1. Introduction

Referring to the book *Statistic*, the authors state that statistics is descriptive because it helps us better understand the prevalence of variability in the real world by providing tools for transferring and reflecting actual phenomena (Witte & Witte, 2017). So, it is important that we can apply statistical tools to collect and analyze what we've observed in the real world.

The database for this project is *"Trends in International Migrant Stock: The 2015 Revision'*, and it was generated by the United Nations. In this project, I will be going to excavate the meaning behind the database.

2. Methods

Data analysis requires data analysts to analyze the data by using different types of data visualization. Meanwhile, different visualization methods contain unique usages. In this project, I've used Bar charts, Scatter graph, Line Graph, Box Plot, and Violin Plot as the data visualization methods. Based on the website, I will the usages of the methods that I've used in this project.

a) Bar chart

   i. A Bar Chart can show discrete and numerical comparisons in either horizontal or vertical forms. Each axis of the chart represents certain categories, and the other axis represents a specific discrete value.

b) Scatter Graph

   i. A Scatter Graph shows the relationships between two variables by showing up all the values between the variables, and each axis represents each variable.

c) Line Graph

   i. A Line Graph displays trends of the data over time by showing up quantitative values within a continuous time period.

d) Box Plot

   i. A box Plot displays how data distributes by showing up median and different quartiles. It is useful for analysts to compare distributions among different groups or databases.

e) Violin Plot

   i. Similarly, a Violin Plot can show the distribution of the data, additionally, it also can show the probability density of the data by displaying the distribution shape of the data. Usually, a Violin Plot carries more information than a Box Plot.

Meanwhile, I've also corporately used multiple Python plotting libraries which are Matplotlib and Seaborn for betting visualizing the observations and the results.

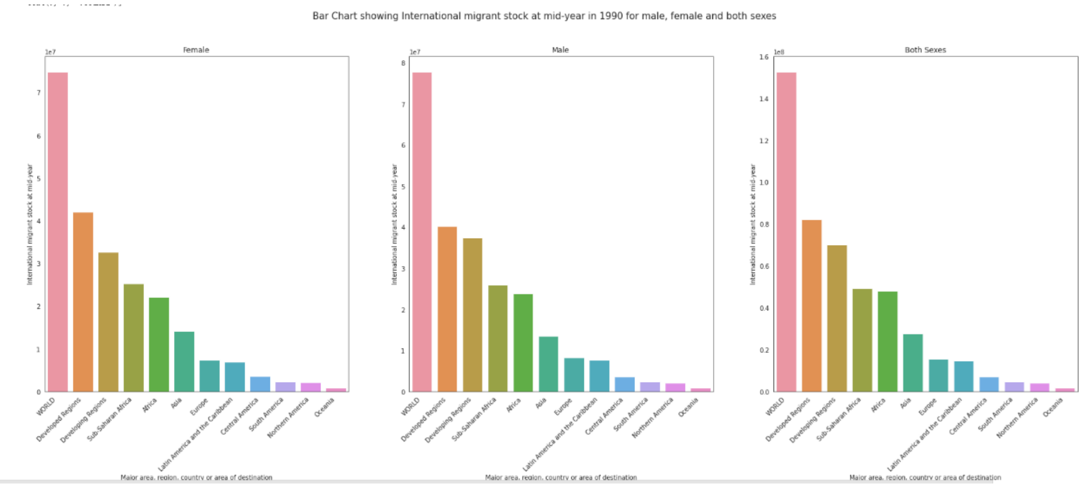
3. Results

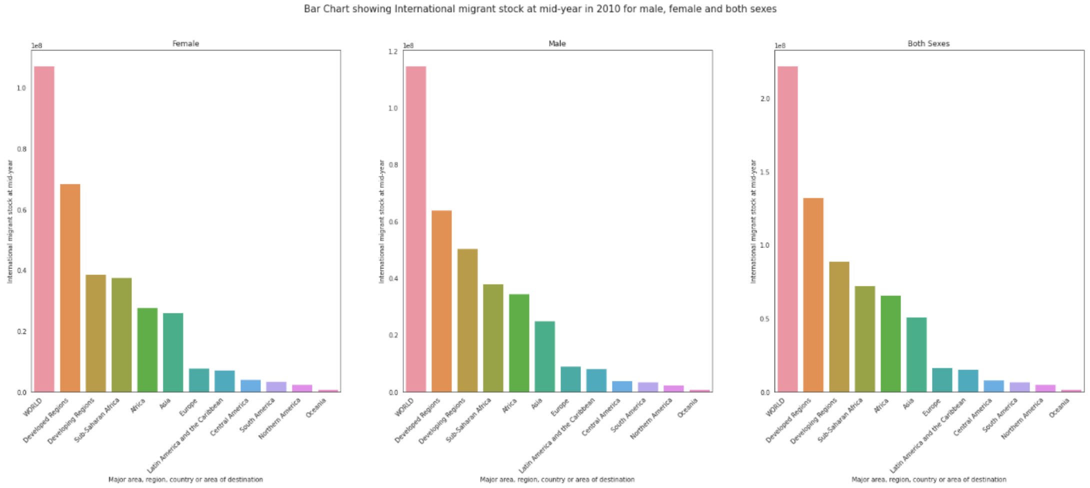
In this part, I will introduce my results based on the graphs and followed Tufte's Data Visualization Principles:

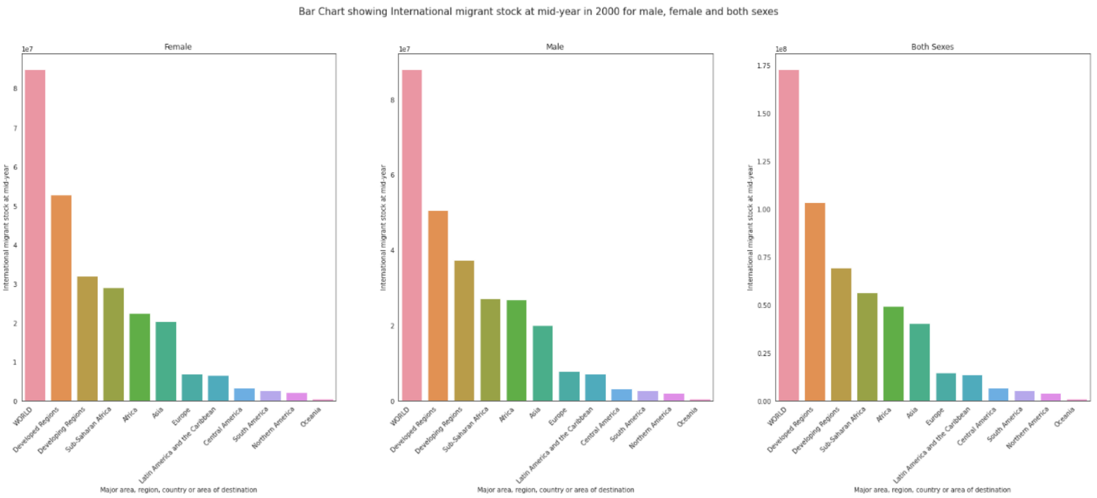
Comparison
Causality
Show multivariate data
Use appropriate encoding
Maximize data-ink ratio
Use appropriate scale
Label carefully

For the comparison concern, I've devoted myself to focusing on comparisons between different variables or groups. Those groups or variables might be connected or correlated with each other. In the use of a bar chart, I've titled it "**Bar Chart showing International migrant stock at mid-year in 1990 for male, female and both sexes**", and I've picked up Gender and Major area, region, country or area of destination as my comparison groups in terms of the value of "International Migrant Stock at mid-year" in the years of 1990, 2000, and 2010. Therefore, I've applied both showing multivariate data and labeling my charts carefully. Also, the charts display in different colors for improving readability. The

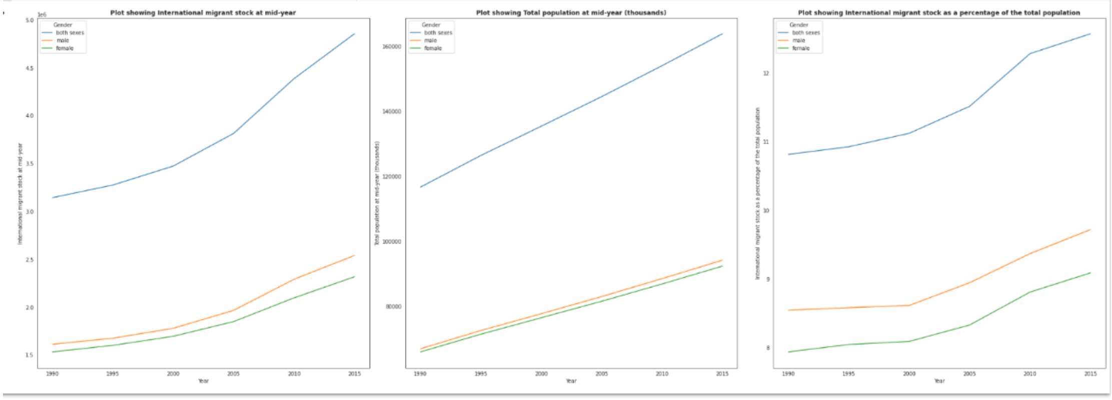
Y-axis and X-axis represent the variable of the International migrant stock and geographical locations, respectively. This chart shows the International migrant stock values for different major regions and areas in 1990, 2000, and 2010 for the reader to understand how the International migrant stock differed under different conditions. Based on the charts, I can see the trends that are similar in 1990, 2000, and 2010 for males, females, and both sexes in different geographical locations on the same scale. To be specific, they remain largely unchanged in the distribution in terms of different geographical locations. I can also compare specific values among different geographical categories when other categorical features remain the same. For example, the gap in the number of female migrants between developed regions and developing was increasing from 1990 to 2000.


Bar Chart showing International migrant stock at mid-year in 1990 for male, female and both sexes

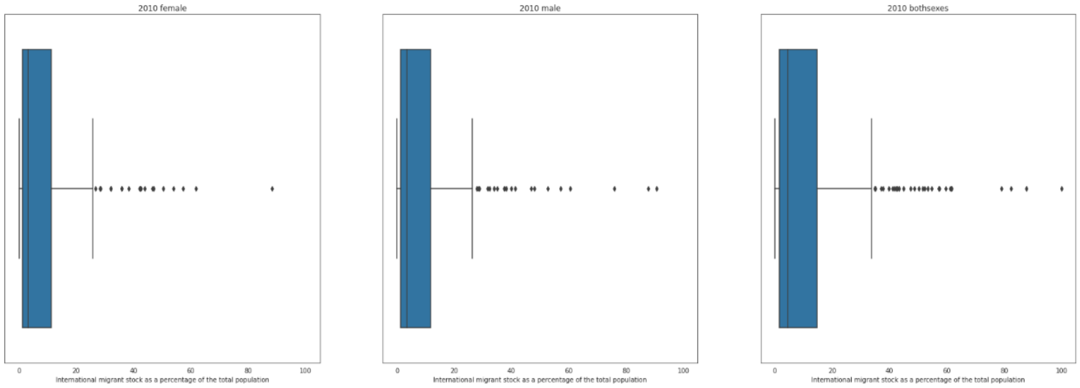

Bar Chart showing International migrant stock at mid-year in 2010 for male, female and both sexes



In the use of a line chart, I've titled it "**Plot showing International migrant stock at mid-year**", "**Plot showing Total population at mid-year (thousands)**", and "**Plot showing International migrant stock as a percentage of the total population**". I've mainly aimed to compare the trends of different variables in 1990, 1995, 2000, 2005, 2010, and 2015. The Y-axis for the first, second, and last charts represents the International migrant stock, Total population, and International migrant stock as a percentage of the total population, respectively. The X-axis for all charts is Year. As the plots are shown, I could see the three variables that have the same trend in a continuous time period from 1990 to 2015. Referring to the last chart, I can observe that the International migrant stock as a percentage of the total population had largely increased around 2010 which might be triggered by the financial crisis in 2008. I also use different colors to represent different genders for improving readability.



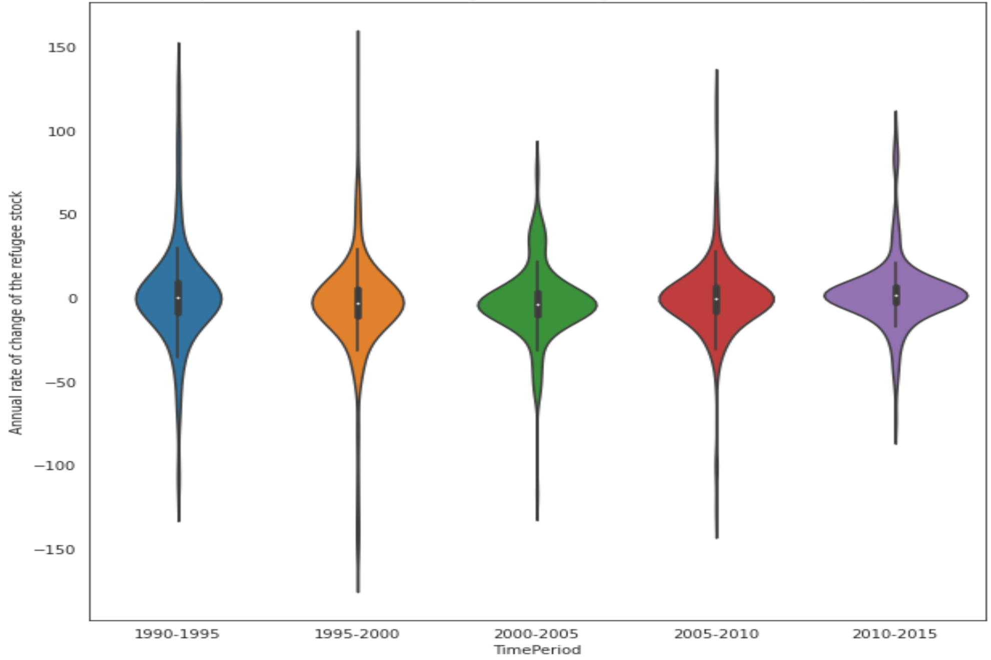
In the use of a box plot, which I've titled it as "**Box Plot showing International migrant stock as a percentage of the total population in 2010 by different genders**", I've devoted to comparing how the International migrant stock as a percentage of the total population differs in different years and genders. The X-axis for all charts is the International migrant stock as a percentage of the total population. Based on the plots, I can see that the International migrant stock as a percentage of the total population for males contains more outliers than females in 2010, within the same scale of 100. Meanwhile, I can see the medium and mean of the variable for male in 2010 is similar for female.



In the use of a violin plot, it is similar to a box plot, and it also provides me with a summary statistic. I've labeled it as "**Violin plot for Annual rate of change of the refugee stock in different time periods**". I've compared the performance of the annual rate of change of the refugee stock in the different time periods which are 1990-1995, 1995-2000, 2000-2005, 2005-2010, and 2010-2015. I've also used different colors to represent the values for different years. The Y-axis is the Annual rate of change of the refugee stock, and the X-axis is TimePeriod. Referring to the plots, I can see in 1995-2000, the annual rate of change of the refugee stock has the most extreme outliers compared to others. Additionally, I can see the median annual rate of change of the refugee stock for each time period is similar.

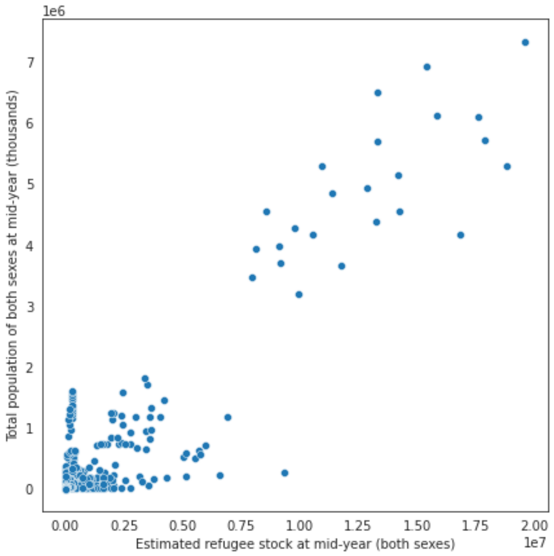
```
Text(0, 0.5, 'Annual rate of change of the refugee stock')
```

Violin plot for Annual rate of change of the refugee stock in different time periods



In the use of scatter plot, I’ve titled it “Scatter plot showing the relationship between Estimated refugee stock and Total population”. I’ve believed they are related in that an increase in the total population will cause an increase in the estimated refugee stock for the future. The Y-axis is the Total population of both sexes at mid-year (thousands), and the X-axis is the Estimated refugee stock at mid-year (both sexes). As the chart shows, I can see the relationship between those two variables is positive which estimated refugee stock increases as the total population increases.

Scatter plot showing the relationship between Estimated refugee stock and Total population



In the end, for the Maximize data-ink ratio concern, all of my charts avoid using unnecessary elements as I've shown above.

4. Discussion

In conclusion, I've found certain connotations behind the database by using statistical tools. During the process of data analysis, I also realized the significance of data visualization. The use of charts in data visualization is critical because it can help data analysts efficiently convey the information related to the database. Moreover, data cleaning and wrangling are also important for data visualization because I need to transform the database into an appropriate form that I need for data visualization. I've faced some problems in both data wrangling and cleaning and visualizing processes. Firstly, I was struggling to efficiently pick up certain columns or rows. Secondly, I was struggling to find strong causality between two compared groups from the given database.

5. Reference

1. Witte, R., & Witte, J. (2017). *Statistics* (11th ed.). John Wiley & Sons Inc.
2. Ribecca , S. (2022). *The Data Visualisation Catalogue.* Retrieved December 15, 2022, from https://datavizcatalogue.com/