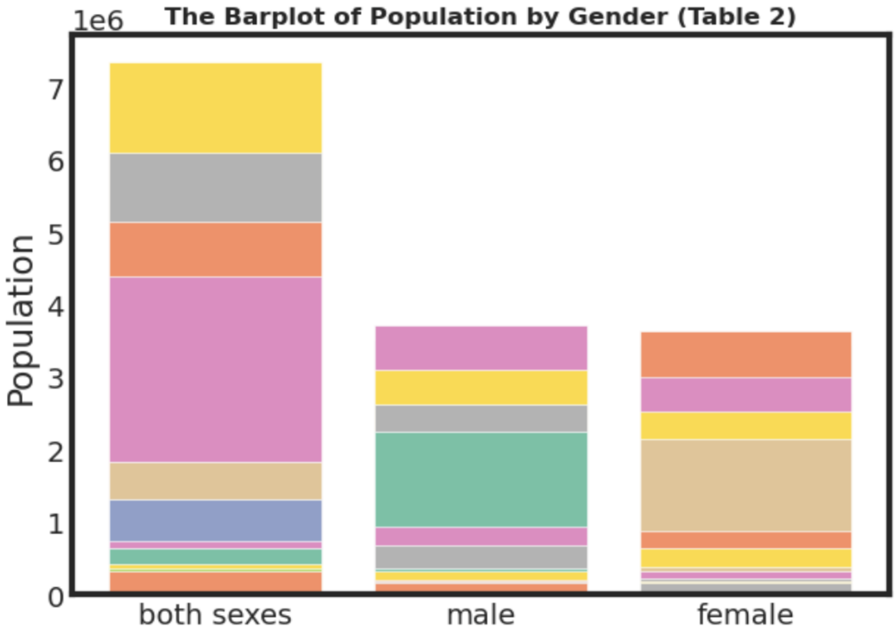

Figure 2: The Barplot of Population by Gender Between 1990 - 2015 From Table 2

The Figure 2 shows the statistics of the total population (thousands) based on gender between 1990 and 2015. The information is obtained from Table 2. From this barplot, it can be seen that the total male population is higher than the total female population, about 3.6×10^{6}, compared to about 3.5×10^{6} for females. It is also visible from Figure 2 that each gender is overlaid with a different color bar plot. These can reflect both the overall population and the individual population. Therefore, it is clear to compare which categories have the highest number of people in the same period.

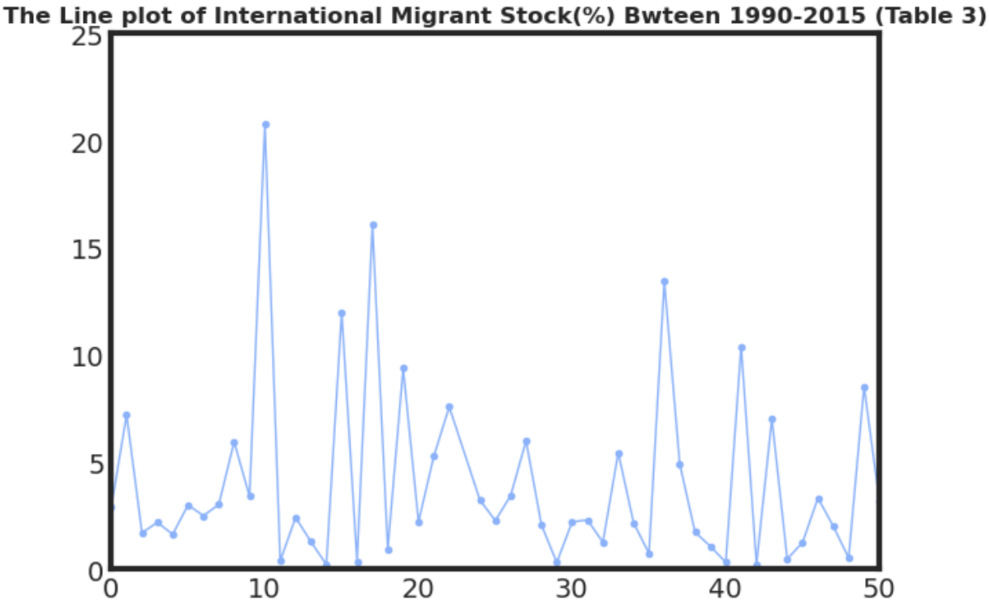

Figure 3: The Lineplot of International Migrant Stock (%) Between 1990 - 2015 From Table 3

The Figure 3 is about the percentage of International migrant stock trend between 1990 and 2015, according to all genders and total population. The linear plot illustrates that the movement of International migrant stock has been up and down. It goes through several troughs before coming to a peak. It shows that the percentage of International migrant stock is not always stable between 1990 and 2015.

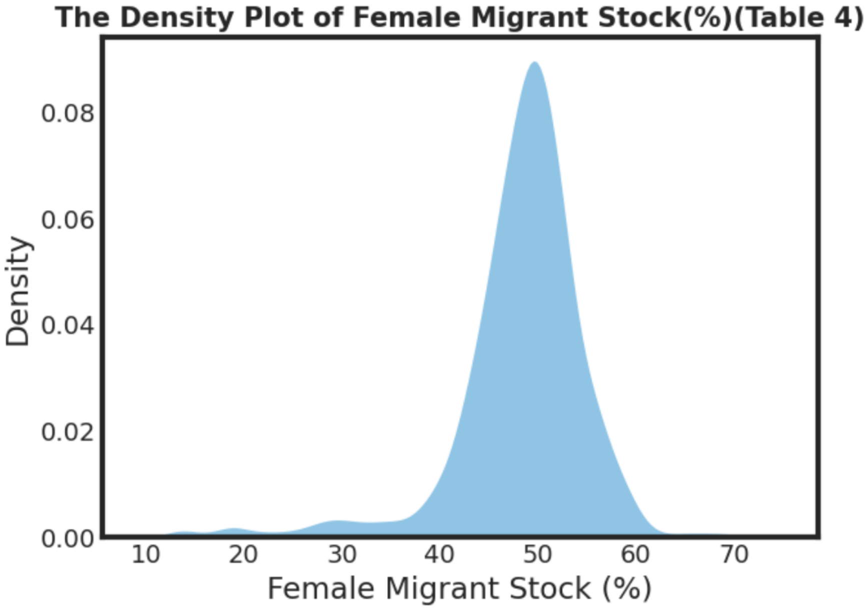

Figure 4: The Density Plot of Female Migrant Stock (%) Between 1990 - 2015 From Table 4

The Figure 4 shows the distribution of the percentage of female migrant stock between 1990 and 2015. From the density plot, the peak value of the rate of female migrant stock is around 50%. This indicates that between 1990 and 2015, 50% was the most highly concentrated position.

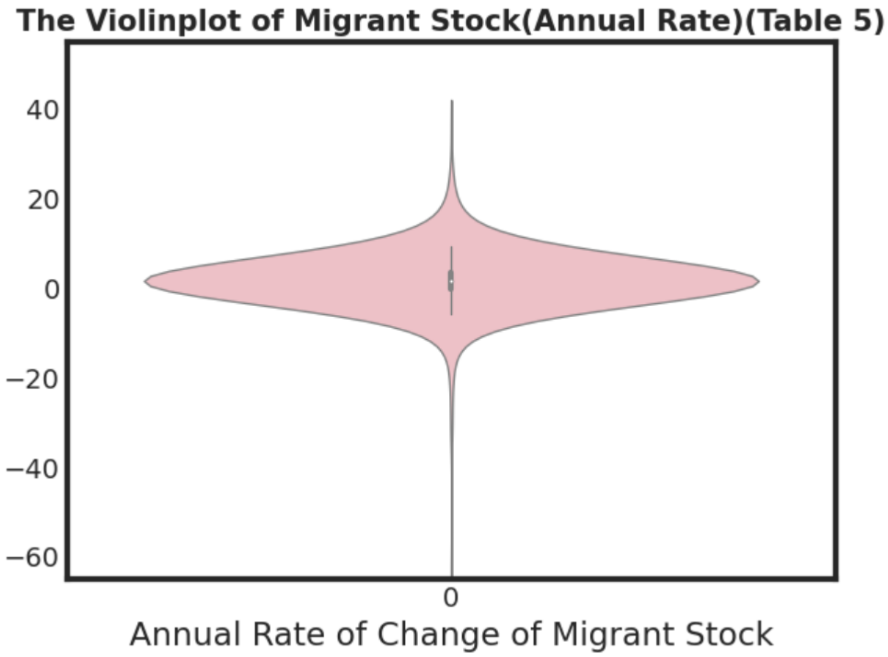

Figure 5: The Violinplot of Migrant Stock (Annual Rate) Between 1990 - 2015 From Table 5.

The Figure 5 shows the annual rate of change of the migrant stock based on gender between 1990 and 2015. This violin plot shows the median at approximately 0%. Then the upper adjacent value is about 22% and the lower adjacent value is about -22%. Furthermore, this violin plot also demonstrates the presence of some outlier points, about 25%.

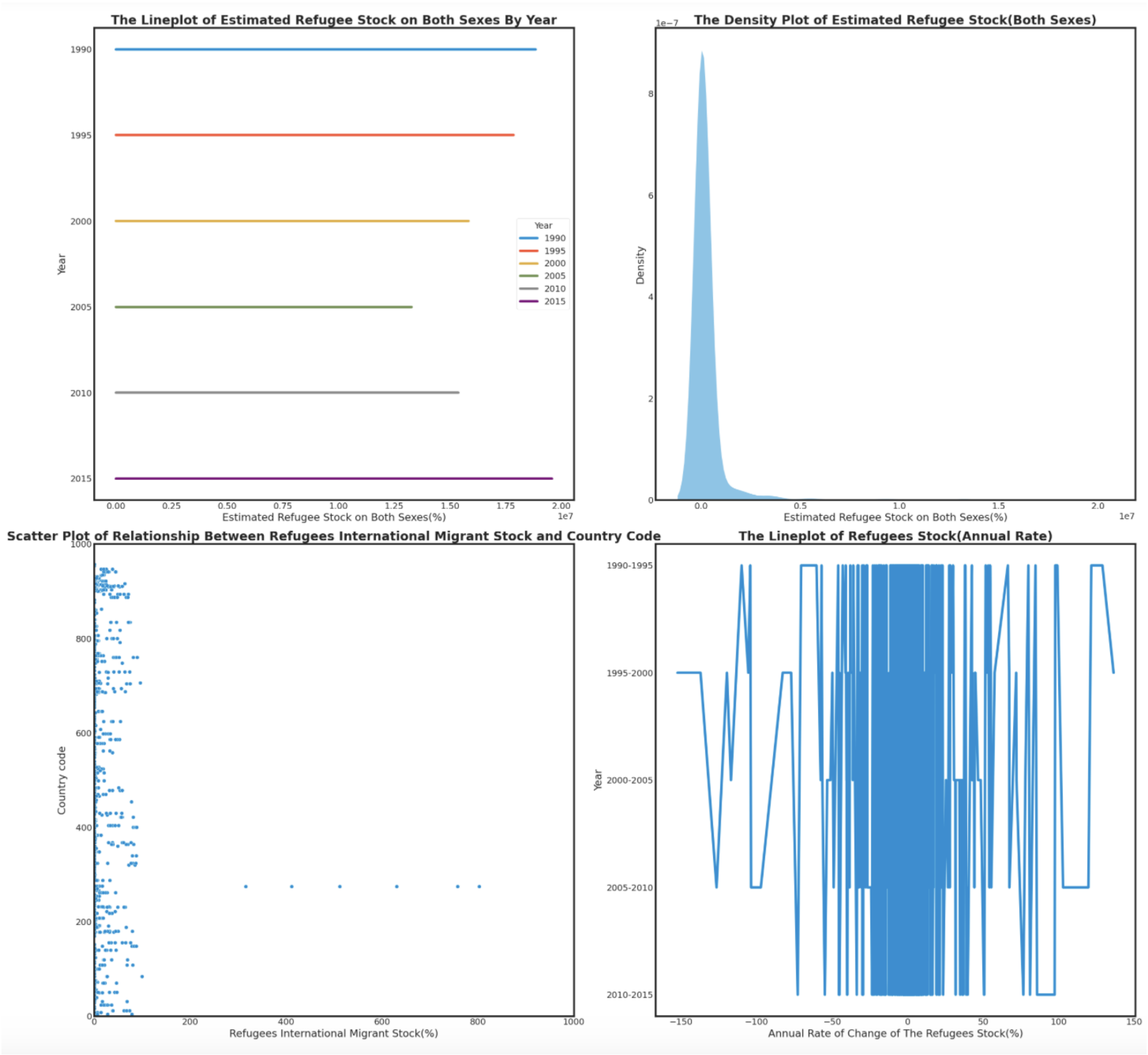

Figure 6: The Four Plot From Table 6.

The Lineplot of Estimated Refugee Stock on Both Sexes By Year.

The Density Plot of Estimated Refugee Stock(Both Sexes).

The Scatter Plot of Relationship Between Refugees International Migrant Stock and Country Code.

The Lineplot of Refugees Stock(Annual Rate).

The figure 6 contains four different distribution plots. The information was obtained from Table 6. The first line plot shows that between 1990 and 2015, the highest estimated refugee stock for both sexes was 2015, about 2×10^{7}, and the lowest was in 2005, about 1.3×10^{7}. The second graph is about the density of estimated refugee stock for both sexes. It can be seen that the density is most concentrated at about 0.1×10^{7}. The third graph is about the relationship between refugees' international migrant stock and country code. From the chart, it can be seen that each country code corresponds to the refugees' international migrant stock within 100%. In the middle of 200 and 400 for country code, there are some outliers. The fourth graph is about the annual rate of refugee stocks. From the graph, it can be seen that the most dense area is distributed between -50% and 50% between 1990 and 2015.

Discussion

According to my results, I became more familiar with the principles of tidy data and the meaning of Tufte's visualization principles. While plotting the different charts, I was able to filter and group the most prominent and meaningful data from the tables. The representative graphs were chosen to present the data logically and visually. Most importantly, I know how to create a clear and complete Exploratory Data Analysis. The visualization process's more challenging part is giving the audience a visual direction to help them understand the data and information. Since there is no uniform standard for this process, it is essential to understand the most basic rules for making graphs based on a statistical perspective.

Conclusion

Overall, data visualization is designed to make it easier and more intuitive to help with analysis. I have drawn graphs using six tables based on Tufte's visualization principles and then briefly summarized all the charts. Consequently, after the drawing, the data had a better quality and information display. This enabled better communication and strategic decision-making during the analysis process.

Appendix：

- Remove all 0 and N/A date in the dataset for better plotting.