Maidah Shah

Assignment 2: Final Project

INF1340H, Fall 2022

December 15th, 2022

EXPLORATORY DATA ANALYSIS THROUGH VISUALIZATION OF TIDIED UN DATASET

Introduction:

The goal of this final project was to perform exploratory data analysis (EDA) through visualization of the UN dataset, which was cleaned using tidy data principles in assignment 1 (A1). EDA and visualization are important to understand the high-level trends and story that a dataset is telling. These findings can then be leveraged to form more meaningful research questions, and subsequently, better data analysis.

Visualization will follow Tufte's principles which highlight key factors for good EDA, including [1]:

1. **Comparisons:** Show data by comparisons to depict contrasts and differences between dependent variables.
2. **Causality:** Demonstrate how the independent variables impact or influence the dependent variables.
3. **Multivariate:** Combine various data so an audience can easily interpret an otherwise complex narrative.
4. **Integration:** Incorporate various modes of information to show evidence of source data-to-findings.
5. **Documentation:** For credibility, include attribution, detailed titles, and measurements.
6. **Context:** Describe or depict the before and after state. Show trend lines to hint at results in the future.

While there are numerous ways to perform EDA, this report will present a selection of visualizations, along with interpretations to highlight key information and trends from the data. Teachings from the Human-centered Data Science book [2] discussed in the course will also be leveraged to enhance visualization.

Methods:

The methods section of this report will be organized by figure and outline the steps taken to prepare the data visualizations, why they were selected, what data was included, and which of Tufte's principles were applied.

Figure 1:

Figure 1 was developed to depict the difference in international migrant stock at mid-year between developed and developing regions. Data was graphed over a timeframe of 1990-2015 for both males and females. This visual was developed as a first step to depict the macro level trends in the dataset. As world data is comprised of developed and developing regions, both were graphed alongside each other for comparison purposes (principle 1). Data was pulled from tidied table 1.

A double horizontal bar graph was used to show trends in the continuous y-variable, migrant stock, over the categorical x-variable, year, and further broken out into the categorical variable, region (principles 2, 3).

Data frame 1 (df1) from A1 was copied and further cleaned by removing any rows with null values in the migrant stock column, as the null values would prohibit sorting of the column. Two data frames were then created using this data. The first data frame was filtered to only include data for developed regions and both sexes, and the second data frame was filtered for developing regions and both sexes. Only the necessary columns of migrant stock and year were kept in the data frames. Both were then merged to create a new data frame. A layered, horizontal bar graph was developed which by default, which was sorted in descending order by developed region, as desired.

Figure 2:

Figure 2 was developed to depict the difference in international migrant stock at mid-year between males and females worldwide. Data was graphed over a timeframe of 1990-2015. This visual was developed to build upon Figure 1, to depict the macro level trends in the dataset by comparing sex rather than region (principle 1). Data was pulled from tidied table 1.

A line graph was used to show trends in the continuous y-variable, migrant stock, over the categorical x-variable, year, and further broken out into the categorical variable, sex (principles 2, 3).

Similar steps from Figure 1 were applied to the code on the cleaned data frame. In this case, two data frames were created to filter by worldwide and males and females. A line graph was developed.

As can be seen, there were not major differences observed between the two sexes. Although this will be further interpreted in the results and discussion section of the report, it is important to mention here, as this resulted in narrowing the focus of subsequent visualizations to compare differences in time and area, rather than sex (principle 1).

Figure 3:

Figure 3 was developed to depict the difference in international migrant stock at mid-year between the six major areas (Oceania, North America, Latin America & The Caribbean, Europe, Africa, Asia). Data was graphed over a timeframe of 1990-2015 for both sexes. This visual was developed to provide a detailed lens of the trends depicted in Figure 1, which was organized by larger regions (principle 1). Data was pulled from tidied table 1.

A multi bar graph was used to show trends in the continuous y-variable, migrant stock, over the categorical x-variable, year, and further broken out into the categorical variable, major area (principles 2, 3).

Similar steps from Figure 1 were applied to the code on the cleaned data frame. In this case, six data frames were created to filter by major area and both sexes. A multi bar graph was developed.

Figure 4:

Figure 4 was developed to depict the difference in international migrant stock at mid-year between the top 15 countries as of 2015. Data was graphed for both sexes. This visual was developed to provide a more recent and detailed lens of the trends depicted in Figure 3, which was organized by major area over time, whereas this visual focused on data from just 2015 and countries within the major areas (principle 1). Data was pulled from tidied table 1.

A horizontal bar graph was used to show trends in the continuous y-variable, migrant stock, over the categorical x-variable, country (principles 2, 3).

Similar steps from Figure 1 were applied to clean the data frame. In this case, one data frame was created to filter by both sexes. The non-country codes were removed, as well as entries from 1990-2010. The data was sorted in descending order by migrant stock and filtered to keep the top 15 countries in the data frame. A horizontal bar graph was developed.

Figure 5:

Figure 5 was developed to depict the difference in international migrant stock at mid-year between the top 5 countries. Data was graphed over a timeframe of 1990-2015 for both sexes. This visual was developed to build upon the trends depicted in Figure 4 but focused on the top 5 rather than the top 15 countries (principle 1). Additionally, trends were observed over a series of time rather than a point in time, as seen in Figure 4. Data was pulled from tidied table 1.

A series of histograms were used to show trends in the continuous y-variable, migrant stock, over the categorical x-variable, year, and further broken out into the categorical variable, country (principles 2, 3).

Similar steps from Figure 1 were applied to clean the data frame. In this case, the data frame was filtered to keep only the top 5 countries, and the year column was kept in the data frame. Data was sorted in ascending order by year and migrant stock, so that when the histograms were developed, trends could be observed as time increased.

Figure 6:

Figure 6 was developed to depict the difference in international migrant stock percentage at mid-year between the top 15 countries as of 2015. Data was graphed for both sexes. This visual was developed to provide a comparison to Figure 4, which looked at just migrant stock rather than migrant stock as a percentage of the total population (principle 1). Data was pulled from table 3.

A horizontal bar graph was used to show trends in the continuous y-variable, migrant stock percentage, over the categorical x-variable, country (principles 2, 3).

Data from table 3 was used. However, in A1, the country code column was removed before tidying the data frame. As that column is now needed, it cannot be appended back to the data frame and match the format of the tidy table. Thus, the original df3 will be used and the same code from A1 will be applied to tidy it. In hindsight, when tidying tables in A1, it would have been better to perform all the data cleaning and then remove any columns at the end, which would make it easier to append later if needed. After this, similar steps from Figure 4 were applied to the code on the cleaned data frame to filter and sort by the top countries. A horizontal bar graph was developed.

Figure 7:

Figure 7 was developed to depict the difference in estimated refugee stock at mid-year between the six major areas. Data was graphed over a timeframe of 1990-2015 for both sexes. This visual was developed to analyze an additional variable, as it focuses on refugees rather than migrants (principles 1, 3). Data was pulled from table 6.

A scatterplot was used to show trends in the continuous y-variable, estimated refugee stock, over the categorical x-variable, year, and further broken out into the categorical variable, major area (principles 2, 3).

Data from table 6 was used. However, for the same reasons as Figure 6, the original df6 with all columns was used and the same code from A1 was run to tidy it. After this, similar steps from Figure 5 were applied to the code on the cleaned data frame to filter for the major areas. A scatterplot was developed.

Figure 8:

Figure 8 was developed as a supplement to Figure 7 to provide a more detailed lens of the trends. Rather than being consolidated in one graph, a series of line graphs were produced to highlight the individual trends by major area over time. The same data and variables from Figure 7 were used. No additional code was required for the data frame.

Figure 9:

Figure 9 was developed to build upon the findings from Figures 7 and 8 to depict the difference in estimated refugee stock at mid-year in Asian regions, as Asia was shown to have

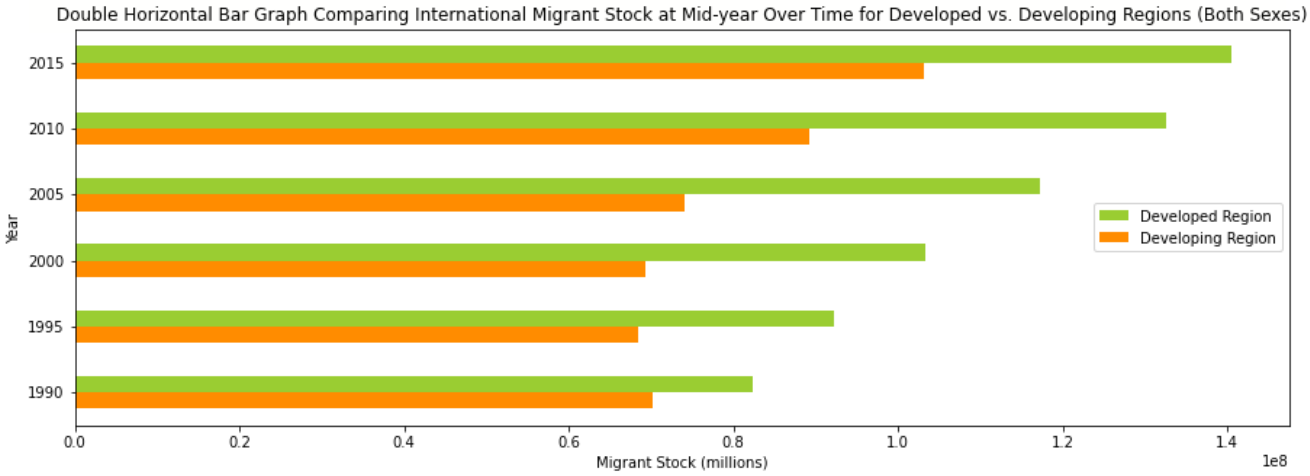
the highest percentage of refugees. Data was graphed over a timeframe of 1990-2015 for both sexes (principle 1). The same data from Figures 7 and 8 was used.

Boxplots were used to show trends in the continuous y-variable, refugee stock, over the categorical x-variable, year (principles 2, 3). Similar steps from Figure 5 were applied to the code to develop the plots.

Results and Discussion:

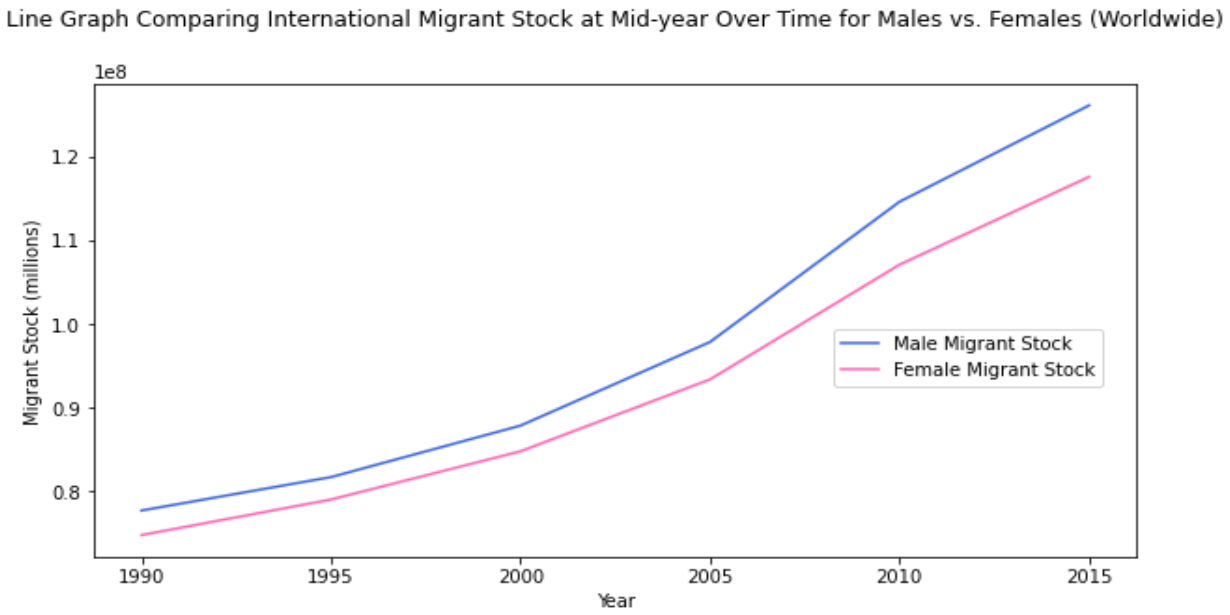
The results and discussion section of this report will also be organized by figure and provide interpretations to highlight key information and trends from each visualization.

Figure 1. Double Horizontal Bar Graph Comparing International Migrant Stock at Mid-year Over Time for Developed vs. Developing Regions (Both Sexes)



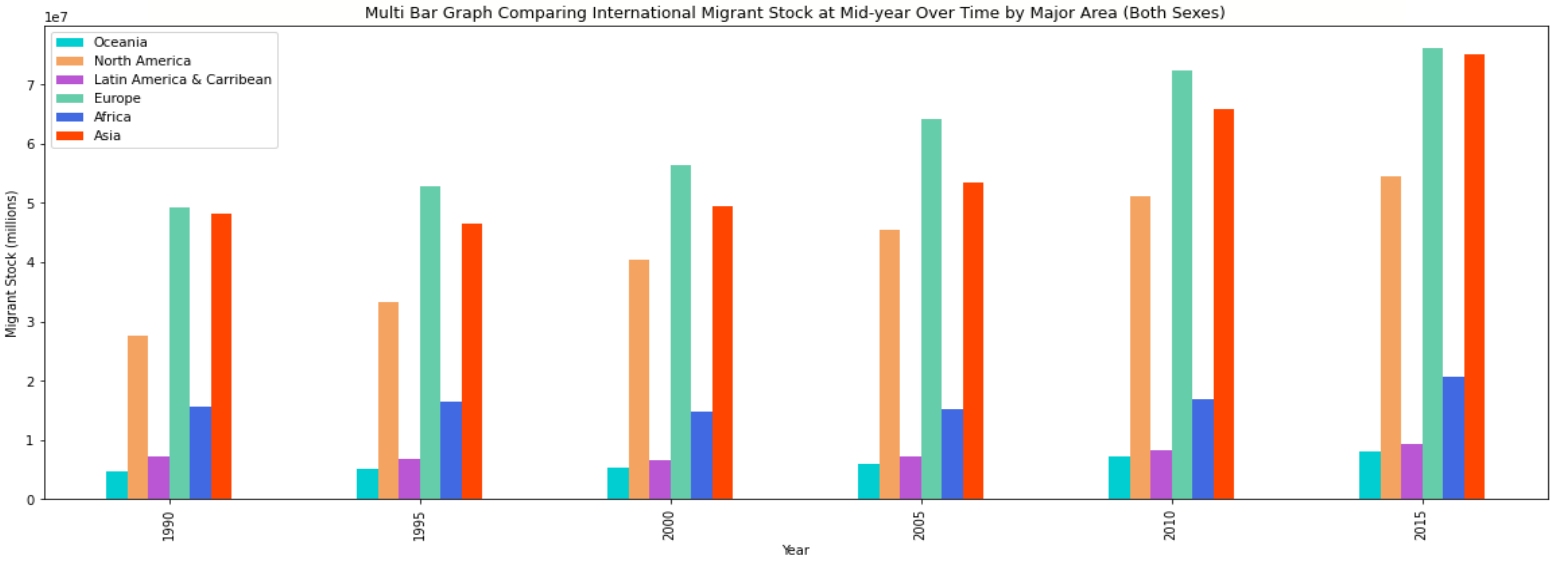
At a macro level, this graph shows that from 1990-2015, there has mostly been a consistent increase in migrants for both developed and developing regions. It also appears that more migrants live in developed regions. While this data only goes up until 2015, this can help to predict that migration will continue to increase over time, with most migrants relocating to developed regions (principle 6). At the micro level, this could be further analyzed to look at what factors, such as work and educational opportunities, and immigration policies, to name a few, draw migrants to developed versus developing regions.

Figure 2. Line Graph Comparing International Migrant Stock at Mid-year Over Time for Males vs. Females (Worldwide)



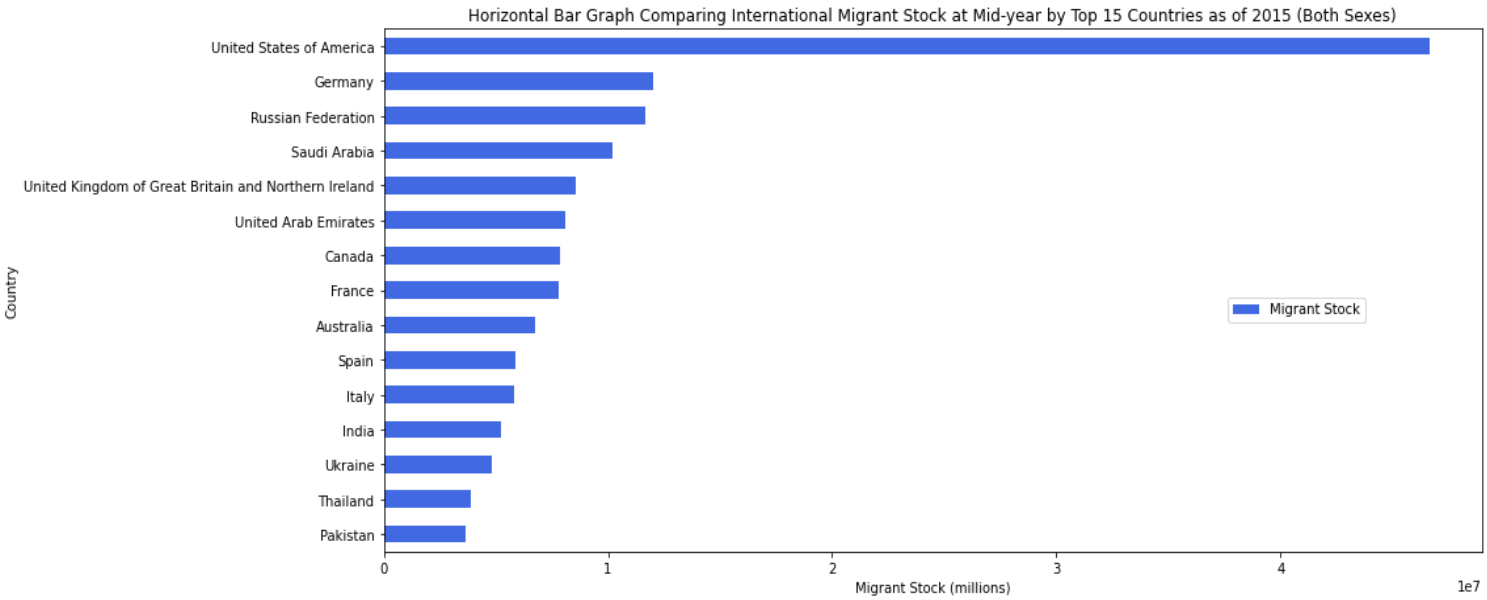
At a macro level, this graph shows that from 1990-2015, there has been a consistent increase in both male and female migrants worldwide. It also appears that more migrants are male, with this gender gap increasing from 2005 onwards. While this data only goes up until 2015, this can help to predict that migrants are slightly more likely to be male than female (principle 6). At the micro level, this could be further analyzed to look at what factors influence more males than females to migrate. For e.g., males may be more likely to migrate to provide for their families in their home countries. Additionally, as a significant sex gap was not observed in migration, the focus of subsequent visualizations will be narrowed to compare differences by time and area, rather than sex (principle 1).

Figure 3. Multi Bar Graph Comparing International Migrant Stock at Mid-year Over Time by Major Area (Both Sexes)



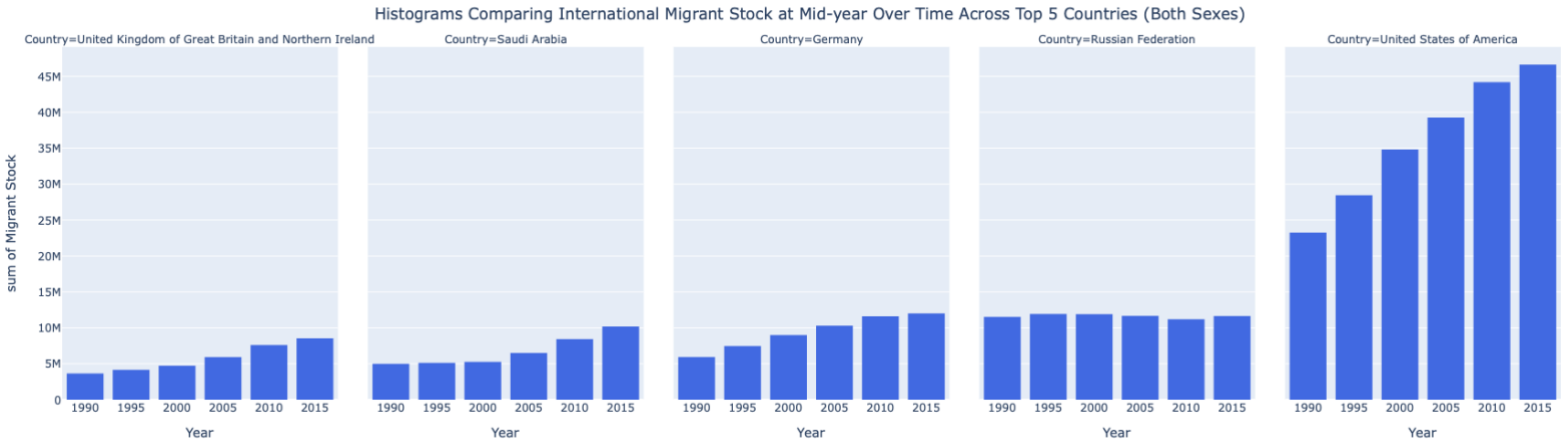
This graph provides a more detailed look into the findings from Figure 1, which looked at migration by region. From 1990-2015, there has been a consistent trend that more people migrate to North America, Europe, and Asia. It also appears that over time, migration levels have remained relatively low in Oceania, Latin America and The Caribbean, and Africa. Additionally, it appears that migration levels in Asia have caught up to those of Europe as of 2015. This data can help to predict that migrants will continue to prefer relocating to North America, Europe, and Asia (principle 6). At the micro level, this could be further analyzed to look at what factors have caused North American migration to grow slower than those of Europe and Asia. For e.g., by looking at the countries that border these major areas and how many people are migrating over. Trends such as migrant workers can also be explored, as the majority work in Arab countries [3], which may have contributed to the increase in migration in Asia. Additionally, government bodies in Oceania, Latin America and The Caribbean, and Africa can look to areas with higher migration to replicate factors that draw more migration.

Figure 4. Horizontal Bar Graph Comparing International Migrant Stock at Mid-year by Top 15 Countries as of 2015 (Both Sexes)



This graph provides a more detailed look into the findings from Figure 3, as it looks at top migration population by country as of most recent data in 2015. The United States of America (USA) has significantly higher migration than any other country. Additionally, building upon the findings from Figure 3 which showed that most people migrate to North America, Europe, and Asia, 14 of the top 15 countries above are in one of the mentioned major areas, further supporting the findings from Figure 3. Like previous recommendations, at a micro level, this could be further analyzed by looking at trends such as bordering country migration, work and educational opportunities, and migrant labour.

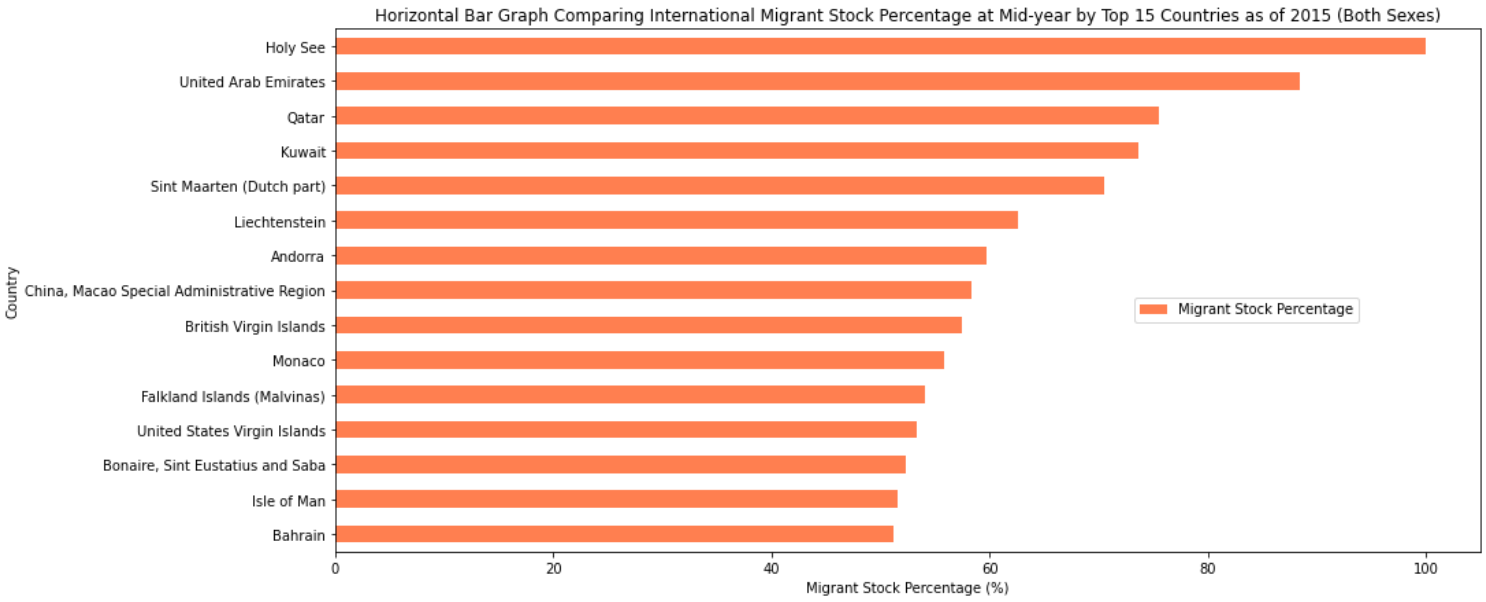
Figure 5. Histograms Comparing International Migrant Stock at Mid-year Over Time Across Top 5 Countries (Both Sexes)



Note: For a larger visual, please refer to the graph output in the code.

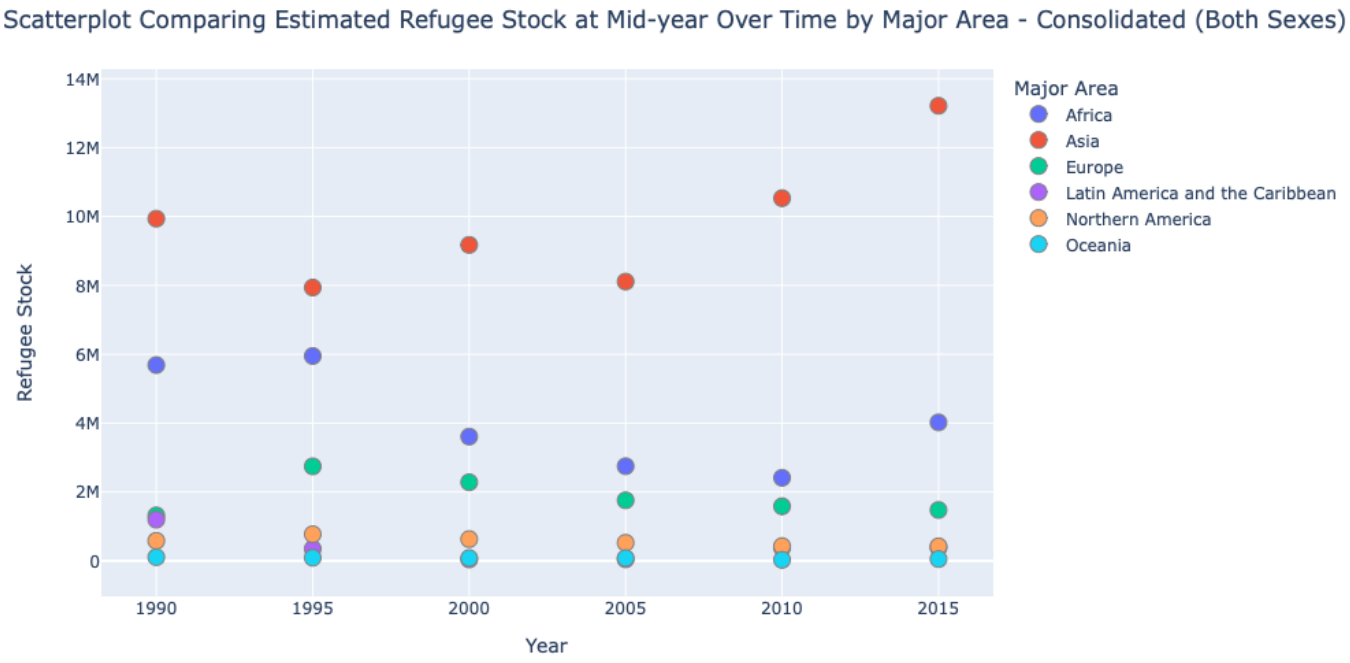
This graph provides a more detailed look into the findings from Figure 4, as it further breaks down the top 15 countries as of 2015 into the top 5 and looks at trends over 1990-2015. The USA has had a steady increase in migration over time with much higher levels than any other country, while the United Kingdom, Saudi Arabia, and Germany have had similar patterns of growth. At the micro level, this data could be further analyzed to look at what factors have caused Russian migration to remain stagnant.

Figure 6. Horizontal Bar Graph Comparing International Migrant Stock Percentage at Mid-year by Top 15 Countries as of 2015 (Both Sexes)



This graph provides an alternate view to Figure 4 by looking at migrant stock as a percentage of the total population, rather than just the number of migrants. This is important to consider, as while more populated countries may have higher migrant stock, the impacts of migration may be more pronounced in less populated countries, as shifts in population will be more drastic. The top 15 countries by migrant stock percentage appear to be either small countries or Arab countries, which further supports findings from Figure 3, which shows high levels of migration in Asia. Like previous recommendations, at a micro level, this could be further analyzed by looking at trends such as migrant labour, as well as the need for migration to boost small ethnic populations observed in some Arab countries [3].

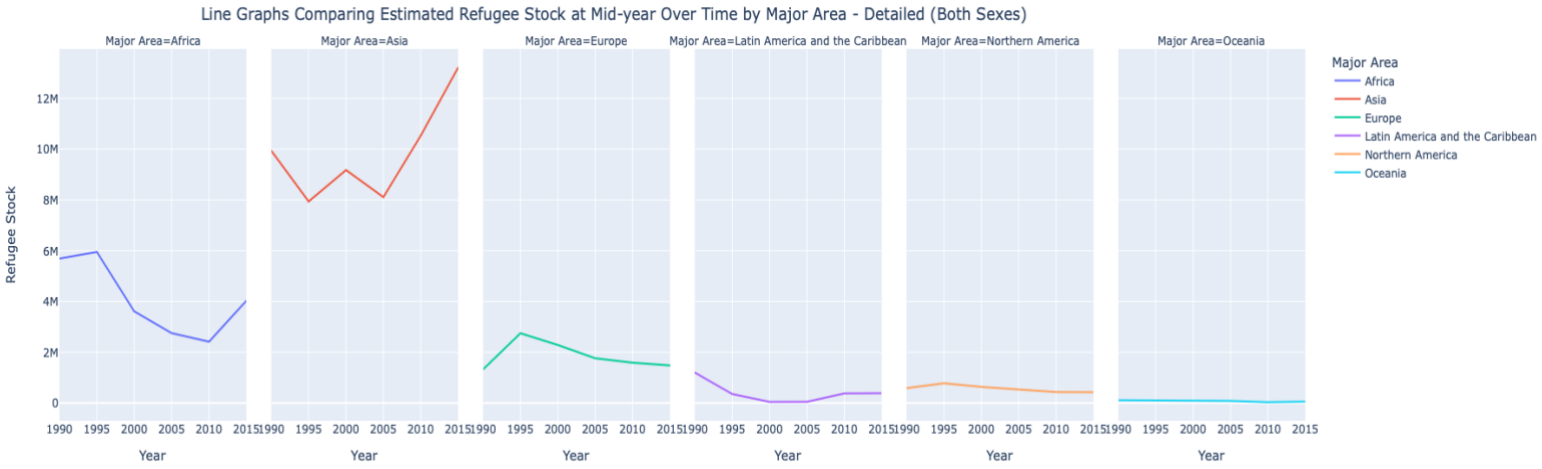
Figure 7. Scatterplot Comparing Estimated Refugee Stock at Mid-year Over Time by Major Area - Consolidated (Both Sexes)



This graph introduces an additional variable into the analysis, which is refugee stock. It is important to consider that refugees are also forms of migrants. This can provide an alternate view to Figure 3, which looks at migrant stock by major area. This will enable comparison of if refugees and migrants display similar patterns across the major areas. From 1990-2015, there has been a consistent trend that more refugees relocate to Asia, Africa, and Europe. Figure 3

showed that more migrants relocate to North America, Europe, and Asia, so two of the major areas, Europe and Asia, appear to be popular locations for both migrants and refugees. It also appears that over time, migrant levels have remained relatively low in Oceania, North America, and Latin America and The Caribbean. Comparing to Figure 3, both migrant and refugee levels remain low in two of the major areas, Oceania and Latin America and The Caribbean. This data can help government bodies predict the areas that both migrants and refugees will continue to relocate to which can be leveraged for growth planning (principle 6).

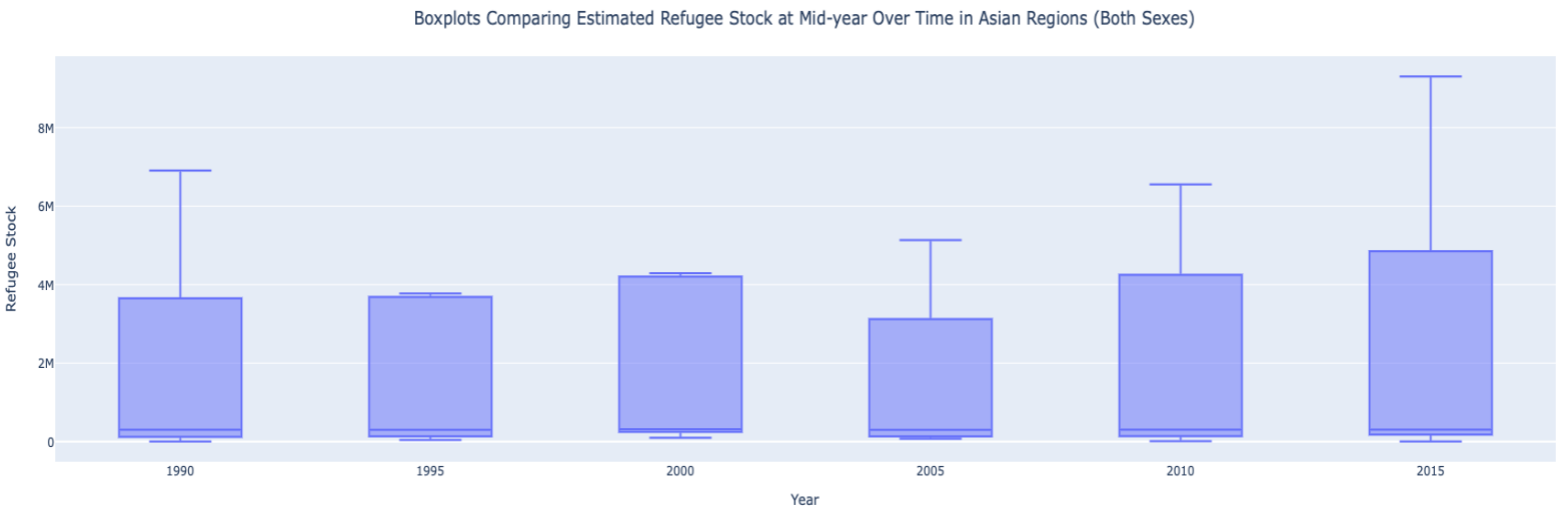
Figure 8. Line Graphs Comparing Estimated Refugee Stock at Mid-year Over Time by Major Area - Detailed (Both Sexes)



Note: For a larger visual, please refer to the graph output in the code.

This graph provides a different and more detailed way to look at the trends observed in Figure 7, by looking at the individual trends of refugee stock across time in each major area. The same findings from Figure 7 apply.

Figure 9. Boxplots Comparing Estimated Refugee Stock at Mid-year Over Time in Asian Regions (Both Sexes)



Note: For a larger visual, please refer to the graph output in the code.

This graph builds upon the findings from Figures 7 and 8, which showed that most refugees relocate to Asia. It looks broadly at the spread of the number of refugees across the 5 Asian regions (central, eastern, south-eastern, southern, western) over time. In 2005, the levels of refugees were the most similar across most regions and differed the most in 2015. Additionally, from 1995 onwards, the highest level of refugees across the regions consistently increased. At the micro level, this data could be further analyzed to look at refugee stock prior to 1990 to observe if there was a random spike in 1990 in the number of refugees across at least one region, or if this was consistent with previous trends, as well as a more detailed look at which regions have the highest and lowest rates of refugees.

Other Considerations:

While the application of Tufte's principles has been noted throughout the methods and results and discussion sections of the report, it should be noted that principles 4 and 5 were applied throughout all the visualizations. Principle 4, integration, can be seen throughout the variety of visuals presented and findings that were drawn from the data [1]. Principle 5, documentation, can be seen throughout the labelling of the visuals [1].

Visuals were created to be minimalistic, yet impactful, and to build upon and further interrogate the findings from each previous graph, as suggested in the Human-centered Data Science book [2].

Tables 2, 4, and 5 from the UN dataset were not explicitly visualized, however, may have been leveraged for other visualizations. Table 2 looked at the total population, which was decided was not as meaningful as looking at refugee or migrant populations for the scope of this report. However, the data from Table 2 was used for calculations in Table 3 to show refugee stock as a percentage of the population, which was visualized. Table 4 looked at female migrant stock, which as mentioned in the report, was already addressed in Figure 2. As significant differences between sexes were not shown, the focus of subsequent visualizations was narrowed to compare differences by time and area, rather than sex. Table 5 looked at the rate of change of migrant stock, however, as several visuals depicted change in migrant stock over time, it was decided that these trends were sufficiently covered for the scope of this report.

Conclusion:

Having performed EDA through visualization of the tidied UN dataset, it is clear that there are numerous ways to apply Tufte's principles and in general, principles for good visualization to interrogate cleaned data. The process outlined above is just one proposed solution.

As next steps, several recommendations have been proposed throughout the results and discussion section of the report to further analyze and understand the trends presented in the data, which can support the development of more meaningful research questions, and subsequently, better data analysis.

References

[1] "Tufte's 6 Principles for Graphical Integrity (Adopted for Service Design)," *International Service Design Institute*, 03-Sep-2020. [Online]. Available: https://internationalservicedesigninstitute.com/tuftes-6-principles-graphical-integrity-adopted-service-design/. [Accessed: 11-Dec-2022].

[2] C. R. Aragon, S. Guha, M. Kogan, M. Muller, and G. Neff, *Human-centered Data Science: An introduction*. Cambridge, MA: The MIT Press, 2022.

[3] "Labour migration in Asia and the Pacific (ILO in Asia and the Pacific)," *International Labour Organization*. [Online]. Available: https://www.ilo.org/asia/areas/labour-migration/lang--en/index.htm. [Accessed: 14-Dec-2022].