Fatima Tuz Zohra (1004167876)
INF1340: Programming for Data Science

Data Visualization Assignment

1.Introduction:

Data Visualization is the process of representing data and information through various graphical representations such as boxplots, scatter plots, line plots, and bar charts. The primary purpose of data visualization is to make it easy to better understand and analyze complex data. Edward Tufte, a statistician and professor at the University of Yale, laid out the principles of data visualizations. The first principle is the concept of sorting – which involves the process of organizing data in a way that makes it easy to compare and understand. This method allows individuals to see patterns and trends in the raw data. The second principle is the concept of grouping – which involves the process of organizing the data into logical groups/categories. The third principle is idea of subsets – which involves the process of breaking down large groups of data into smaller, more manageable parts that are easier to understand and analyze. The fourth principle is the idea of comparison in the form of small multiples – which is the process of showing different aspects of data side-by-side in order to make meaningful conclusions. In this project, we worked to create various visualizations that adhere to Tufte's principles for the UN Migrant Stock (2015) dataset that was previously cleaned.

2. Methods:

The dataset that was used was the UN Migrant Stock Total (2015) which was previously tidied in the Midterm assignment. The detailed coding steps for the visualizations of each of the 6 tables can be found in the attached IPYNB file.

2.1: Tufte's Principle of Sorting

In order to apply Tufte's principle of sorting to Table 1, the 'sort_values' function was used for the International Migrant Stock at Mid-Year data to filter the data in descending order, and it was to sort the Location data in alphabetical order. This sort function was used to better understand the relationship between International Migrant Stock at Mid-Year and Location.

2.2: Tufte's Principle of Grouping

The executive summaries were created in the midterm assignment to separate the large dataset into smaller, more manageable sections. It includes the data for five locations; World, Developed Regions, Least Developed Countries, Less Developed Regions excluding Least Developed Countries, and Sub-Saharan Africa. In order to visualize the executive summaries, we grouped the data by the year and gender using the groupby function.

For example, when visualizing the executive summary of Table 1, we grouped the International Migrant Stock (IMS) at Mid-Year data by year and gender. We also took the sum of the IMS data. This code aligns with this principle because it reduces the clutter and redundancy of the data by grouping similar data together in order to represent the unique combinations of the Year and Gender values. In addition, by summing the IMS data for each group, we further make the data more concise and make the visualization more simple and easier to interpret.

2.3: Tufte's Principle of Subsets

In our main tables, we had many locations that were separated into three categories: Major Area, Region and Country/Area. In order to better understand and visualize this data, we separated the visualizations for the location by Major Area and Region. Using these subsets, we were able to better understand the trends that were occurring.

For example, when generating a bar chart for Table 1, the database was first filtered to include only the rows where the category column was equal to major area. A list was then created for each of the major areas and a for loop was used to iterate over the list of the major areas. For each of the major areas in the list, the code filters the data frame to include only the rows in which the location column is equal to the current major area. A bar plot is then generated for each of the major areas. The hue function is used to separate the data based on gender. This aligns with Tufte's Principle of Subsets as we represent the data for each of the major area separately and we can focus on this small subset of the locations and gain a better understanding of the trends.

2.4: Tufte's Principle of Comparison

In our tidied dataset, we have a table that combines the executive summaries of Table 1 and 2 in order to generate a table that has data for International Migrant Stock and Total Population at Mid-Year. In order to represent this data in a bar plot visualization, we grouped the year and gender and summed the values in the International Migrant Stock and Total Population at Mid-Year for each group. A figure with two axes was then created using the 'plt.subplots' function that assigns the axes to the variables ax1 and ax2.

This visualization aligns with Tufte's Principle of Comparison, that states it is more effective at times to display comparisons between different groups rather than showing them separately. In this visualization, we compare the international migrant stock at mid-year and total population at mid-year for the different locations on the same plot. This was it is easy to compare how the data changes over time. In addition, we can better analyze the trends and patterns.

3. Results & Discussion:

3.1: Visualizing Table 1 Executive Summary

Figure 1-6 represents the visualizations that were created for the executive summary of Table 1.

Figure 1 represents the International Migrant Stock at Mid-Year (IMS) for World from 1990 to 2015. As we can see in the Bar Chart, there is an overall increase in the International Migrant Stock at Mid-Year for both sexes, with males being higher than females.

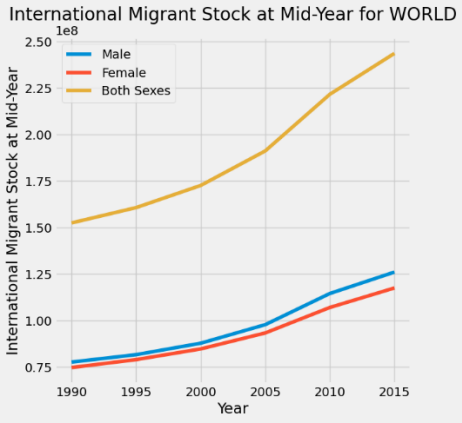

Figure 1. Line plot representing International Migrant Stock at Mid-Year for World

Figure 2 represents the International Migrant Stock (IMS) at Mid-Year for Developed Regions from 1990-2015. There is a general increase over the years both sexes with the migrant stock leveling off from 2010 to 2015. The female IMS was higher than the males.

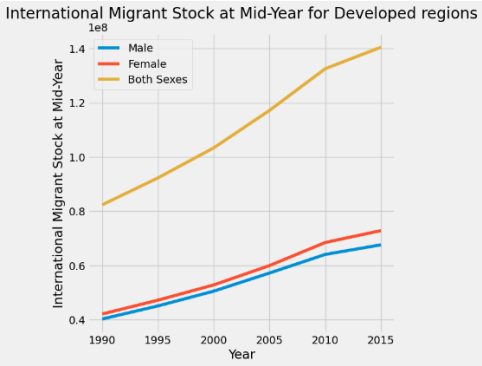

Figure 2. Line plot representing International Migrant Stock at Mid-Year for Developed Regions

Figure 3 represents the International Migrant Stock (IMS) at Mid-Year for Developing Regions from 1990-2015. From 1990-2005, the migrant stock stays constant for both sexes. There is a slight increase from 2000 to 2005 and a rapid increase from 2005-2015. From both sexes, the male IMS is higher than the female.

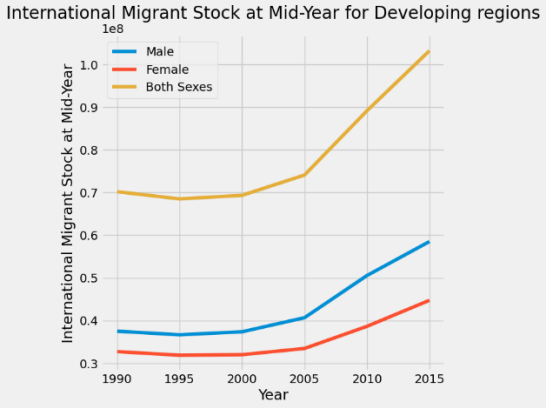

Figure 3. Line plot representing International Migrant Stock at Mid-Year for Developing Regions

Figure 4 represents the International Migrant Stock (IMS) at Mid-Year for Least Developed Countries from 1990-2015. For both sexes, IMS increased from 1990-1995, followed by a decline in IMS from 1995-2005. From 2005, we see a slight increase in IMS with a rapid increase from 2010-2015. From both sexes, the male IMS was higher than the female.

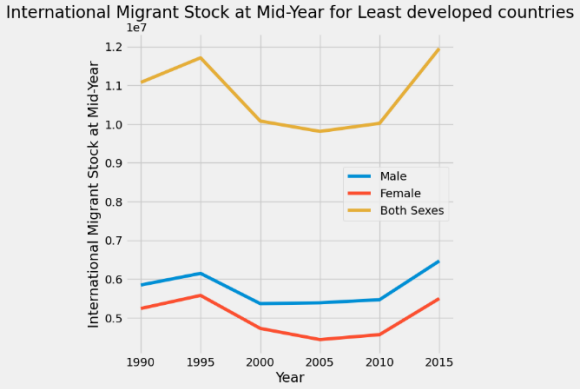

Figure 4. Line plot representing International Migrant Stock at Mid-Year for Least Developed Countries

Figure 5 represents the International Migrant Stock (IMS) at Mid-Year for Less developed regions excluding least developed countries from 1990-2015. For both sexes, IMS decreased from 1990-1995 and increased from 1995-2015. The males IMS was higher than females.

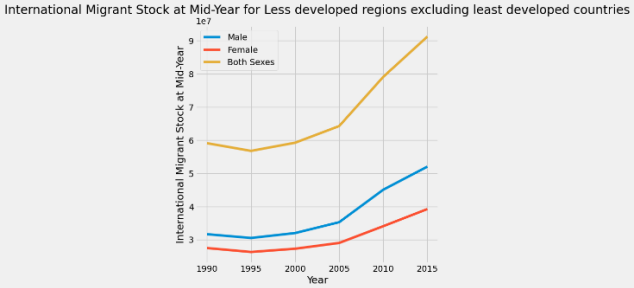

Figure 5. Line plot representing International Migrant Stock at Mid-Year for Less Developed Regions Excluding Least Developed Countries

Figure 6 represents the International Migrant Stock (IMS) at Mid-Year for Sub-Saharan Africa from 1990-2015. For both sexes, IMS increased from 1990-1995, decreased from 1995-2000 and then increased again from 2000-2015. The female IMS was higher than the males.

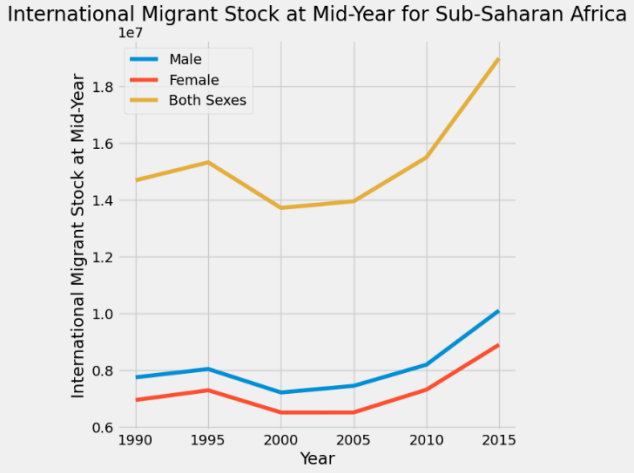

Figure 6. Line plot representing International Migrant Stock at Mid-Year for Sub-Saharan Africa

3.2: Visualizing Table 1 by Major Area

Figure 7 is a boxplot that represents the IMS at Mid-Year for the Major Areas (Africa, Asia, Europe, Latin America & Caribbean, Oceania). When examining all the Major Areas for both sexes, we find that Europe has the highest IMS at Mid-Year. For Asia and Africa, we found that the male IMS was higher than females. For the remaining major areas, the female IMS was higher than the male.

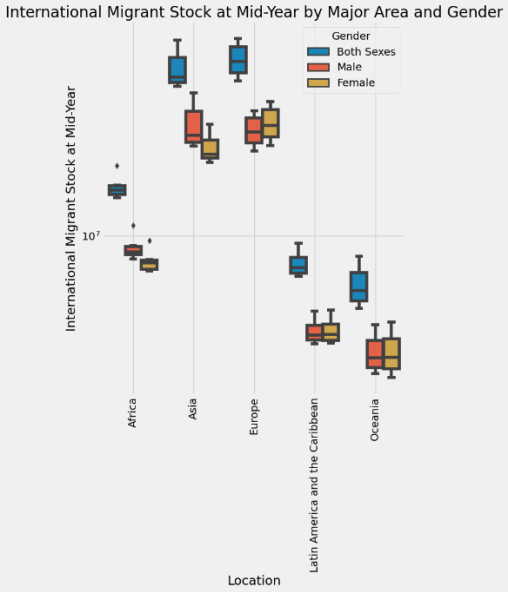

Figure 7. Boxplot representing International Migrant Stock at Mid-Year for Major Areas

3.3: Visualizing Table 1 by Region

In Figure 8-13, bar plots were generated to represent the International Migrant Stock (IMS) at Mid-Year by Region for each year (1990, 1995, 2000, 2005, 2015).

In Figure 8, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 1990.

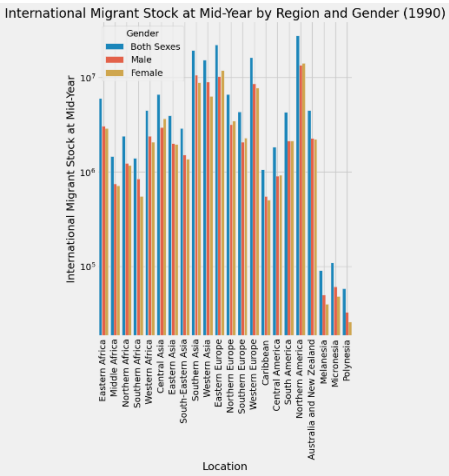

Figure 8. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (1990)

In Figure 9, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 1995.

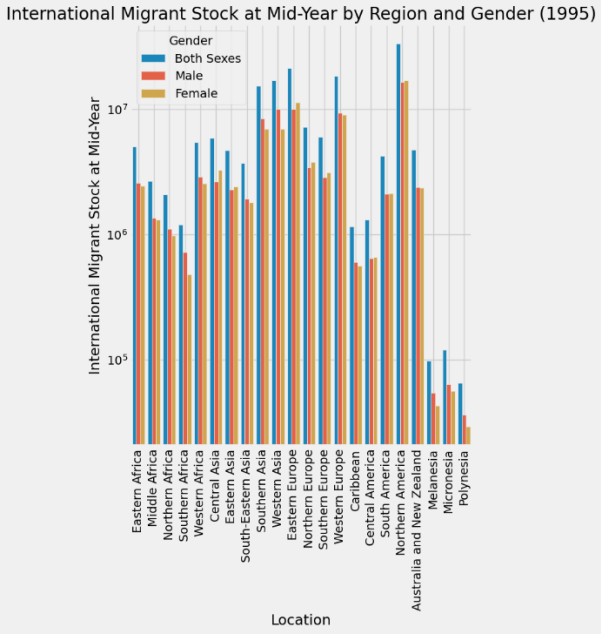

Figure 9. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (1995)

In Figure 10, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 2000.

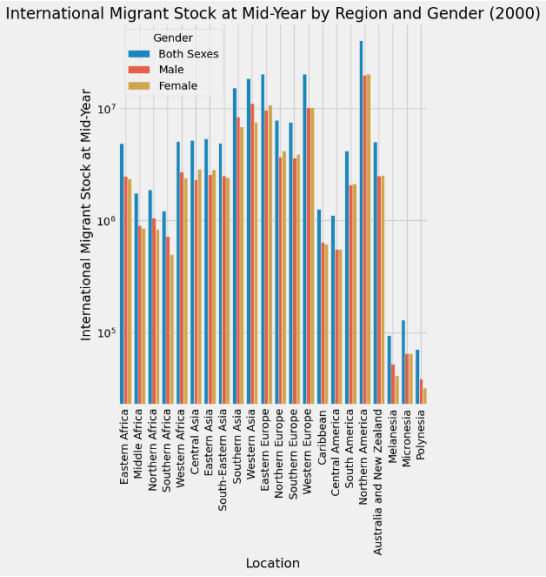

Figure 10. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (2000)

In Figure 11, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 2005.

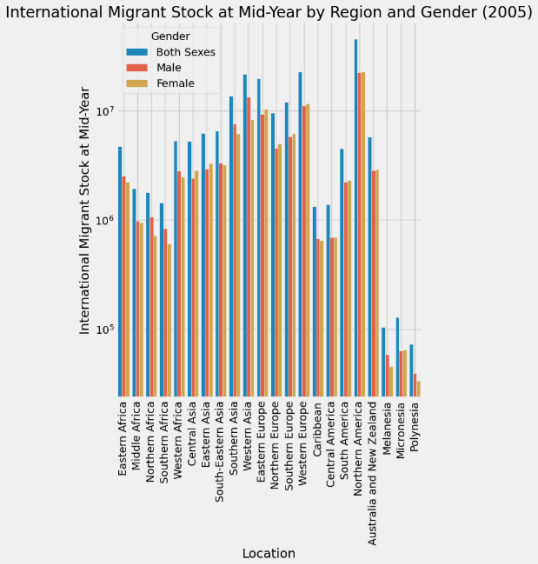

Figure 11. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (2005)

In Figure 12, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 2010.

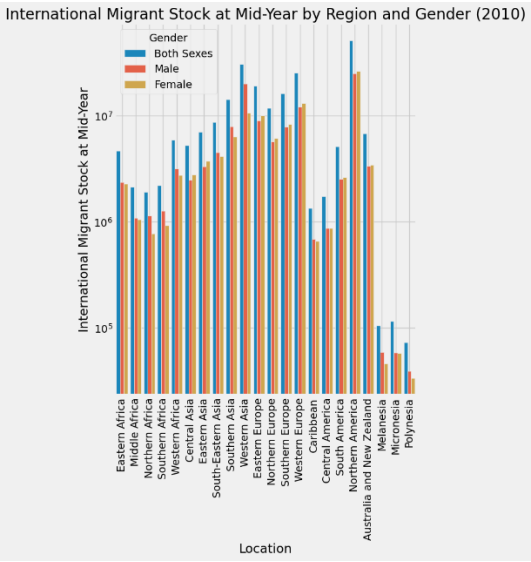

Figure 12. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (2010)

In Figure 13, we can see that Northern America had the highest IMS for both sexes, with females being higher than male during 2015.

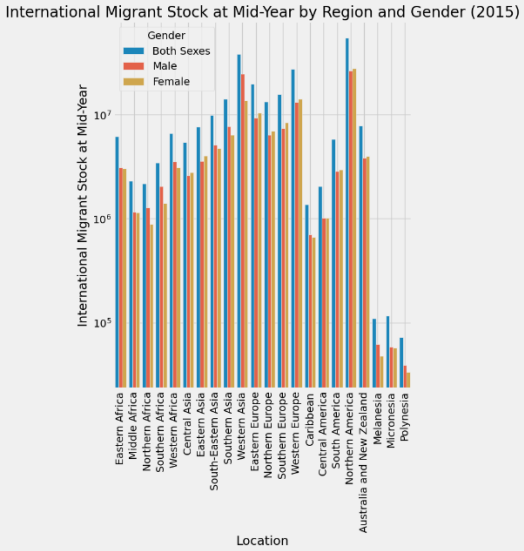

Figure 13. Bar Plot representing International Migrant Stock at Mid-Year by Region and Gender (2015)

3.4: Visualizing Table 2 Major Area

In Figure 14-18, bar charts were generated to represent the total population at Mid-Year for the Major Areas from 1990-2015.

In Figure 14, it can be seen that both sexes had the highest population at Mid-Year with their being equal amounts of male and females during 2015 in Africa. There was a general increase in population numbers.

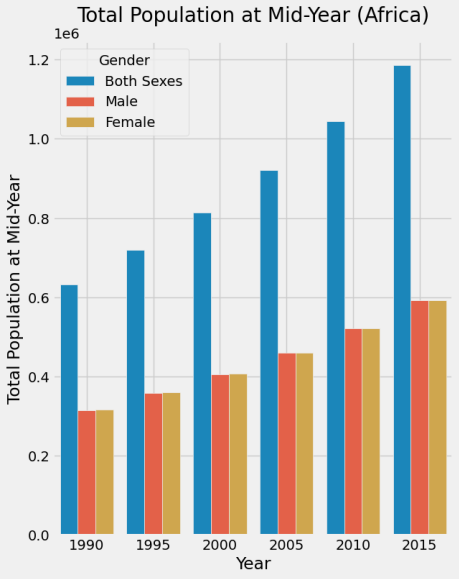

Figure 14. Bar Chart representing Total Population at Mid-Year (Africa)

In Figure 15, there is a general increase in the total population at mid-year from 1990-2015 in Asia. The highest total population for both sexes was in 2015, with males being higher than females.

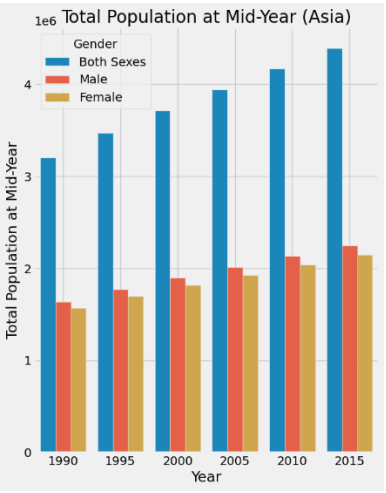

Figure 15. Bar Chart representing Total Population at Mid-Year (Asia)

In Figure 16, there is a slight increase in the total population at mid-year from 1990-2015 in Europe. The highest total population for both sexes was in 2015, with females being higher than males.

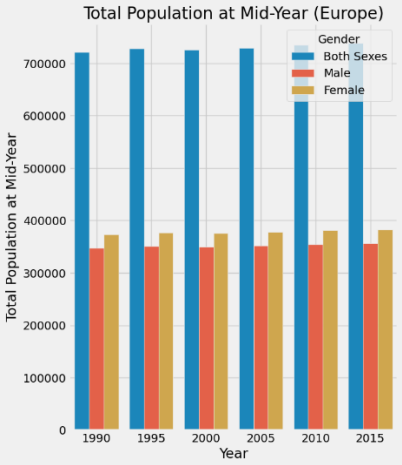

Figure 16. Bar Chart representing Total Population at Mid-Year (Europe)

In Figure 17, there is a slight increase in the total population at mid-year from 1990-2015 in Latin America and the Caribbean. The highest total population for both sexes was in 2015, with females being higher than males.

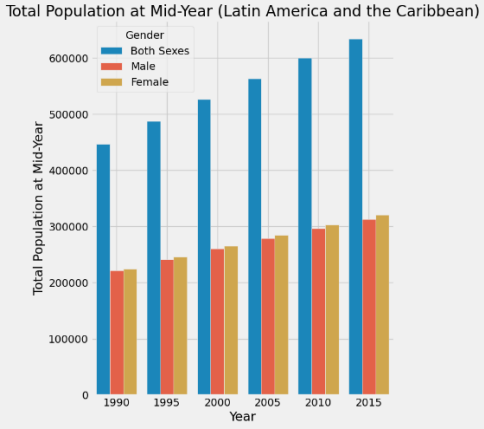

Figure 17. Bar Chart representing Total Population at Mid-Year (Latin America and the Caribbean)

In Figure 18, there is a slight increase in the total population at mid-year from 1990-2015 in Oceania. The highest total population for both sexes was in 2015, with males being higher than females.

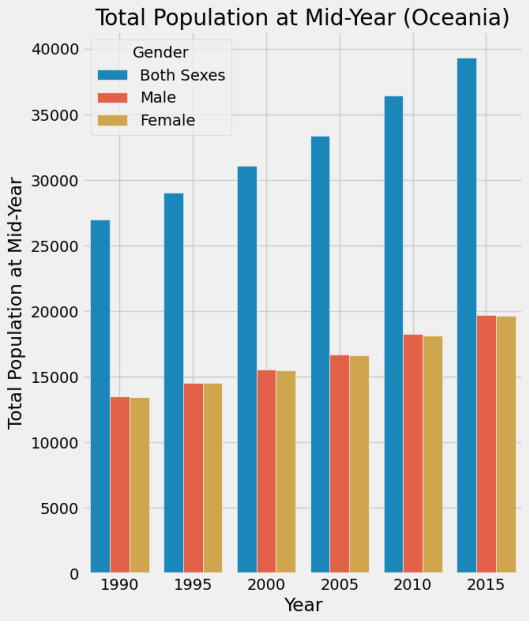

Figure 18. Bar Chart representing Total Population at Mid-Year (Oceania)

3.5: Visualizing Table 1 & 2 Executive Summary

In Figure 19-24, side-by-side bar charts were generated for International Migrant Stock at Mid-Year and Total Population at Mid-Year using the executive summary tables with year and gender grouped.

In Figure 19, IMS and Total Population for the World was highest in 2015 for both sexes, with females being higher than males.

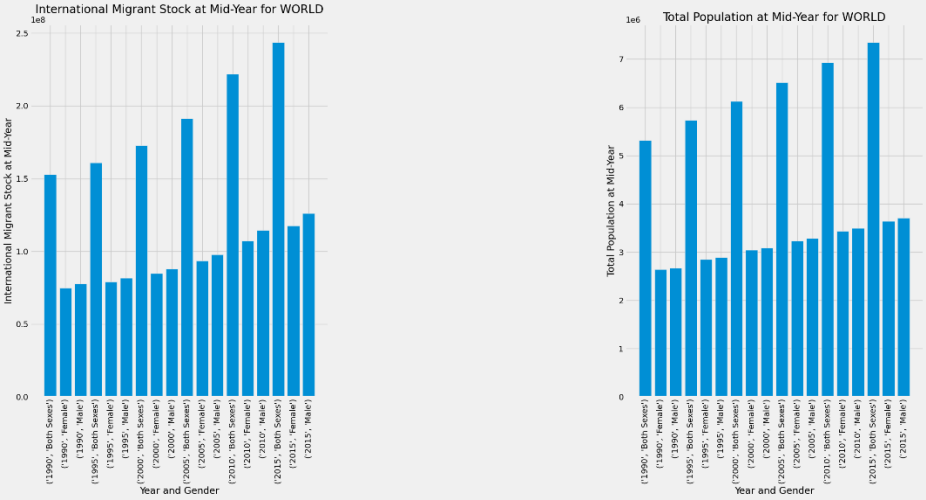

Figure 19. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for World by Year & Gender

In Figure 20, IMS and Total Population for the Developed regions was highest in 2015 for both sexes, with females being higher than males.

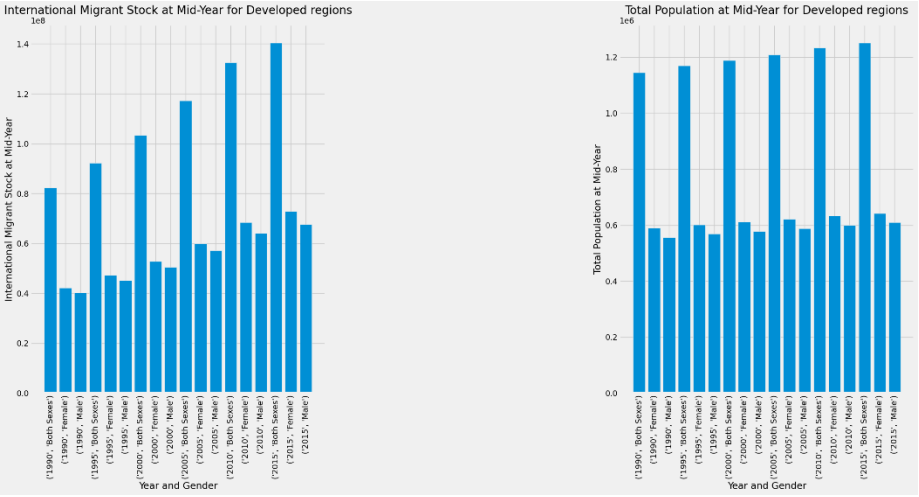

Figure 20. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for Developed Regions by Year & Gender

In Figure 21, IMS and Total Population for the Developing regions was highest in 2015 for both sexes, with males being higher than females.

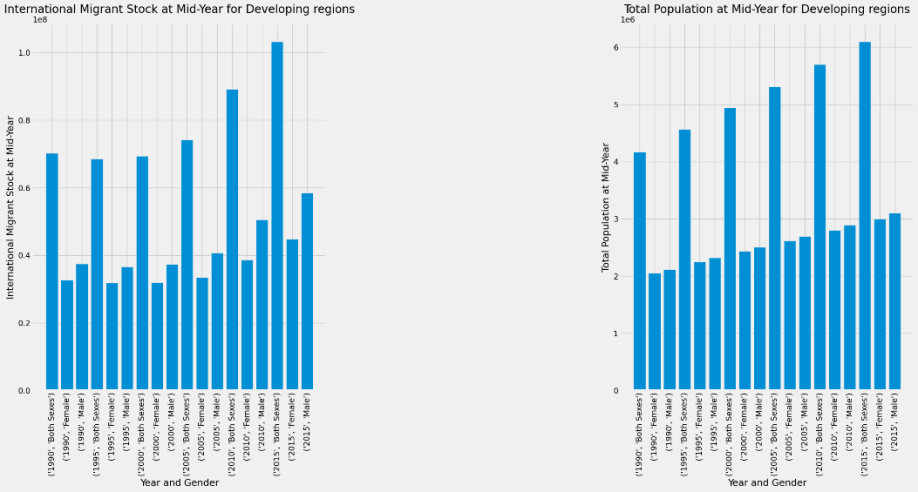

Figure 21. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for Developing Regions by Year & Gender

In Figure 22, IMS and Total Population for the Least developed countries was highest in 2015 for both sexes, with males being higher than females.

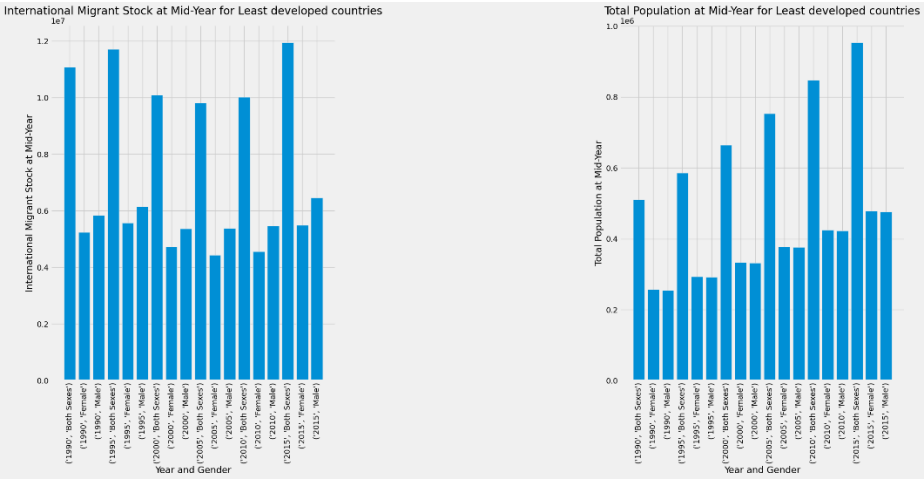

Figure 22. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for Least Developed Countries by Year & Gender

In Figure 23, IMS and Total Population for the less developed regions excluding least developed countries was highest in 2015 for both sexes, with females being higher than males.

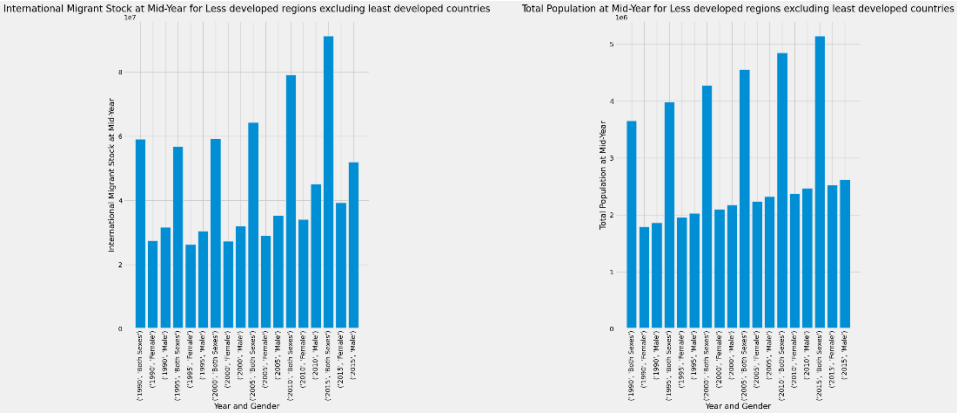

Figure 23. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for Less developed regions excluding least developed countries by Year & Gender

In Figure 24, IMS and Total Population for the World was highest in 2015 for both sexes, with the IMS for males being higher than females but the total population being the same for both genders.

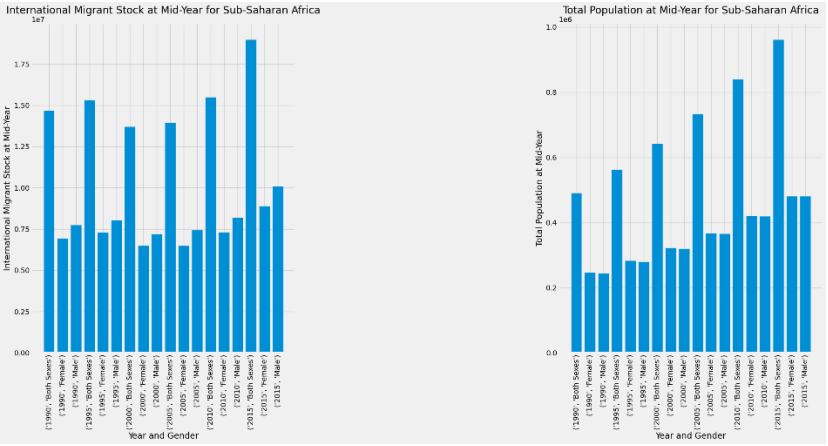

Figure 24. International Migrant Stock at Mid-Year (right) and Total Population at Mid-Year (left) for Sub-Saharan by Year & Gender

3.6: Visualizing Table 3 Executive Summary

In Figure 25, a Scatterplot was generated to represent the International Migrant Stock as a Percentage of Total Population for each the locations in the executive summary. For the World, we found that the highest percentage was in 2015 for males. Similar figures were generated for the remaining locations in the executive summary that can be found in the IPYNB file.

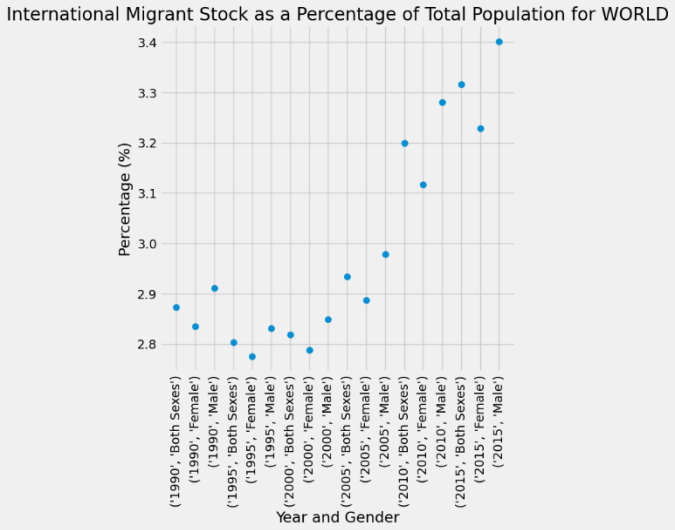

Figure 25. Scatterplot representing International Migrant Stock as a Percentage of Total Population for World by Year & Gender

3.7: Visualizing Table 3 Major Area

We generated line plots for each major areas to represent the IMS as a percentage of total population grouped by year and gender. Only one example was shown but the remaining can be found in the IPYNB file.

In Figure 26, we represent the data for Africa and find that the percentage followed a general decline from 1990-2010 and then increased from 2010-2015. The percentage from both sexes was the highest for males.

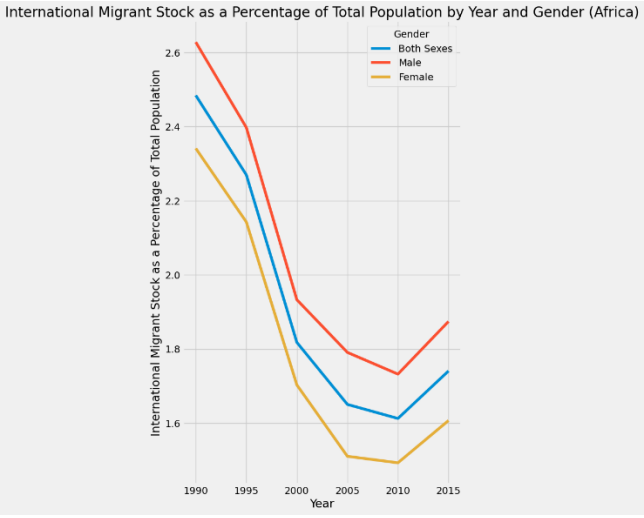

Figure 26. Line Plot representing International Migrant Stock as a Percentage of Total Population by Year and Gender (Africa)

3.8: Visualizing Table 4 Executive Summary

We generated scatter plots for each location in the executive summary that represent the migrants as a percentage of the international migrant stock from 1990-2015. Only one example was shown but the remaining can be found in the IPYNB file.

In Figure 27, the scatter plots generated to represent the migrants as a percentage of total population grouped by year and gender for World. The percentage increased from 1990-1995 and then decreased from 2000-2015.

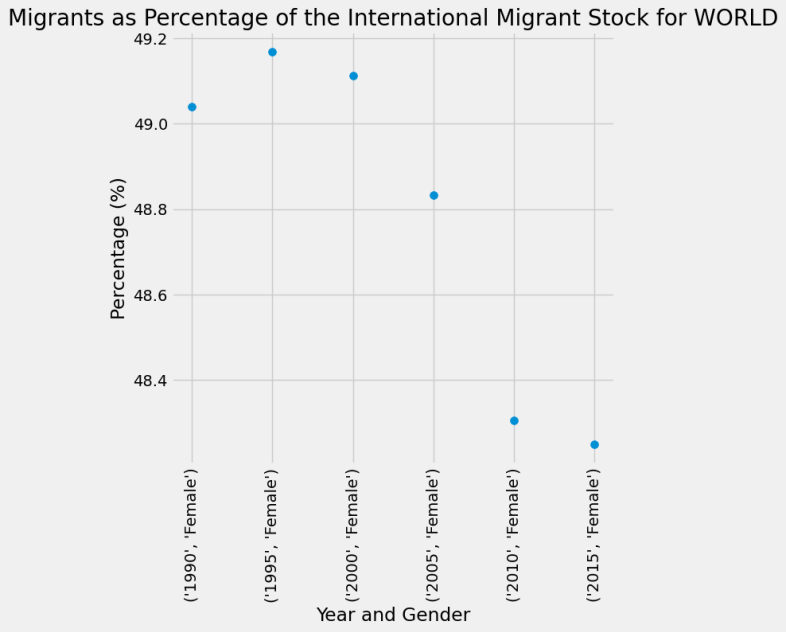

Figure 27. Scatterplot representing Migrants as a Percentage of International Migrant Stock for World

3.9: Visualizing Table 4 Major Area

In Figure 28, a boxplot was generated representing the female migrants as a percentage of the International Migrant Stock for each of the major areas. The percentage was the highest in Europe and lowest in Asia.

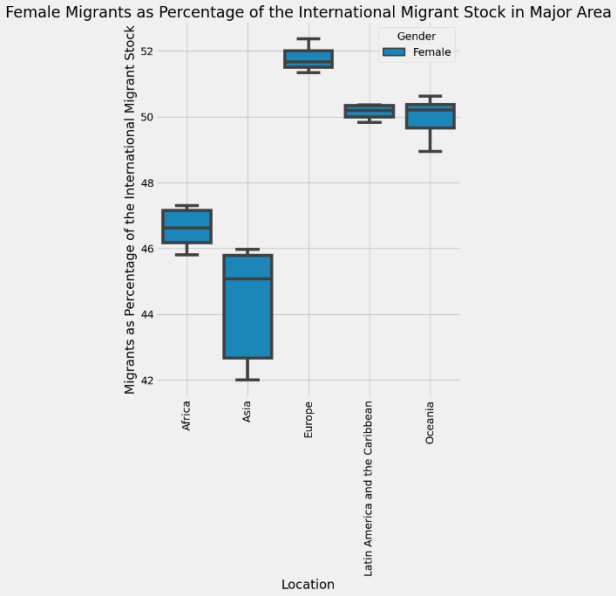

Figure 28. Boxplot representing female migrants as a percentage of the International Migrant Stock in Major Area

3.10: Visualizing Table 4 Region

In Figure 29, a boxplot was generated representing the female migrants as a percentage of the International Migrant Stock for each of the regions. The percentage was the highest in Central Asia and lowest in Western Asia.

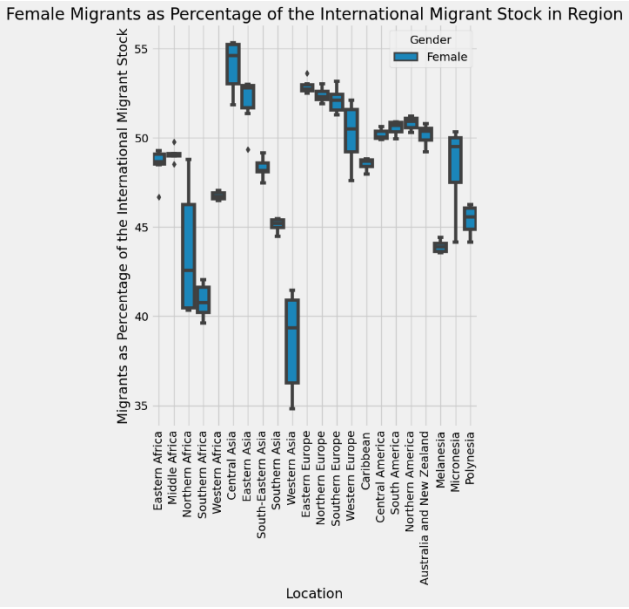

Figure 29. Boxplot representing female migrants as a percentage of the International Migrant Stock in Region

3.11: Visualizing Table 5 Executive Summary

We generated line plots for each location in the executive summary that represent the annual rate of change of the migrant stock for the various year ranges; 1990-1995, 1995-2000, 2000-2005, 2005-2010, and 2010-2015. Only one example was shown but the remaining can be found in the IPYNB file.

In Figure 30, a line plot represents the annual rate of change of migrant stock for world. The migrant stock increased from 1990-1995 to 2005-2010. The annual rate of change of the migrant stock declines from 2005-2010. We found that between both sexes, male rates were higher than female.

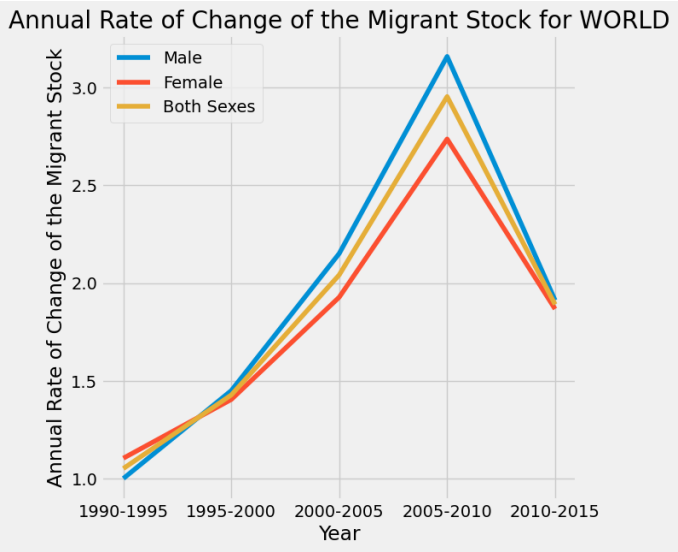

Figure 30. Line Plot representing Annual Rate of Change of Migrant Stock for World

3.12: Visualizing Table 5 Major Area

In Figure 31, a boxplot was generated representing the annual rate of the migrant stock in the major areas. Asia had the highest annual rate of change of the migrant stock with male being the highest between both sexes.

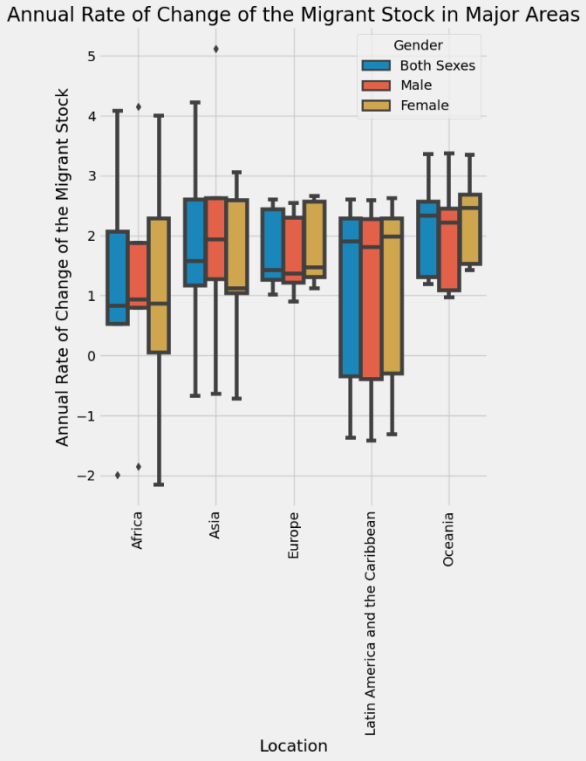

Figure 31. Boxplot representing Annual Rate of the Migrant Stock in Major Area

3.13: Visualizing Table 6 Estimated Refugee Stock Executive Summary

We generated bar charts for each location in the executive summary that represent the estimated refugee stock at mid-year for each location with year and gender grouped. Only one example was shown but the remaining can be found in the IPYNB file.

In Figure 32, the bar chart shows that the estimated refugee stock was the highest for both sexes for the World during 2015 and lowest in 2005.

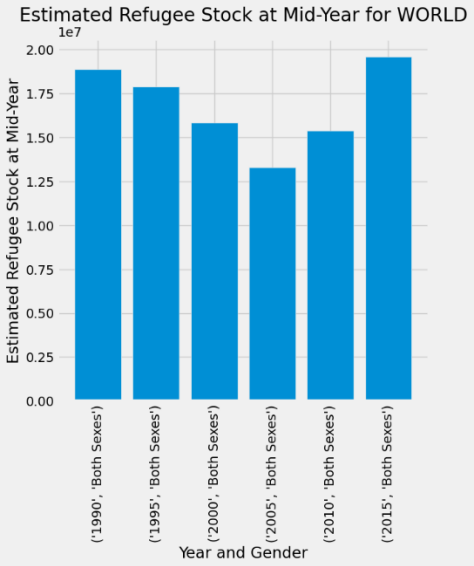

Figure 32. Bar Chart representing Estimated Refugee Stock at Mid-Year for World by Gender & Year

3:14: Visualizing Table 6 Refugees as a Percentage of the International Migrant Stock Major Area

In Figure 33, a boxplot was generated representing the refugees as a percentage of the IMS in the major areas. Africa had the highest percentage and Oceania had the lowest percentage.

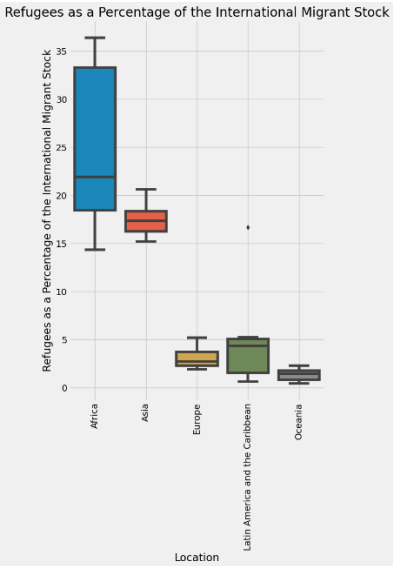

Figure 33. Boxplot representing Refugees as a Percentage of the International Migrant Stock for Major Area

3:15: Visualizing Table 6 Annual Rate of Change of the Refugee Stock Major Area

In Figure 33, a boxplot was generated representing the annual rate of change of the refugee stock in the major areas. Asia had the highest percentage and Oceania had the lowest percentage.

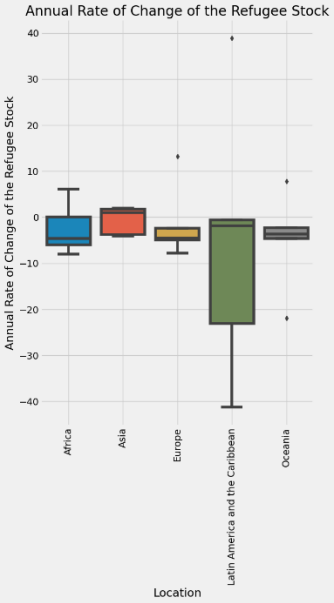

Figure 34. Boxplot representing Annual Rate of Change of the Refugee Stock by Major Area

3:16: Visualizing Table 6 Estimated Refugee Stock at Mid-Year Major Area

In Figure 35, a boxplot was generated representing the estimated refugee stock at mid-year in the major areas. Asia had the highest percentage and Oceania had the lowest percentage.

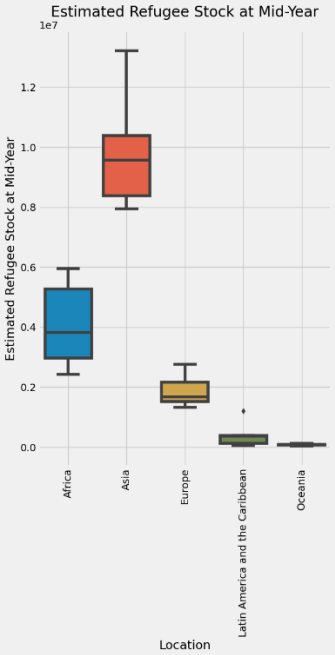

Figure 35. Boxplot representing Estimated Refugee Stock at Mid-Year by Major Area

Conclusion and Future Works

In conclusion, data visualization is a very important and effective method of understanding and analyzing the trends and patterns that exist in a dataset. Applying Tufte's principles of sorting, grouping, subsets and comparing, we can generate effective and clear visualization that are easy to interpret and visualize.

One of the limitations in this analysis was that it was difficult to visualize the data for the individual countries/areas because of the risk of overcrowding the visualizations. In future works, it would be interesting to generate subsets of the countries and areas and then create the relevant visualizations in order to understand and analyze that data.