INF 1340 Final Project

Exploratory Analysis on Cleaned UN International Immigrant Stock 2015 Dataset

Student Name: Jiazhou Bi

Student Number: 1008537255

Professor: Dr. Shion Guha

Introduction

Data visualization is an extremely important part of a data scientist's job, because the model is only as good and acceptable as the presentation of the model itself, as well as how people can understand it well enough to utilize it. Therefore, delivering the data clearly can be very beneficial for anyone who is presenting data related material. As mentioned in the textbook, data visualization is one of the most effective ways of representing data when used properly as it can help convey our knowledge to others . Furthermore, we will be following Tufte's Principles in our data visualization to ensure our visualizations can be used in an effective, non-misleading manner. Tufte's Principles include the following: graphical integrity, data-ink, chartjunk, data density, and small multiples (Andrew, 2020). Throughout the discussion section, these principles will be mentioned again when discussing visualizations that follow these principles.

Methods

Dataset

The dataset used in this project is the tidy data I got from our midterm project. The original dataset is the UN International Immigrant Stock 2015 dataset, and it was cleaned following the five tidy data principles taught in during this course. The tidy data contains the same information as the original dataset, but in a way that is much more convenient for conducting data analysis.

Because this dataset contains a multitudinous amount of information, it is impossible to deeply analyze the whole dataset. Therefore, I must narrow down on my research question so I can focus on the part of this dataset which interests me the most.

Research Question

I would like to see how the international immigration stock of Canada, the United States of America, and North America (Bermuda, Canada, Greenland, Saint Pierre and Miquelon, and United States of America) are changing relative to the world's immigrant stock from 1990 to 2015. Are there any differences in years? What role does North America play in the world's international immigrant stock? Are there any differences between Canada and the USA? Within Canada, are both sexes behaving the same in their international immigrant patterns?

Main Variables

International migrant stock at mid-year: the number of international migrants who is currently living in the country/region/area.

Year: the year when the data was collected.

Country/region/major area: the country/region/major area where the data was collected.

Migrants as percentage of international migrant stock: the percentage of international immigration stock relative to their total population.

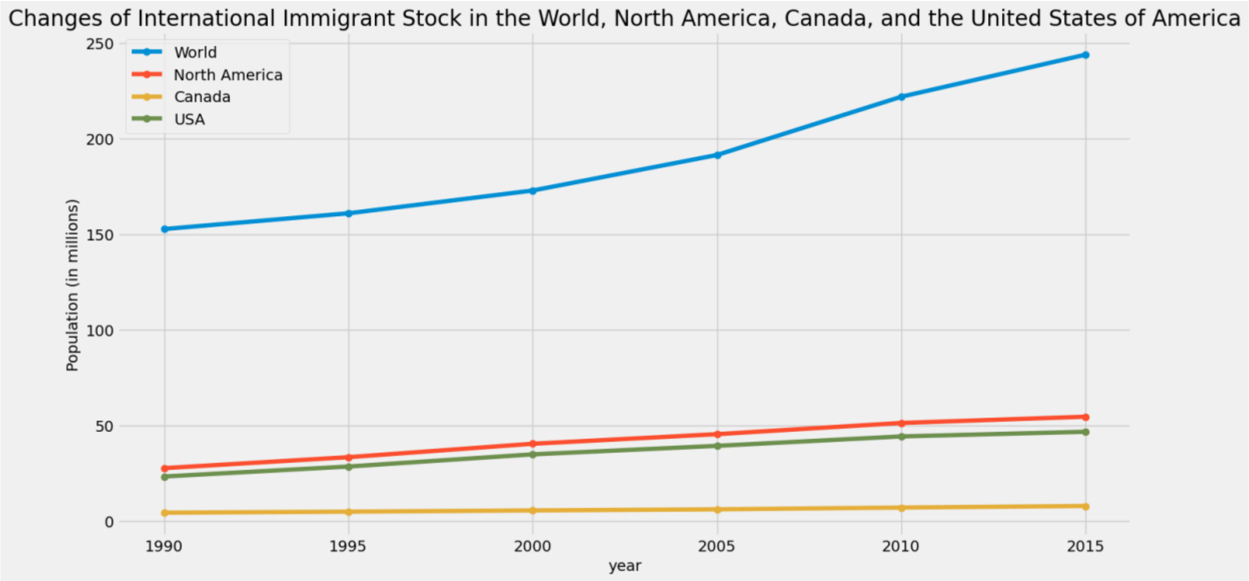
Sex: the biological sex of the individual being recorded, coded as either male or female. This binary coding of sex is decided by the original data collector, and it does not reflect the perspective of me on sex.

Analysis Methods

I have imported my dataset into Python. I have also installed some packages, including Pandas, Numpy, Matplotlib, and Seaborn to facilitate my data analysis and visualization process. My data frames from my midterm project were further wrangled in this project to suit my research questions.

Results

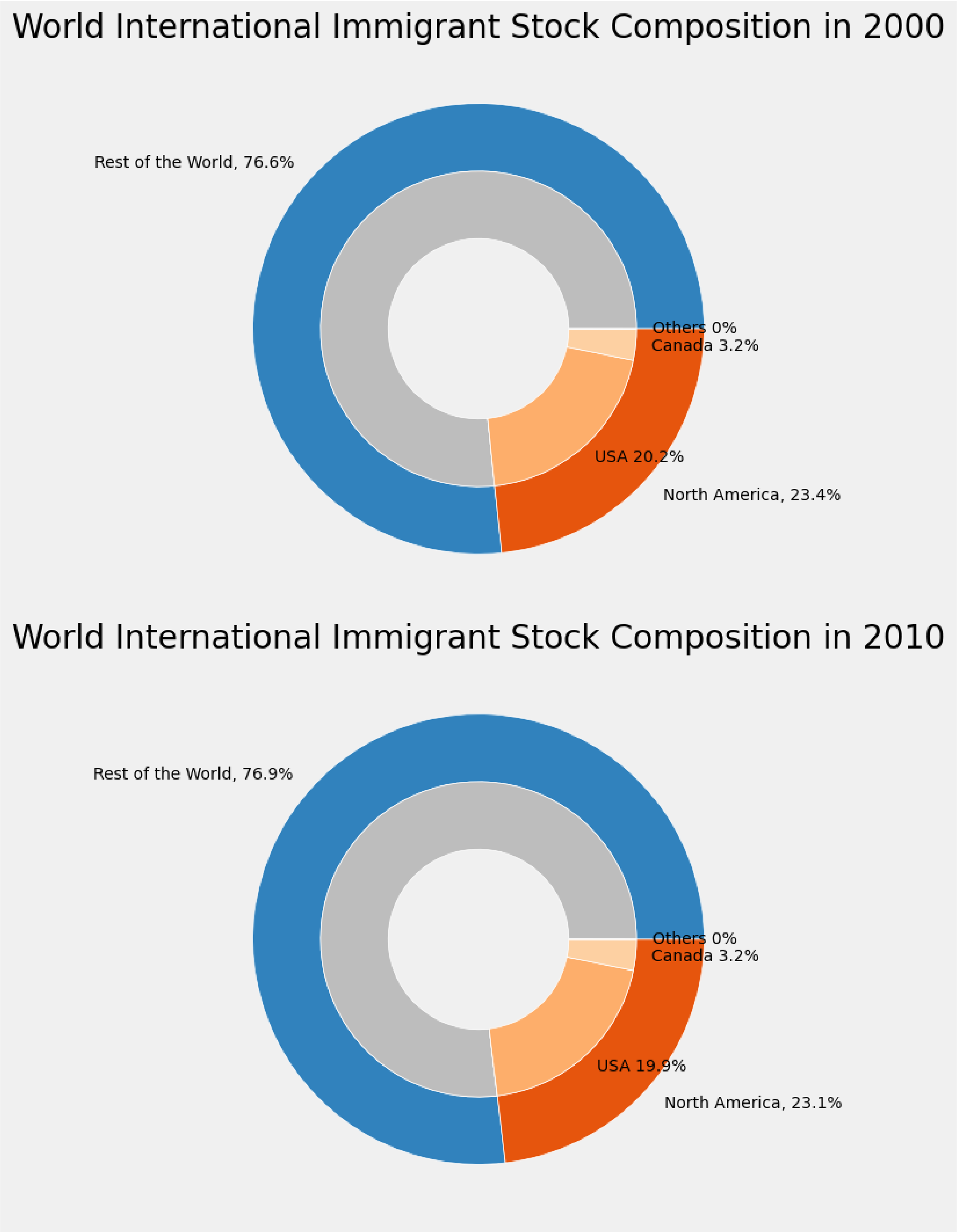
First, I had look at the trend of the overall international immigrant stock amount worldwide from 1990 to 2015, as well as the international immigrant stock amount for North America, Canada, and the United State respectively. From 1990 to 2015, the world international immigrant stock has increased for 59.74%. This number is 97.31% for North America, 80.82% for Canada, and 100.54% for the USA. These numbers are suggesting that in North America, particularly for Canada and the USA, their international immigrant stock is growing at a much quicker pace then the world in general. The overall trend is shown below as a line graph.



From the graph above, it seems that overall international immigrant stock is increasing from 1990 to 2015 for both the world and for North America, with North America's international immigrant stock almost doubled in its numbers in the 25-year period. We can also see that from 2000 to 2010, there was a significant spike in world international immigration stock amount, but this phenomenon was not directly observed in North America, where the increase in international immigrant population seems to be increasing at a steady pace. Therefore, we choose the year 2010 as our new research interest and do more analysis on the year 2010.

Because I observed a rapid increase in the world's international immigrant stock number from 2000 to 2010, I conducted further analysis to see if North America, including Canada, the USA, as well as Bermuda, and Greenland, Saint Pierre and Miquelon are playing the same role in the world's overall international immigrant stock in both 2000 and 2010. In 2020, North America accounts for 23.4% of the world's international immigrant stock. Within this 23.4%, USA occupied 20.2% of the world's international immigrant stock, this number is 3.2% for Canada, and 0% for the other areas in North America. In 2010, these number remained quite like 2000, with North America accounts for 23.1%, and 19.9%, 3.2%, and 0% for the USA, Canada, and other areas in North America, respectively. Although Dr. Guha suggested not to use pie charts, I find using a nest "donut chart" (pie chart with hollow center) can be effective in conveying the proportion relation between different parties, as well as the detailed breakdown of each party. I will explain more about this in the discussion section of this paper. The graphs were attached on the next page.

Comparing these two pie charts, despite the dramatic increase in world international immigrant stock during the 2000-2010 period, North American international immigrant stock population remained at slightly over 23%. These two graphs accurately depicted the roles played by the North American countries in the world’s international immigrant stock.

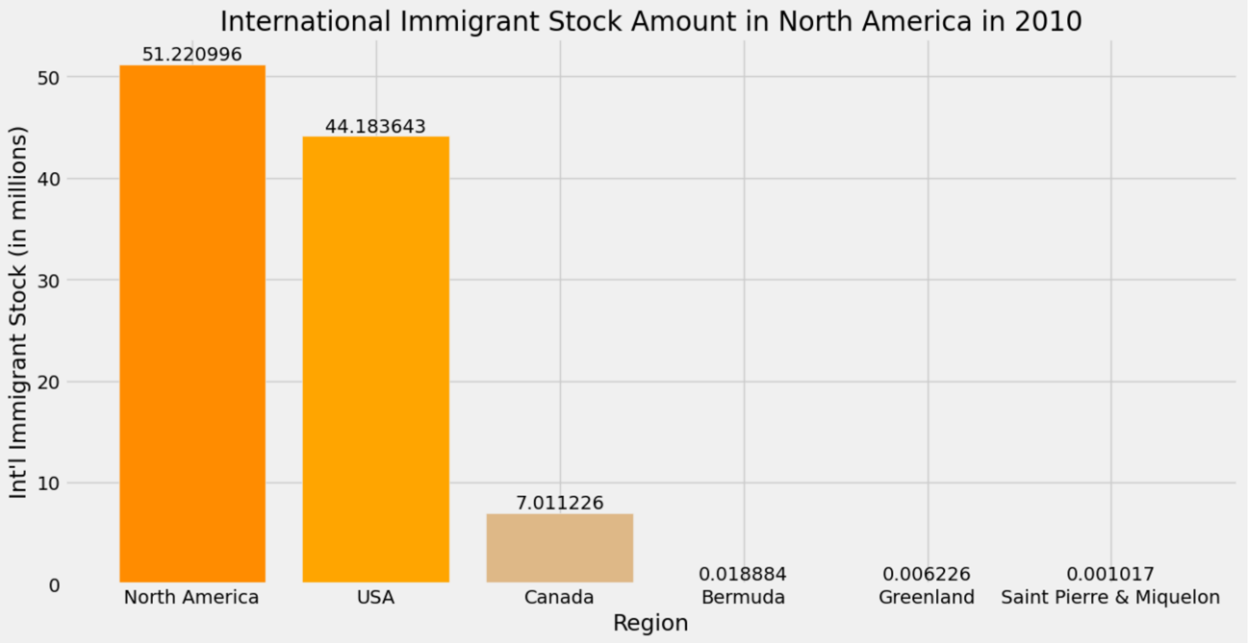


World International Immigrant Stock Composition in 2000

World International Immigrant Stock Composition in 2010

Because North America has more than 23% of the world's international immigrant stock, I would like to focus on North American only in 2010. I had a closer examine at the international immigrant stock population in North America, as well as the number for each country/area, individually to see who the major players in North America are.

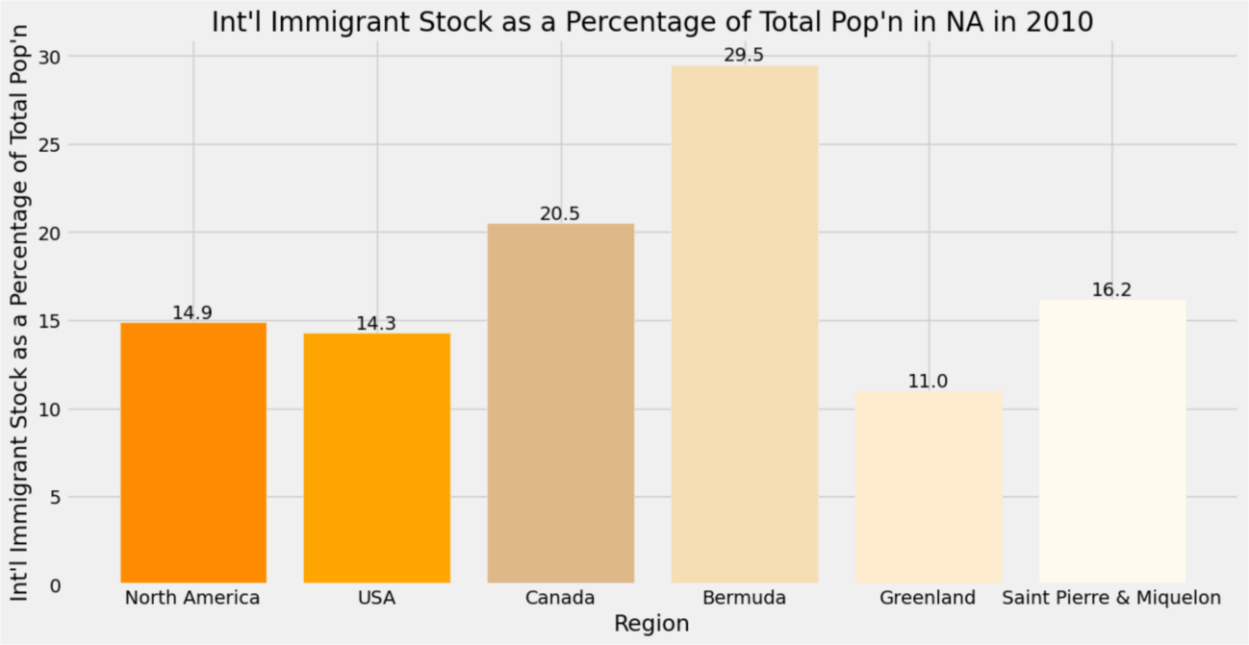
Overall, there were 51.220996 million international immigrants in North America in 2010. The USA itself had 44.183643 million, which is 86.26% of the North American number. Canada is ranking at the second place as it has 7.011226 million 13.69%. Other areas (Bermuda, and Greenland, Saint Pierre and Miquelon) had only 0.05% combined, making them neglectable in the number of international immigrants they have, relatively speaking. The bar chart is attached below.



However, we also know that the USA and Canada both have a much larger population than Bermuda, Greenland, and Saint Pierre and Miquelon. As a result, if we want to know more about their population composition, it is more meaningful to compare their percentage of international immigration stock relative to their total population, respectively.

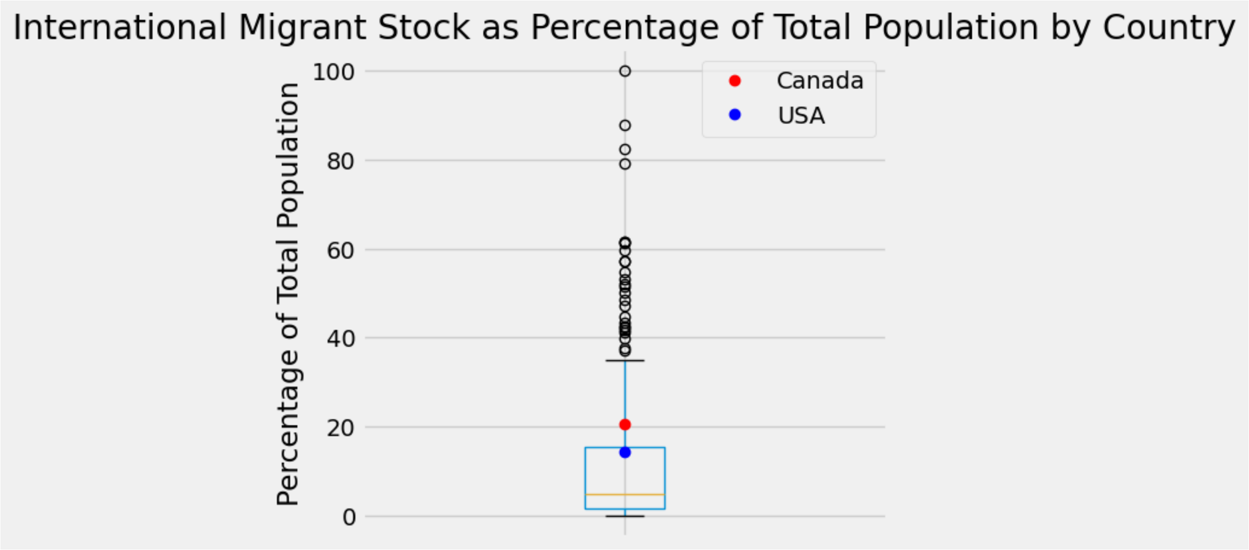
When we compare the percentage of international immigration stock relative to their total population, Bermuda is leading in North America, with 29.5% of its total population being international immigrants. The second place is Canada with 20.5%. The third is Saint Pierre and

Miquelon with 16.2%. The USA had 14.3% of its total population being international immigrants, and Greenland had 11%. The graph is pasted below.



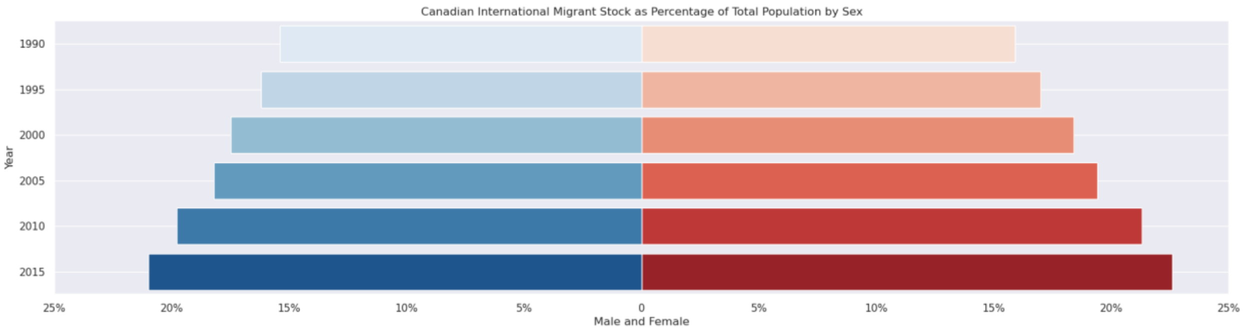
Now I know that Canada and the USA have around 21% and 14% of their total population being international immigrant stock, but I would like to know how these two countries are behaving, comparing to all the other countries/areas in the world. So, I wrangled the data and did some descriptive statistical analysis.

The mean of international migrant stock as percentage of total population is 13.12% and the standard deviation is 17.93%. The range is from 0.06% to 100%, with the IQRs being 1.58% and 15.36%. I have drawn a boxplot with all the countries/areas' international migrant stock as percentage of total population, and highlighted Canada and the USA in the same plot to make it more informative. The graph is on the next page.



The graph above is suggesting that international migrant stock as percentage of total population by country is significantly skewed to the right, with many outliers highlighted in black circles in the box plot above. Comparing to other countries, Canada, and USA both have a higher-than-average international migrant stock as percentage of total population but remained in a "normal" range.

Finally, I took a closer look on Canadian international migrant stock as percentage of total population by sex from 1990 to 2015. Are the patterns similar for both sexes, or different? It turned out that for both male and female, their international migrant stock as percentage of total population in Canada from 1990 to 2015 were increasing at around 7% per 5 years. For a direct comparison, I have plotted a pair of six-observation data into a population pyramid. I have attached it below.



From the graph, both sexes are increasing at similar speeds, which confirmed our data. Both sexes are behaving in the same pattern in their change in the international migrant stock as percentage of total population in Canada from 1990 to 2015. However, because the data was too small (in terms of five-year-period), we cannot conduct a meaningful t-test.

Discussion

For better illustrative purposes, I will divide this section into five different subsections, as there are seven different (five different types of) graphs in this paper. If we exclude the two donut graphs, there are five graphs in this project. I will talk about them individually and mention how they are echoing with Tufte's Principles.

Line Graph

We found that both the world and North America are increasing in their international immigrant stocks number. The world has increased almost 70% in this number during the 25-year period, whereas North America almost doubled their number. The line graph did a great job in delivering this information as it can show how the numbers are changing. With no manipulation on any of the axes, this graph has a high graphical integrity, maximized data-ink, and high data-density. There were also four lines in the graph, making it still legible but also has some small multiples.

Donut Graph

Donut graph, or pie chart, was strongly not recommended by the professor as the amount of information it can convey is low and may be better present with a table. However, with the example I presented in this paper, I argue that if we would like to show the composition of a certain concept with sub-parties involved, a nested donut graph can be efficient as well because different layers can show the sub-composition, as well as their belonging relationship in a direct manner. From the two nested donut graphs, we saw that despite the dramatic increase in world international immigrant stock during the 2000-2010 period, North American international immigrant stock population remained at slightly over 23%. Indeed, donut graph can have lower data-density, and higher chartjunk. In my case, it has enough complexity so that it has a high graphical integrity and small multiples.

Bar Graph

With the bar charts, we effortlessly compared the international immigrant stock number, as well as international immigrant stock as percentage of total population within North America. Although the USA and Canada have the most international immigrant stock number, Bermuda has the highest percentage of its population being international immigrants. Bar charts are straight forward in comparing the difference among groups. Without intentional manipulation, according to Tufte's principle, bar chart can have very low chartjunk (Sigdel, 2020), and high data-density and data-ink.

Boxplot

Boxplot is frequently used in statistical analysis as it shows the mean, IQR, as well as potential outliers within the data. We can also highlight the spot we are interested in to make the graph more informative. In our boxplot, it shows that international migrant stock as percentage of total population by country is significantly skewed to the right, with many outliers highlighted in black circles in the box plot above. Comparing to other countries, Canada, and USA both have a higher-than-average international migrant stock as percentage of total population but remained in a "normal" range. Boxplot is a great visualization tool because it has a high graphical integrity, high data-ink, low chartjunk, high data density, as well as high small multiples if we have a large amount of data. Because it fits all the Tufte's Principles, it is an excellent way of visualizing data.

Population Pyramid

The last graph of this paper is a population pyramid. It is essentially two bar charts that were glued together to form on graph for direct comparison on two different independent variables at the same time. From our population pyramid, we concluded that both sexes follow the same pattern. We can conclude that Canadian international migrant stock as percentage of total population for both sex from 1990 to 2015 are steadily increasing. Population pyramid can be useful in delivering information because it has high data-ink, high data-density, and very high small multiples. However, it can also be slightly higher on chartjunk. Therefore, population pyramid should be used with caution so that it is not distracting our audience from the information we want to convey.

Conclusion

In conclusion, we have used 5 different types of visualizations to show some of our findings from the UN International Immigrant Stock dataset. We first used line graph to see the overall trend of international immigrant stock number of the world, North America, USA, and Canada. Then we used pie chart to show the role of North American international immigrant stock, compared to the rest of the world. After that, we drew bar charts to see the number of, as well as the percentage of international immigrant stock relative to the total population of each country/area in North America. Following that, we used box plot to see the percentage of international immigrant stock relative to the total population of both Canada and USA, compared to other countries/areas all over the world. Lastly, we used population pyramid to compare the patterns for both male and female international immigrant stock in Canada from 1990 to 2015. Each type of data visualization has its strengths and weaknesses. The best practice is to follow Tufte's Principles and combine different tools to effectively communicate our information to the audience.

References

Andrew (andrewtk). (2020). Tufte's Principles. *https://thedoublethink.com/tuftes-principles-for-visualizing-quantitative-information/* Retrieved Dec. 2022.

Aragon, C., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-Centered Data Science: An Introduction*. MIT Press.

Sigdel, Rajesh. (2020). Improve Your Visualization Skills Using Tufte's Principles of Graphical Design. *https://medium.com/nightingale/improve-your-visualization-skills-using-tuftes-principles-of-graphical-design-3a0f40a53a2c* Retrieved Dec. 2022.