Bryan Zhu
1004892342
INF1340 Final Writeup

Exploratory Data Analysis Using a United Nations Dataset

Introduction:

This final project was a continuation of the midterm project, where we were given a dataset from the United Nations consisting of trends in international migrant stock. This UN dataset consists of six tables in an Excel sheet, with additional tables describing notes and an annex. The six tables were generally similar in structure although they measured different variables. The UN dataset first had to be cleaned by following the principles of tidy data since it violated those principles in multiple manners. From there, we can begin to conduct our exploratory data analysis.

In order to create our visualizations, this exploratory data analysis will be guided by Tufte's visualization principles. The first principle involves sorting; this means that the data needs to be organized in a way that's easy and intuitive to understand. The second principle is grouping; this is the idea of grouping certain data into logical categories when applicable. The third principle involves the concept of subsets, in that larger data groups can be broken down into smaller constituent parts to make them easier to comprehend. The fourth principle is comparison; this means showing aspects of data side by side with each other to demonstrate similarities and contrasts. This report will present a series of visualizations that follow very similar structure in accordance with the similarities seen between the tables in the UN dataset.

Methods:

The first thing that can be done with the first data table is a histogram can be constructed to observe the sex, year, and international migrant stock variables simultaneously.

```
[ ]  import plotly.express as px

[ ]  px.histogram(data_frame = table1, x = 'Sex', y = 'Migrant Stock', facet_col = 'Year'
                  , title="International Migrant Stock per Year at Mid-Year")
     #making histogram to see migrant stock for for each sex
```

This follows the first and fourth of Tufte's principles, in that it is sorted to visualize these three variables simultaneously, and the bars on the histograms are shown next to one another to provide visual comparison between categories.

Next, a construction of a visual that allows us to see some comparisons in migrant stock between regions would be useful. The first thing to do is to focus on the major regions (Africa, Asia, etc.).

```
[ ]  major_region = (903, 904, 905, 908, 909, 935)
     table1m = table1[table1['Country Code'].isin(major_region)]
     table1m
     #isolating major regions
```

This follows the second of Tufte's principles. Rather than attempting to visualize every single region provided in the dataset at the same time, they're grouped together into 6 major regions to better be able to see comparisons between a fewer number of variables.

Both sexes will also be grouped together in order to get a complete overview of the entire region as a whole.

```
[ ]  both_sexes = ['both']
     table1ms = table1m[table1m['Sex'].isin(both_sexes)]
     table1ms.head()
     #selecting only "both sexes" to get total migrant stock for eeach region
```

This again follows Tufte's second principle since the sexes are being grouped into one wholistic category.

Finally, we can construct a line plot consisting of the migrant stock of these major regions to observe changes over time.

```
[ ] px.line(data_frame = table1ms, x = 'Year', y = 'Migrant Stock', color = 'Country/Region/Area', labels = {"Country/Region/Area": "Region"}
              , title="International Migrant Stock at Mid-Year for each Major Region")
    #line plot to see trends in migrant stock for each region
```

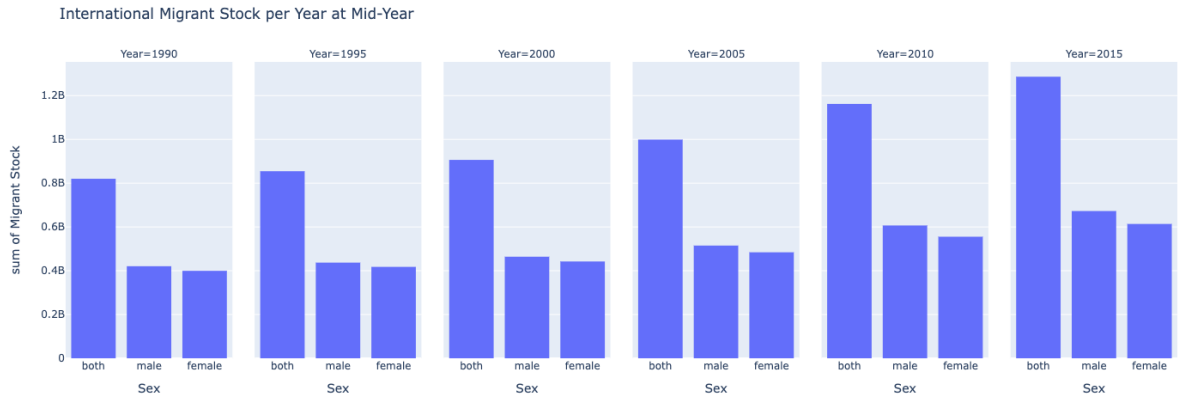
This line plot will align with Tufte's fourth principle to not only provide comparisons between the major regions with each other but also to see how the migrant stock varies from year to year, and also to be able to visually observe any trends.

Since tables 1, 2, 3, and 5 are very similar to each other structurally, the same principles above can be applied to each of these tables, which will be shown later.

Table 4 is slightly different than the other tables. Instead of containing both sexes, this table just contains migrant data for females. We can construct a box plot instead of a line plot to show variations in deviations in the data between each major region. Depending on perspective, this can either be in accordance with Tufte's second or third principle. The categories are there to be of assistance in better comprehension.

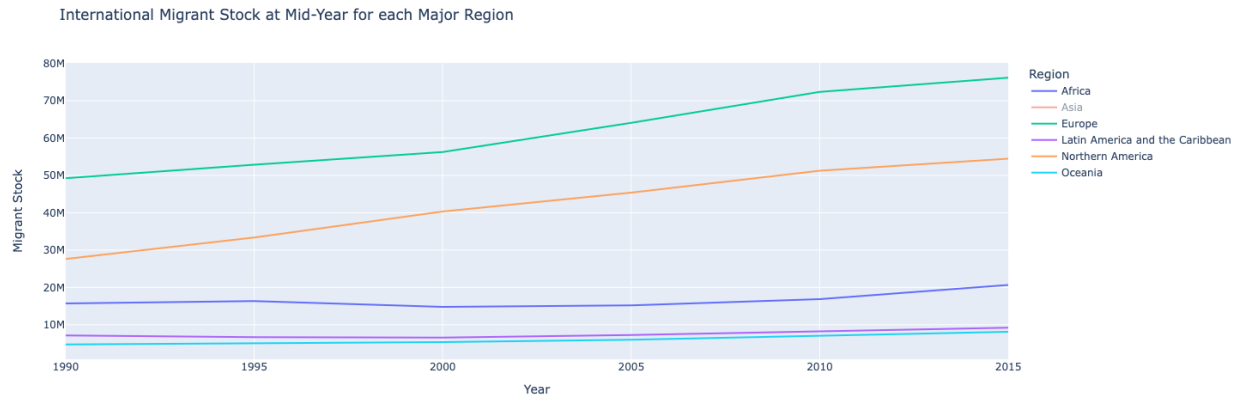
Results:

Figure 1:



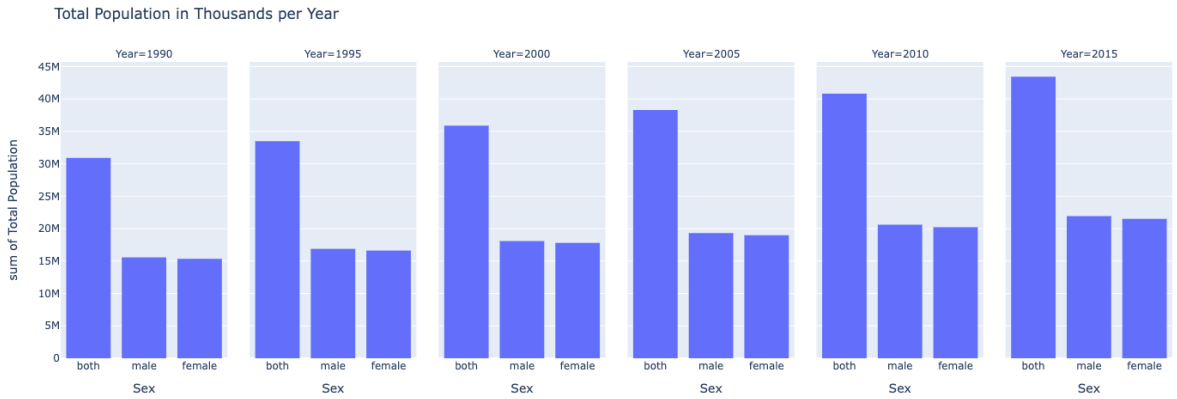
This figure from table 1 shows that the total international migrant stock globally has been steadily increasing from 1990 to 2015. Each year there is a slightly greater proportion of male migrants, but generally the male-female ratio isn't outrageously skewed.

Figure 2:



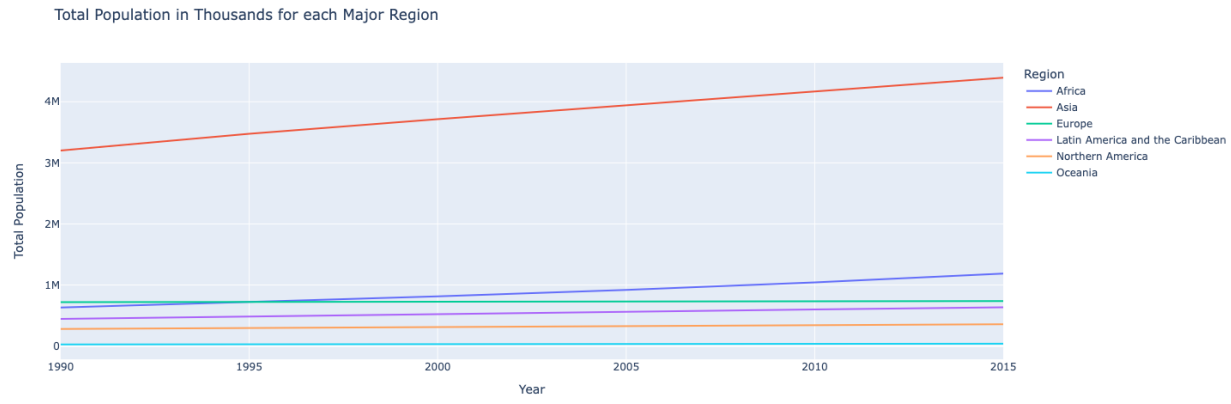
This figure from table 1 shows the total international migrant stock for each major region, with the labels on the right. Europe and North America have the greatest migrant stock, and also have the greatest rate of increase of migrant stock, as shown by the fact that they appear to have greater slopes in comparison to the other four major regions.

Figure 3:



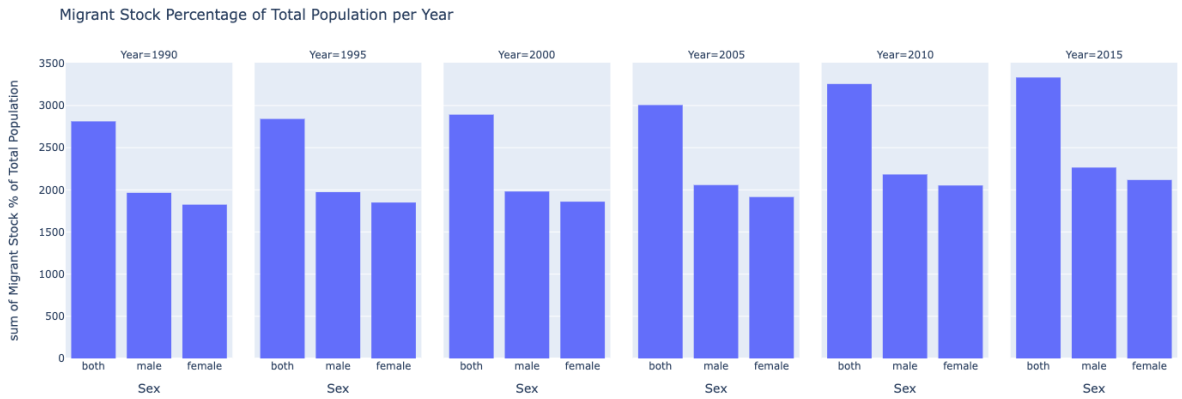
This figure from table 2 shows the total population in thousands compared to year and sex. As one might expect, the total population has been increasing steadily from 1990 to 2015. The proportion of males to females in the population remains relatively consistently equal.

Figure 4:



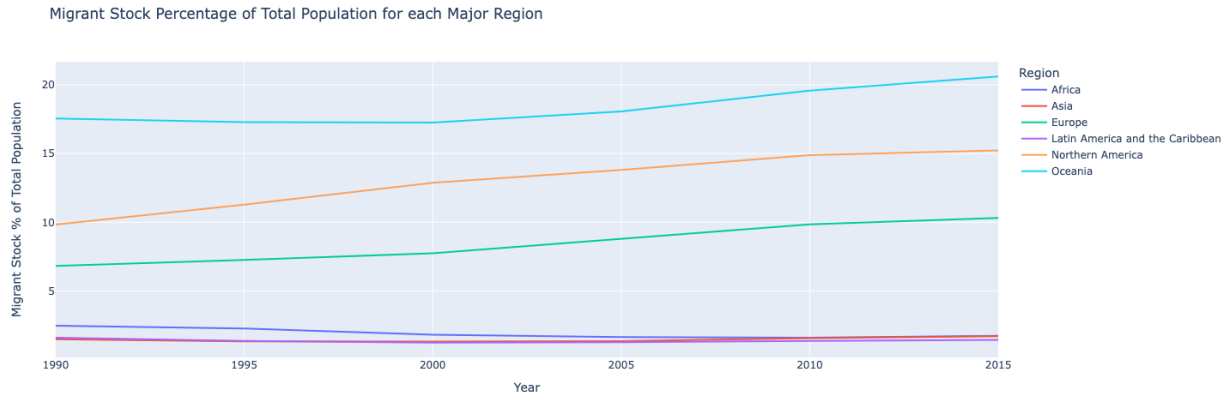
This figure from table 2 portrays total population for each of the major regions. While each of the major regions has had increases in total population, Asia's population is noticeably much higher than that of any of any other region, and is also increasing at the greatest rate.

Figure 5:



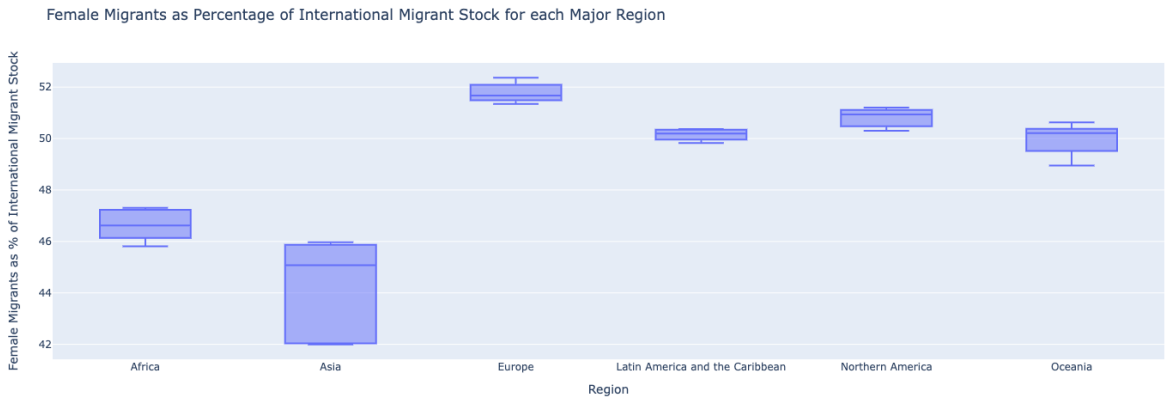
This figure from table 3 shows the migrant stock as a percentage of the total population. There has been a steady increase in this, meaning that globally migrants are becoming more prevalent into within nations' populations.

Figure 6:



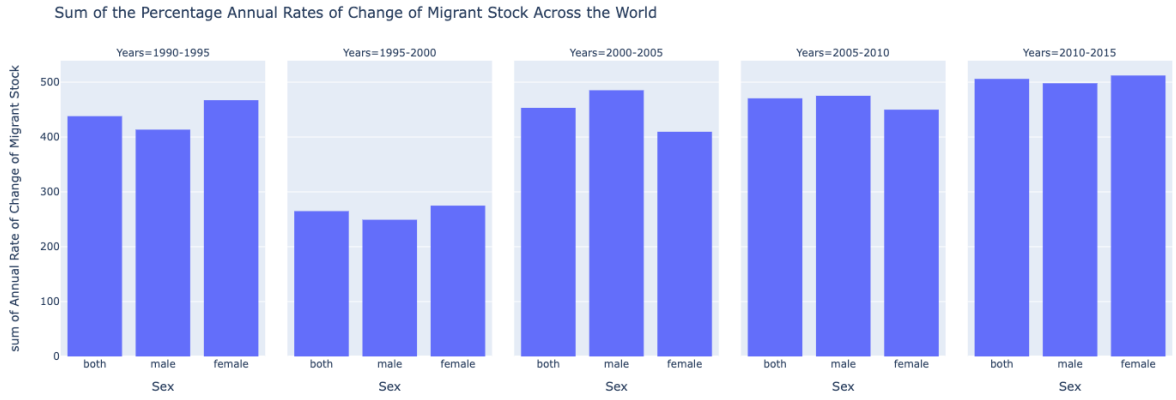
This figure from table 3 illustrates the migrant stock percentage for the major regions. Africa, Latin America/Caribbean, and Asia have very minimal proportion of migrants as their population. Conversely, Oceania has quite a high proportion of migrants.

Figure 7:



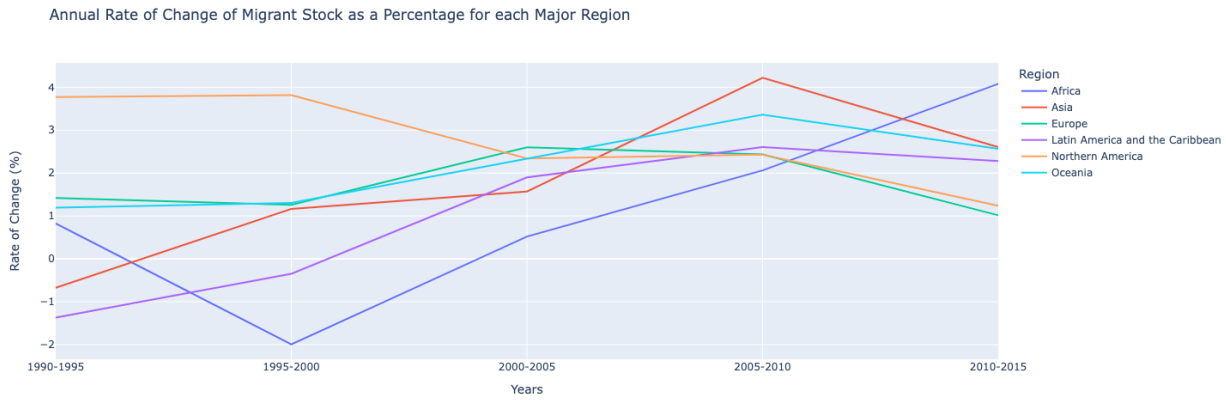
This figure from table 4 shows the proportion of migrants that are females for each of the major regions. Asia has had the most variation over the years, while Latin America/Caribbean has had the least variation. It is also interesting to note that even though Asia has had the greatest variation, the proportion of female migrants has always been far below 50%.

Figure 8:



This figure from table 5 shows the total sum of all rates of change of migrant stock across the world. There is a glaring difference between the years 1995-2000. The sum is far below all the sums for the other years, which are relatively consistent. Just by visual comparison, the sum for 1995-2000 is almost half of that for 2010-2015.

Figure 9:



This figure from table 5 shows the annual rate of change of migrant stock for the major regions. The general trend appears to be that the rate of change increased until a plateau in 2005-2010, and then the rate slightly decreased. Africa is the only region in which its annual rate of change continued to increase.

Discussion:

This portion of the project primarily focused on creating visuals for the dataset that was already cleaned for the midterm. There were minor adjustments that needed to be made in order to create better visuals, but the bulk of the cleaning was already previously done. As a general overview of what was seen, there appears to be greater rates of immigration around the globe. This can be due to a variety of factors, which from this dataset alone cannot be inferred.

With the exception of Figure 7, most of the figures bear very similar structural and visual resemblance to other figures, as they were constructed using similar code, but with variable names altered to coincide with the variables in the appropriate table.

The visuals for table 6 of the dataset have been omitted from this report. This is because it was realized that the data cleaning process for table 6 was incomplete and thus not properly organized for data visualization. I was unable to figure out how to do this, so to avoid presenting illegible figures they have been omitted to retain clarity in the report.

One thing to note is that because there were so many countries included in this UN dataset, the visuals weren't able to be constructed to show data from individual countries. This would have overcrowded the figures and reduce their legibility greatly. The alternative would have been having over a hundred figures for each of the individual countries. So, this report and these visualizations were limited to grouping into the major regions.

Conclusion:

To conclude, data visualization is crucial in order to convey messages from a dataset. A large set of individual data values on their own provides little value unless they can be comprehended. Visuals and figures are what allow data scientists to be able to show important conclusions from a data set. Using Tufte's principles provides a guideline to create effective visuals that are easier to interpret intuitively.

It is prominent to note that the data cleaning process was more tedious that the data visualization process. Exploratory data analysis was far more straightforward once the dataset had already been prepared for it. I think that this class provides a solid foundation in the important concepts of data science, and this project offers a way to practice some of those skillsets and give roots to build upon for further advancement in data science.