Descriptive Statistics and Data Visualization

Name: Emily Clarke

Introduction

"A model is only as good as its representation and how well it can be communicated to the people who will use it" (Aragon et. al, 2022, pg. 69). This describes data visualization, which represents data in the form of graphs. On the surface level, it is simply a graph, though it is much more than this. The creation of data visualizations is computer generated, but the data is analyzed by people. Therefore, the aim of this report is to create data visualizations and analyze them to comply with human centered data science.

Within the scope of this project, we will be analyzing a large dataset called the *UN Trends in International Migrant Stock: The 2015 Revision*. This dataset consists of six main tables. As I create and analyze the visual findings, I will be looking through the lens of Tukey's exploratory data analysis (EDA) principles and Tufte's visualization principles. Tukey has three key EDA principles, which include sorting the data, grouping the data, and subsetting the data into smaller datasets. Tufte has two core visualization concepts, which include chart junk which emphasizes the importance of minimalism within graphs and small multiples which take subparts of a dataset and split into their own graph. Arguably, small multiples adheres to Tukey's EDA principles, which is why this report will primarily be looking through the lens of Tufte's two visualization principles.

Methods

The methods that will be utilized will revolve around visualizations and descriptive statistics. The visualizations provide graphic representation of the data and information from the UN dataset. The graphic visualizations chosen to portray the data include histograms and boxplots. Histograms were a chosen graphic tool to measure interval level data, as this was the most convenient tool to widely distribute a large scale of continuous data. Histograms are often confused with bar charts, it is necessary to define histograms as graphs which summarize quantitative/numerical data. Boxplot was chosen as a graphing tool because it is a chart that provides a descriptive statistical summary of a set of data. The descriptive statistics that this boxplot includes is the minimum score, lower quartile, median, upper quartile, and maximum score. The method of analysis that will be used is multivariate analysis, where multiple variables

are compiled and analyzed. The variables that will mainly be used include the International Migrant Stock (varies by Table in the UN dataset) interval data, sex, and year.

Results and Visualizations

Figure 1

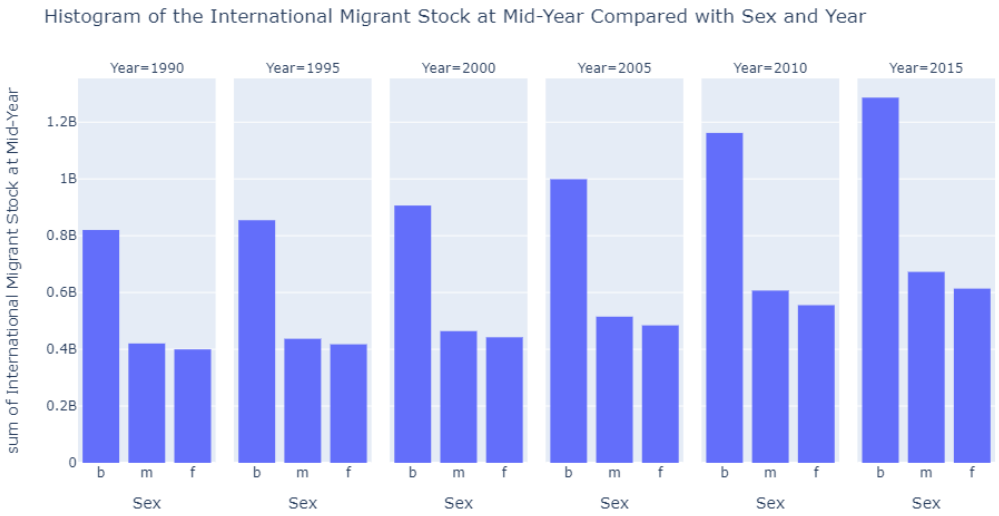


Figure 2

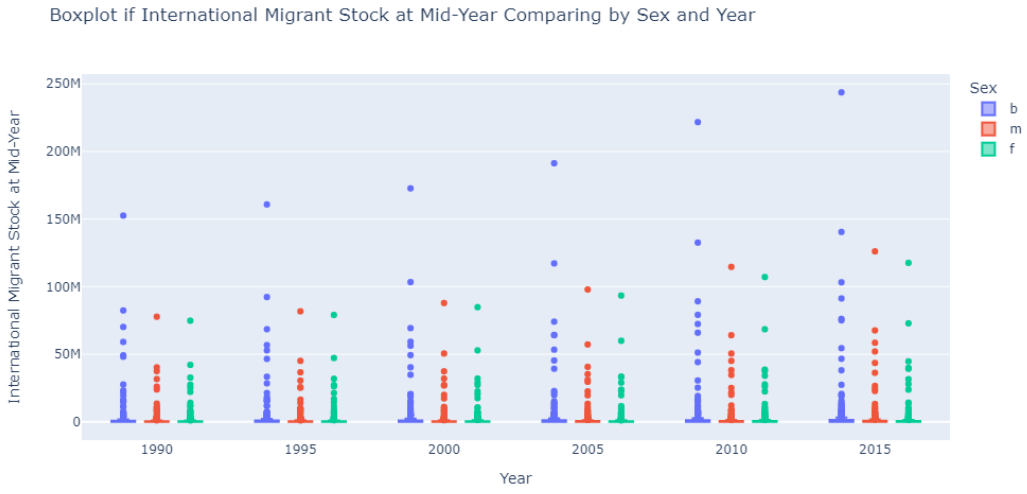


Figure 3

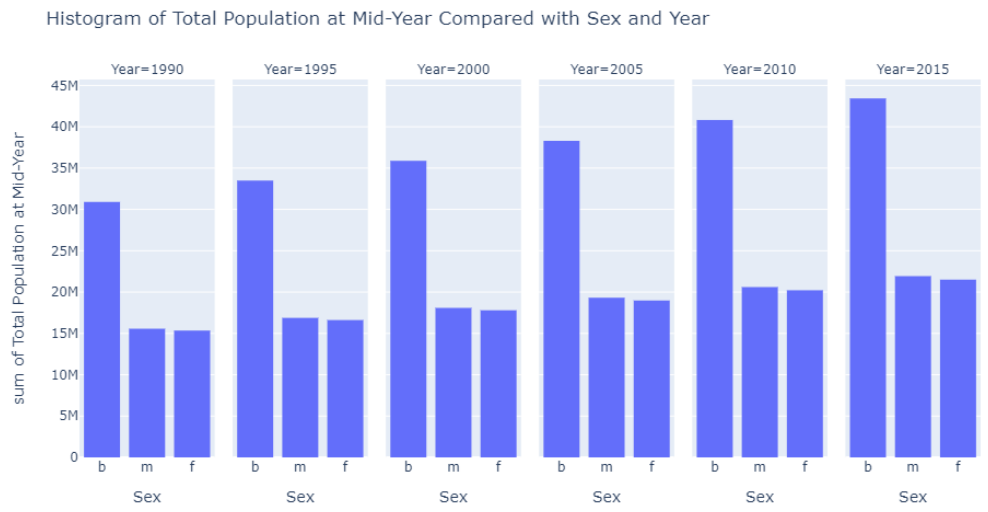


Figure 4

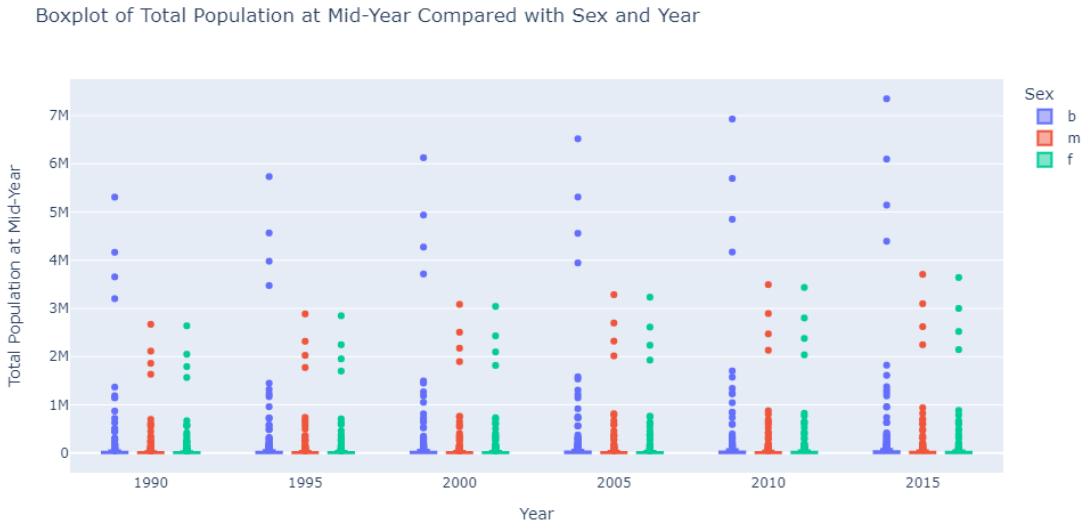


Figure 5

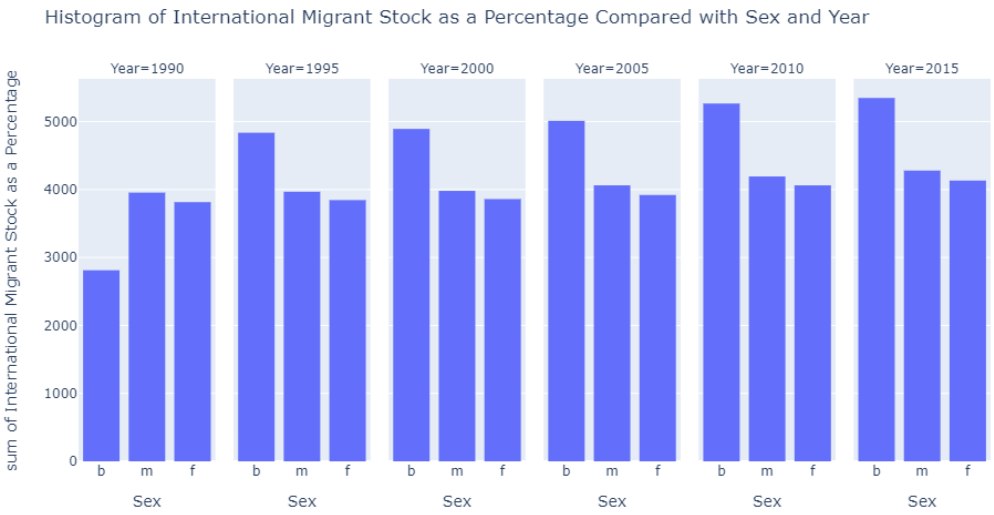


Figure 6

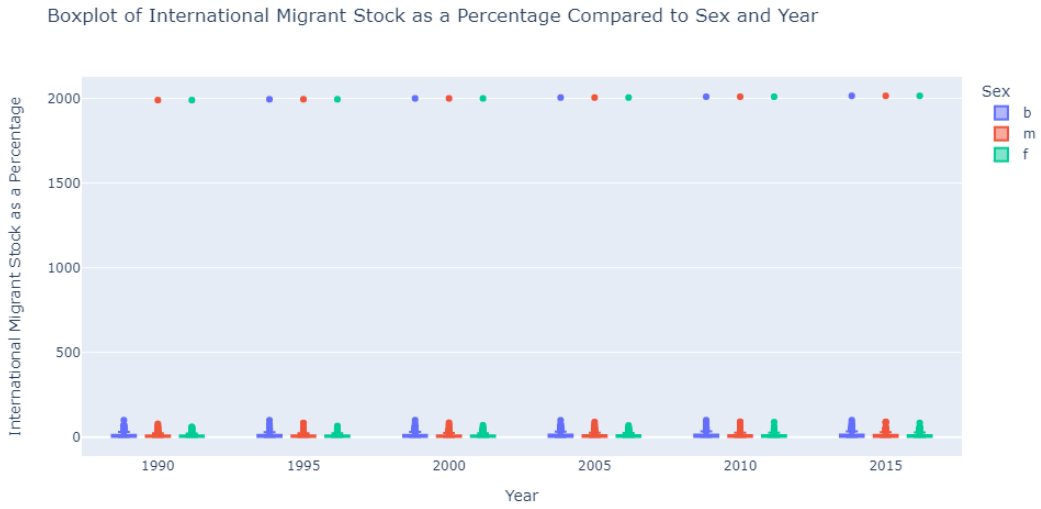


Figure 7

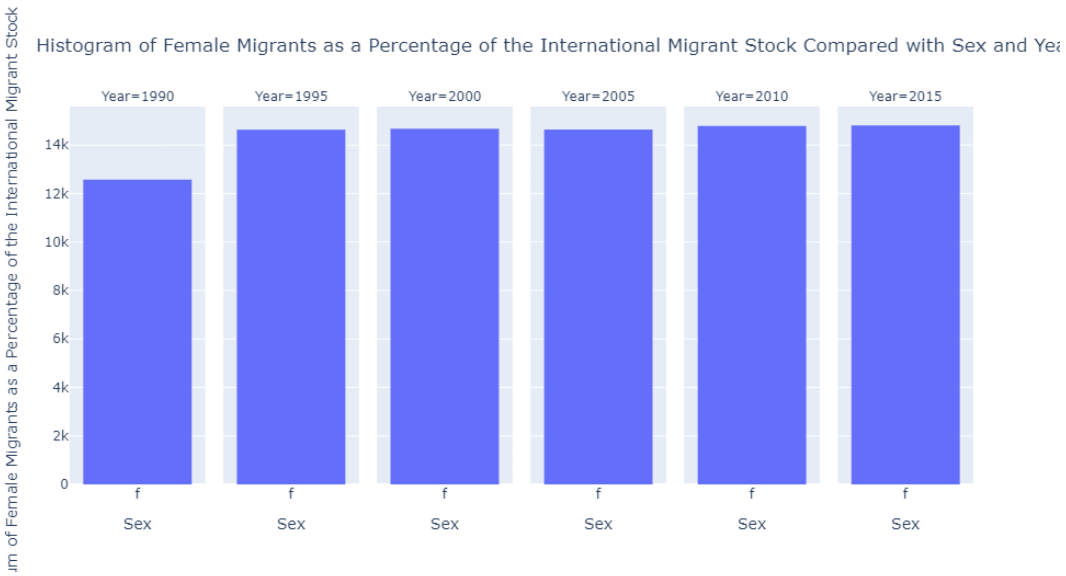


Figure 8

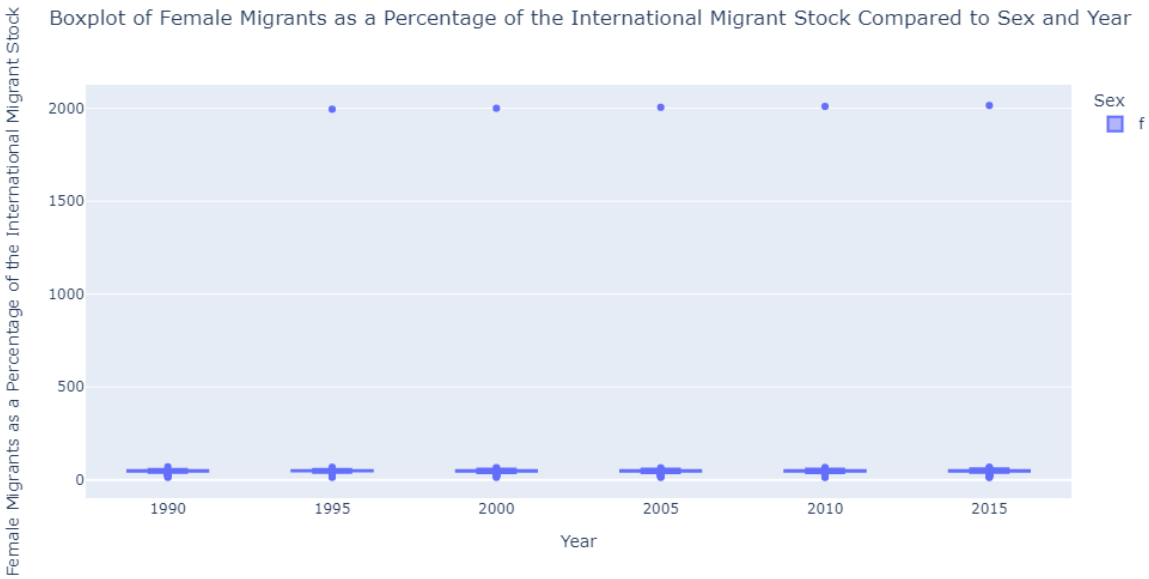


Figure 9

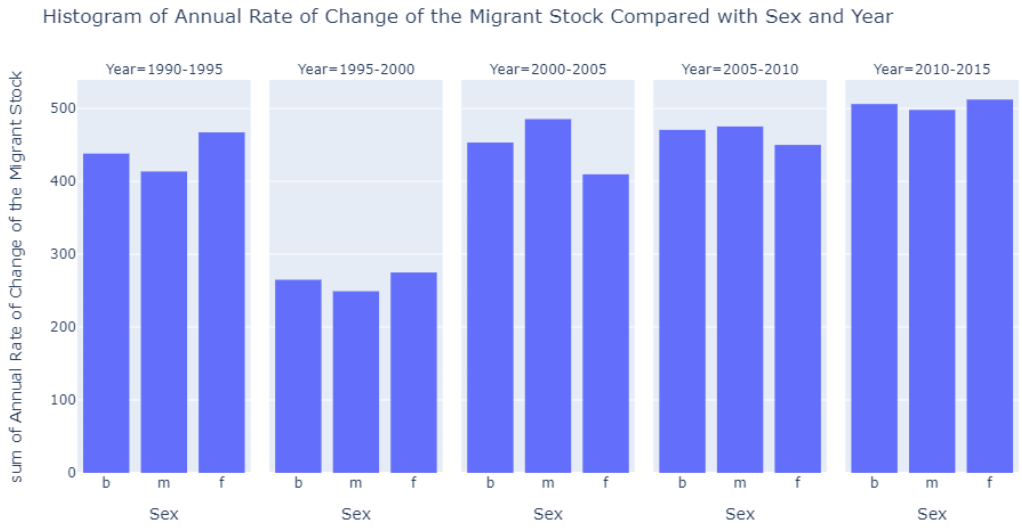
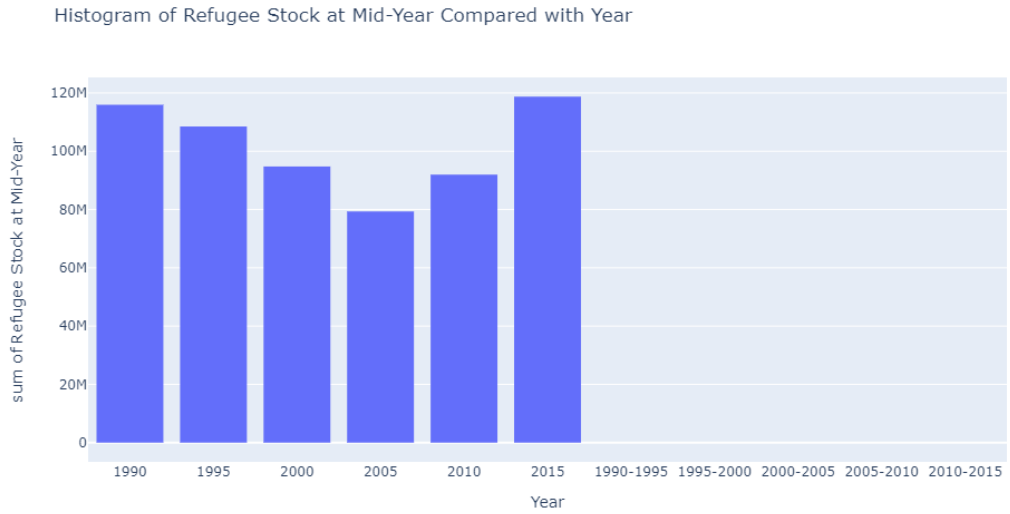


Figure 10



Discussion

Figure 1 is a histogram, looking at variables 'Sex' and 'Year' as compared to the 'International Migrant Stock at Mid-Year' from Table 1 of the UN Dataset. Migrant stock refers to "the number of people born in a country other than that in which they live, including refugees" (The World Bank, 2022). The first variable, 'Year=1990', displays the sum of both sexes at 820.0771 million, male at 420.4716 million, and female at 399.6128 million. The last variable, 'Year=2015", displays the sum of both sexes at 1.286222 billion, male at 672.5891 million, and female at 613.6396 million. The first and last 'Years' will be analyzed to identify the trends of

the graph as there was a continuous increase. There is a positive relationship between year and the total population considering there was a gradual increase in the migrant stock throughout the years. It would be correct to say as years increase, migrant stock population increases. One notable difference between the sexes is that there are more male migrant stock populations than females.

Figure 2 is a boxplot, analyzing variables 'Sex' and 'Year' as compared to the 'International Migrant Stock at Mid-Year' from Table 1 of the UN Dataset. In the first 'Year=1990', the descriptive statistics of both sexes indicate that the minimum=270, quartile 1 (q1) =26.4 thousand, median=133.545 thousand, quartile 3 (q3)=836.608 thousand, and maximum=152.5632 million. In 1990, men indicated minimum=150, q1=14.261 thousand, median=69.66 thousand, q3=426.147 thousand, and maximum=77.7475. In 1990, females indicated minimum=120, q1=12.332 thousand, median=60.7385 thousand, q3=405.682 thousand, and maximum=74.8157 million. In 'Year=2015', both sexes displayed statistics of minimum=141, q1=36.114 thousand, median=213.51 thousand, q3=1.3879 million, and maximum=243.700 million. Males showed a minimum=78, q1=17.136 thousand, median=102.608 thousand, q3=715.142 thousand, maximum=126.115 million. Females showed a minimum=63, q1=18.204 thousand, median=109.324 thousand, q3=625.165 thousand, maximum=117.584 million. When comparing the latest year (1990) and the most recent year (2015), it is clear that the numerical data shows an increase in the migrant stock.

Figure 3 is a histogram, analyzing variables 'Sex' and 'Year' as compared to the 'Total Population at Mid-Year' from Table 2 of the UN Dataset. 'Year=1990' highlights the sum of both sexes=30.904 million, men=15.565 million, and female=15.432 million. 'Year=2015' displays the sum of both sexes=43.428 million, men=21.927 million, and female=21.502 million. As the numbers suggest, there is a gradual increase in the total population throughout the years (beginning at 1990 to 2015). This displays a positive relationship between years and total population, which means that as years increase the population increases. The male population is slightly larger than the female population, and remains larger throughout the years.

Figure 4 is a boxplot, analyzing variables 'Sex' and 'Year' as compared to the 'Total Population at Mid-Year' from Table 2 of the UN Dataset. In 'Year=1990', both sexes indicated a min=0.77, q1=387.877, median=4,986.705, q3=22.231 thousand, max=5.309 million. Males indicate a min=29.843, q1=776.227, median=3,468.319, q3=16.047 thousand, and max=2.670

million. Females show a min=31.492, q1=803.467, median=3,494.221, q3=16.682 thousand, and max=2.639 million. In 'Year=2015', both sexes show a min=0.8, q1=583.591, median=7,218.885, q3=35.939 thousand, and max=7.349 million. Males display a min=43.889, q1=1,056.775, median=5,041.793, q3=21.2447 thousand, and max=3.707 million. Females show a min=47.605, q1=1,066.513, median=5,223.225, q3=22.172 thousand, max=3.642 million. The results from the boxplot indicate that the year 1990 has smaller values than the year 2015, which indicates that the total population increases throughout the years. Despite the findings for Figure 3, these findings show that females in 2015 have slightly higher values in the boxplot data than the values for men in 2015. The only lower value is the maximum value.

Figure 5 is a histogram, analyzing 'Sex' and 'Year' as compared to the 'International Migrant Stock as a Percentage' from Table 3 of the UN Dataset. 'Year=1990' indicates the sum of both sexes=2,811.88, males=3,954.649, and females=3814.335. 'Year=2015' indicates the sum of both sexes=5,347.518, males=4,279.136, and females=4,131.802. In the year 1990, the population as a percentage started as the lowest population. With time, the population of both males and females has increased, until 2015, where the population is the largest.

Figure 6 is a boxplot, illustrating 'Sex' and 'Year' as compared to the 'International Migrant Stock as a Percentage' from Table 3 of the UN Dataset. In 1990, both sexes show min=0.0325, q1=1.583, median=4.558, q3=13.286, max=100; males show min=0.032, q1=1.383, median=3.949, q3=9.924, and max=100; and females show min=0.032, q1=1.253, median=3.596, q3=9.452, and max=100. In 2015, both sexes show min=0.071, q1=1.457, median=4.767, q3=15.248, and max=100; males show min=0.084, q1=1.357, median=3.722, q3=12.620, and max=100; and females show min=0.056, q1=1.252, median=3.685, q3=12.051, and max=100. The values from 1995 are lower than 2015, as expected. Certainly, the female values in 2015 are lower than males.

Figure 7 is a histogram, illustrating 'Sex' and 'Year' as compared to 'Female Migrants as a Percentage of the International Migrant Stock' from Table 4 of the UN Dataset. Figure 7 is an in-depth look at female migrant stock trends. It begins in 'Year=1990', which reports the sum of females as 12.573 thousand. This is the overall lowest sum in Figure 7, and increases over time. Around 2005, the migrant stock of females slightly declined, with a sum of 14.640 thousand. After this dip, the female percentages increased to sit at 14.811 thousand in 2015, which is the largest female group.

Figure 8 is a boxplot, illustrating ‘Sex’ and ‘Year’ as compared to ‘Female Migrants as a Percentage of the International Migrant Stock’ from Table 4 of the UN Dataset. In 1990, females showed a min=13.856, q1=45.991, median=48.943, q3=51.323, max=70.703. In 2005, females illustrated a min=13.638, q1=45.880, median=49.100, q3=52.126, and max=n/a. In 2015, females showed a min=13.325, q1=45.861, median=49.427, q3=52.090, max=n/a.

Figure 9 is a histogram, depicting ‘Sex’ and ‘Year’ as compared to ‘Annual Rate of Change of the Migrant Stock’ from Table 5 in the UN Dataset. The year 1990-1995 begins at a large rate of change of the migrant stock, as it illustrates both sexes at 437.960, males at 413.454, and females at 467.121. There is a relatively large decrease that occurs in ‘Year=1995-2000’, which leaves the sum of both sexes at 264.876, males at 249.140, and females at 274.905. Within the years 1995-2000, the rate of change of the migrant stock hit its lowest level, which means that there were significantly less refugees. After this decline, there was an expansion in the rate of change of the migrant stock, as the ‘Year=2010-2015’ shows the peak number of sums, as both sexes were at 505.956, males at 497.956, and females at 512.125.

Figure 10 is a histogram, depicting ‘Year’ as compared to ‘Refugee Stock at Mid-Year’. From the visualization, it is apparent that the dataset was not properly cleaned in regards to Table 6 of the UN Dataset, which leaves some discrepancies in the visualization. The tables that are illustrated will still be explained. In ‘1990’, the refugee stock was 115.942 million. Throughout the years, there is a gradual decline, which hits the lowest level in ‘2005’ at 79.334 million. From there, there is an increase in the refugee migrant stock, which hit an overall high in 2015 at 118.732 million.

References

Aragon, Cecilia; Guha, Shion; Kogan, Marina; Muller, Michael; Neff, Gina. (2022). Human Centered Data Science. The MIT Press.

The World Bank. (2022). Metadata Glossary. https://databank.worldbank.org/metadataglossary/jobs/series/SM.POP.TOTL.ZS#:~:text=Short%20definition,It%20also%20includes%20refugees.