INF 1340 Final Project

Yidan Chen
Due Date: Dec 15, 2022

Introduction

Exploratory data analysis (EDA), which was originally introduced by American mathematician John Tukey in the 1970s, has now become widely used by data professionals to understand patterns, identify anomalies, spot outliers or unusual occurrences, discover intriguing relationships between the variables, test hypotheses, and validate assumptions (IBM, 2020). The final project is a continuation of the previous mid-term project, in which I adopted the tidy data method and implement the five principles of tidy data to clean and structure the dataset to facilitate analysis. For this final report, I will utilize various EDA methods as well as apply Edward Tufte's data visualization principles for the datasets that I cleaned and generated in the previous report.

The dataset used for cleaning was collected by the United Nations, the population division, and the department of economics and social affairs. The entire dataset includes 6 tables, and it contains the data and trends in international migrant and refugee stock from 1990-2015.

Method and Results

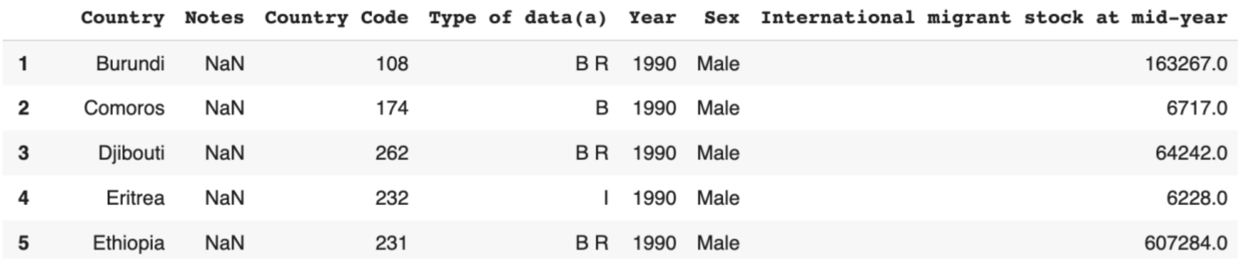
For the project, the programming language used for cleaning was Python, and Seaborn, Matplotlib, and Plotly are the main data visualization libraries for further analysis.

Table 1

In the previous project, I followed the tidy data principle and separated table 1 into 6 sub-tables, which are:

1. International Migrant Stock for both sexes(female/male) in Major Areas
2. International Migrant Stock for both sexes(female/male) in Regions
3. International Migrant Stock for both sexes(female/male) in Countries
4. International Migrant Stock for the sum of both sexes in Major Areas
5. International Migrant Stock for the sum of both sexes in Regions
6. International Migrant Stock for the sum of both sexes in Countries

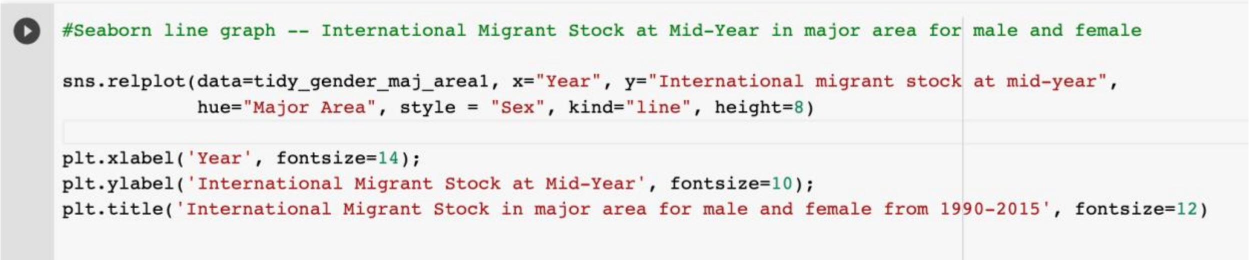
Starts with **Table 1, International Migrant Stock for both sexes(female/male) in Major Areas**



	Country	Notes	Country Code	Type of data(a)	Year	Sex	International migrant stock at mid-year
1	Burundi	NaN	108	B R	1990	Male	163267.0
2	Comoros	NaN	174	B	1990	Male	6717.0
3	Djibouti	NaN	262	B R	1990	Male	64242.0
4	Eritrea	NaN	232	I	1990	Male	6228.0
5	Ethiopia	NaN	231	B R	1990	Male	607284.0

Table 1, International Migrant Stock for both sexes(female/male) in Major Areas

Firstly, I generated a line chart using the seaborn library for this table, see the code below



```
#Seaborn line graph -- International Migrant Stock at Mid-Year in major area for male and female

sns.relplot(data=tidy_gender_maj_area1, x="Year", y="International migrant stock at mid-year",
            hue="Major Area", style = "Sex", kind="line", height=8)

plt.xlabel('Year', fontsize=14);
plt.ylabel('International Migrant Stock at Mid-Year', fontsize=10);
plt.title('International Migrant Stock in major area for male and female from 1990-2015', fontsize=12)
```

The line chart is used to show quantitative data across a continuous range or period of time and is mostly used to show trends and illustrate how the variables have changed over time. From figure 1, we can see that **Europe, Asia, and Northern America** are the three main areas with the most international migrant stock. The trends in **Africa and Oceania** are more steady comparatively. Among the six major areas, we can see that the international migrant stock is gradually increasing worldwide from 1990-2015, and most of them have relatively similar trends for males and female; however, I noticed that there was a significant increase for **Asia-Male** in around the year of 2005.

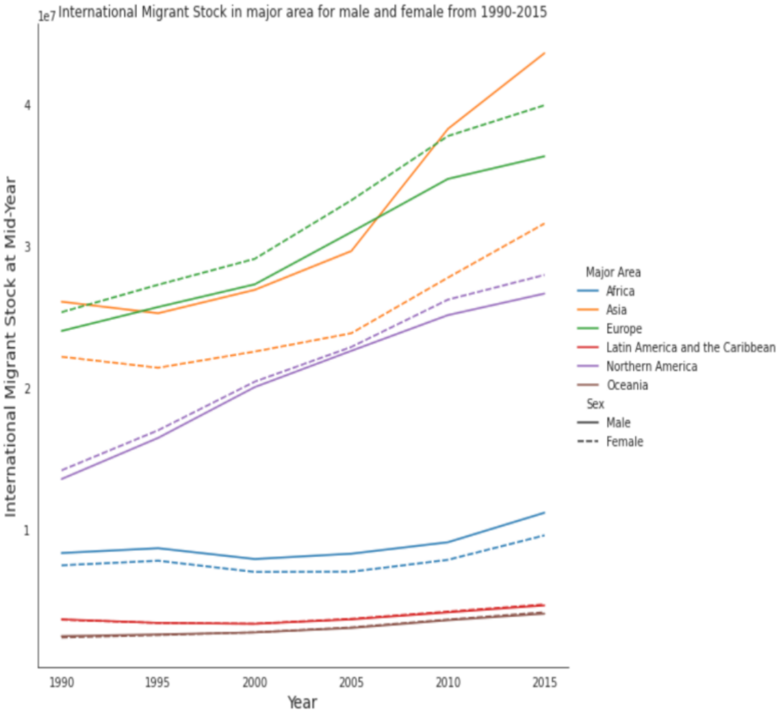


Figure 1

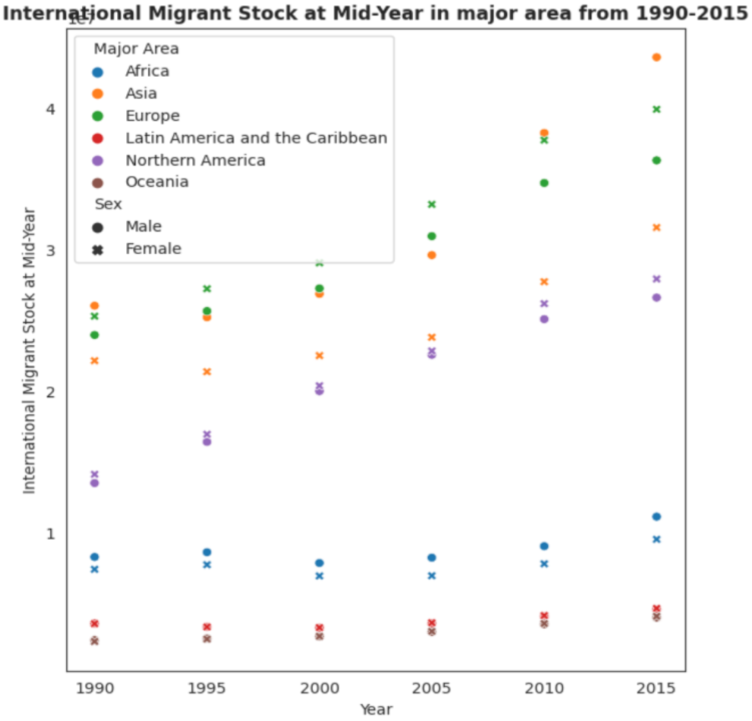
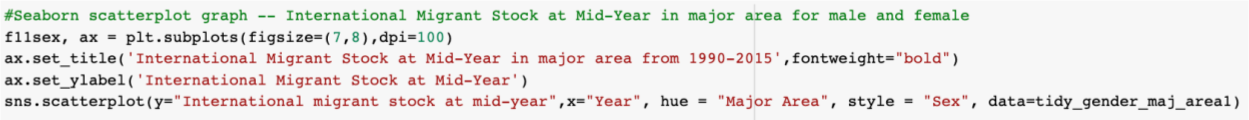
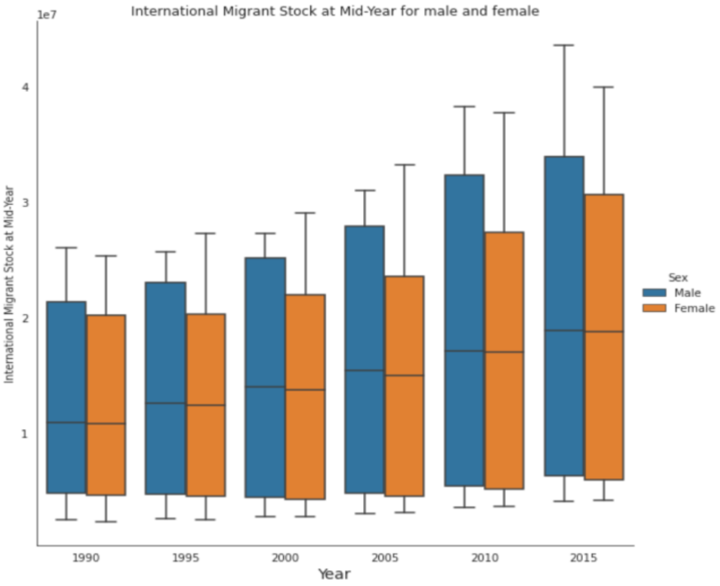


Figure 2

I also generate figure 2 using the scatterplot, by having an axis for each variable, we can observe that there is a **positive linear relationship** between the two variables: “international migrant stock” and “year”.



```
#Seaborn scatterplot graph -- International Migrant Stock at Mid-Year in major area for male and female
f1lsex, ax = plt.subplots(figsize=(7,8),dpi=100)
ax.set_title('International Migrant Stock at Mid-Year in major area from 1990-2015',fontweight="bold")
ax.set_ylabel('International Migrant Stock at Mid-Year')
sns.scatterplot(y="International migrant stock at mid-year",x="Year", hue = "Major Area", style = "Sex", data=tidy_gender_maj_area1)
```



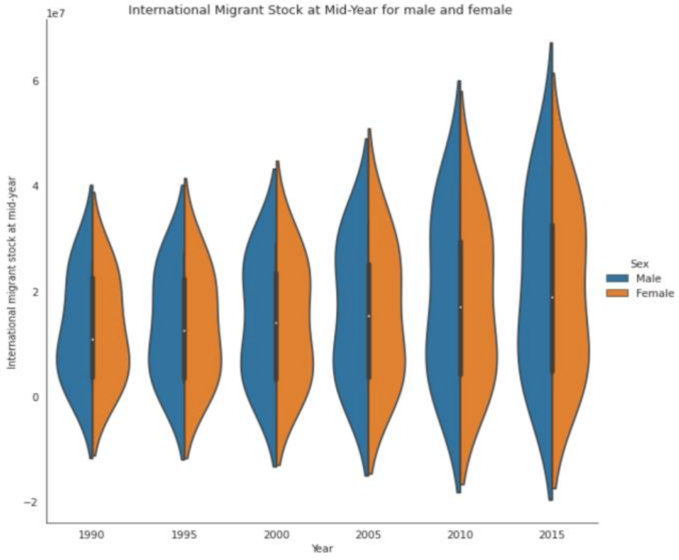
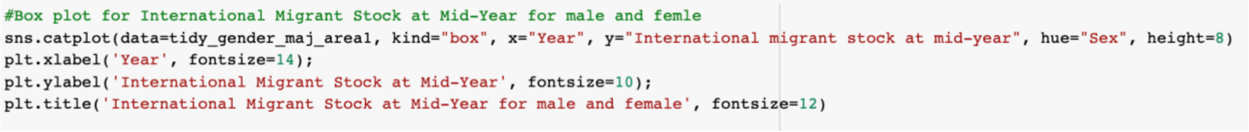


Figure 3

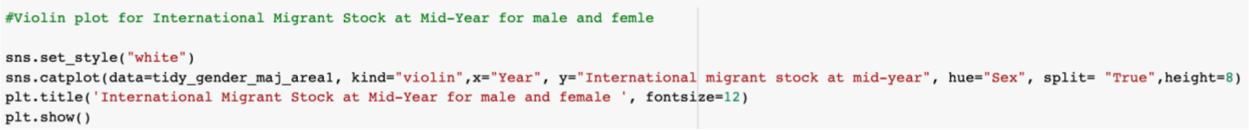
Figure 4

Then, I created the box plot and the violin plot for the international migrant stock for males and females worldwide. By observing the box plot in figure 3, we can tell that the box plot is getting longer and longer as time goes by, meaning that there are **more and more differences** among groups of the variables; we can also see that the lower quartiles are relatively stable, meaning that the there are **not significant changes in the lower end of the groups**, which also reflects what we observed in figure 1 and figure 2, the relatively stable migrant stock in Africa and Oceania.



```
#Box plot for International Migrant Stock at Mid-Year for male and femle
sns.catplot(data=tidy_gender_maj_area1, kind="box", x="Year", y="International migrant stock at mid-year", hue="Sex", height=8)
plt.xlabel('Year', fontsize=14);
plt.ylabel('International Migrant Stock at Mid-Year', fontsize=10);
plt.title('International Migrant Stock at Mid-Year for male and female', fontsize=12)
```

Figure 4 is a violin plot, a hybrid between a box plot and a kernel density plot to illustrate data peaks. Unlike box plots, which can only show summary statistics, violin plots provide how numerical data is distributed as well as the density of each variable. From figure 4, we can see that the shapes of the violin are similar for both males and females over the years, however, the shape of the violin is getting skinnier as time goes by, meaning that there **are more and more differences among variables groups and the data is spreading out gradually**.



```
#Violin plot for International Migrant Stock at Mid-Year for male and femle

sns.set_style("white")
sns.catplot(data=tidy_gender_maj_area1, kind="violin",x="Year", y="International migrant stock at mid-year", hue="Sex", split= "True",height=8)
plt.title('International Migrant Stock at Mid-Year for male and female ', fontsize=12)
plt.show()
```

I also created a facet grid scatterplot using the seaborn library as shown below in figure 5, for better side-by-side comparison, the conclusion also reflects on what we discovered from the previous figures that **Asia, Europe, and Northern America** are major areas with the most international migrant stock and **Asia** is the one with the most differences in gender.

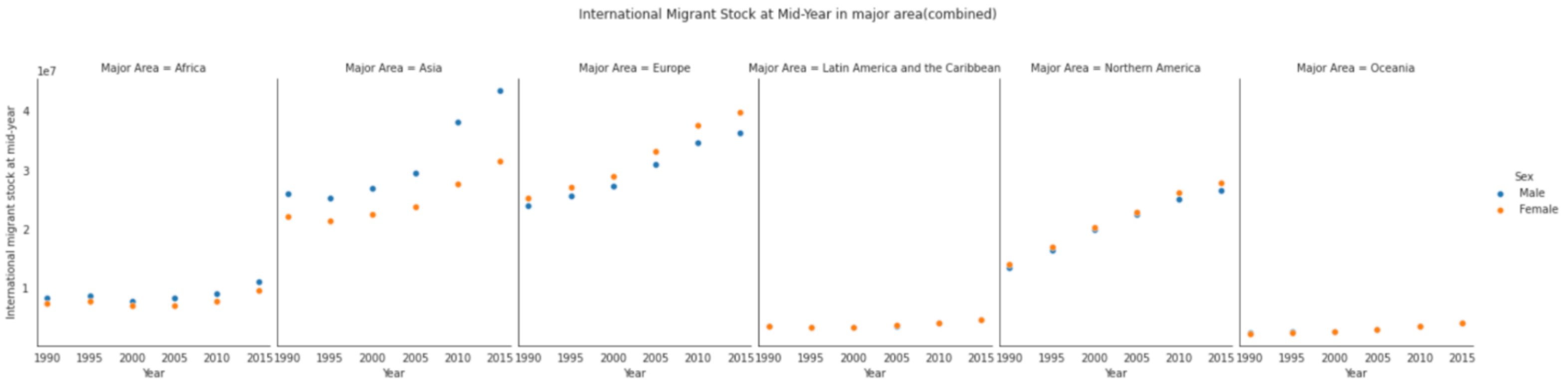
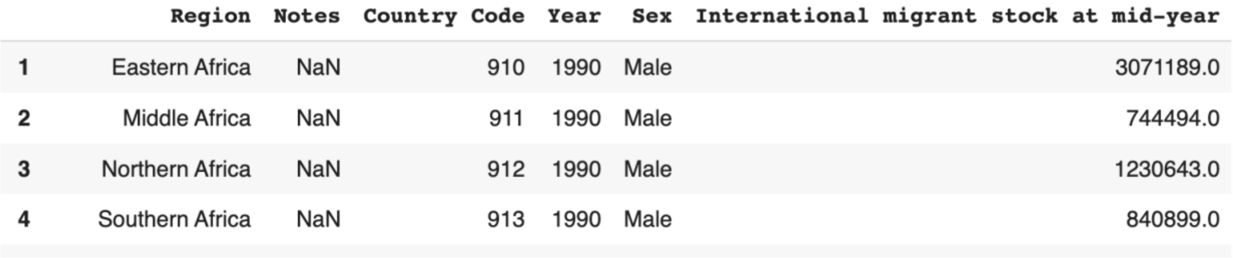


Figure 5

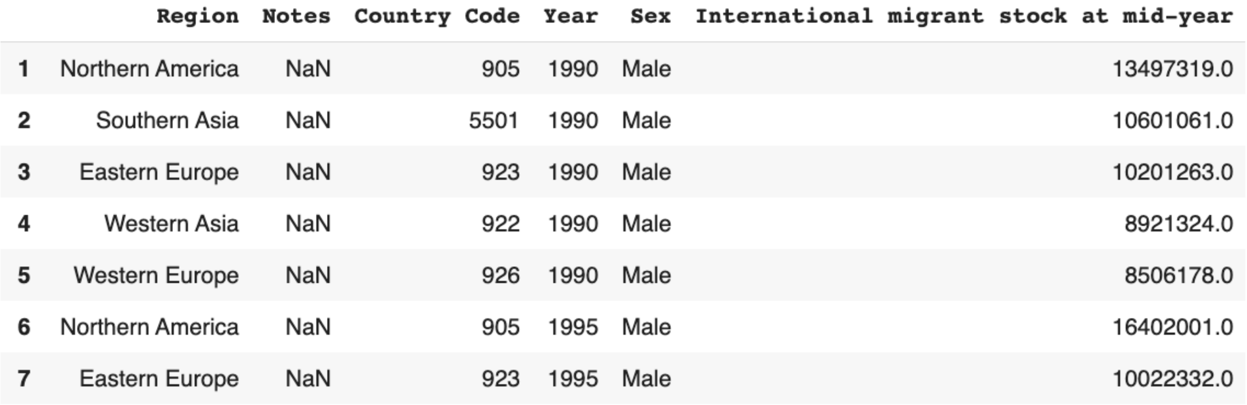
In order to find out what caused the gender differences in Asia and some detailed information on other major areas, I decided to look into table 2 to explore further.



	Region	Notes	Country Code	Year	Sex	International migrant stock at mid-year
1	Eastern Africa	NaN	910	1990	Male	3071189.0
2	Middle Africa	NaN	911	1990	Male	744494.0
3	Northern Africa	NaN	912	1990	Male	1230643.0
4	Southern Africa	NaN	913	1990	Male	840899.0

Table 2, International Migrant Stock for both sexes(female/male) in regions

Because there are 22 major areas and it would be hard to capture important information if we keep all the regions, I created a new table 3 (Below) to show the top 5 regions with the most international migrant stock for each year for both males and females.



	Region	Notes	Country Code	Year	Sex	International migrant stock at mid-year
1	Northern America	NaN	905	1990	Male	13497319.0
2	Southern Asia	NaN	5501	1990	Male	10601061.0
3	Eastern Europe	NaN	923	1990	Male	10201263.0
4	Western Asia	NaN	922	1990	Male	8921324.0
5	Western Europe	NaN	926	1990	Male	8506178.0
6	Northern America	NaN	905	1995	Male	16402001.0
7	Eastern Europe	NaN	923	1995	Male	10022332.0

Table 3, International Migrant Stock for both sexes(female/male) in top 5 regions each year

Firstly, according to **Tufet's principle, data visualization needs to be data-intense, design-simple, and word-sized graphics[2].** Small multiples, which show a series of the same small graph repeated in one visual, can be viewed as a great tool to visualize large quantities of data with a high number of dimensions. Therefore, I created the below small multiple histograms using Plotly Express in figure 6.

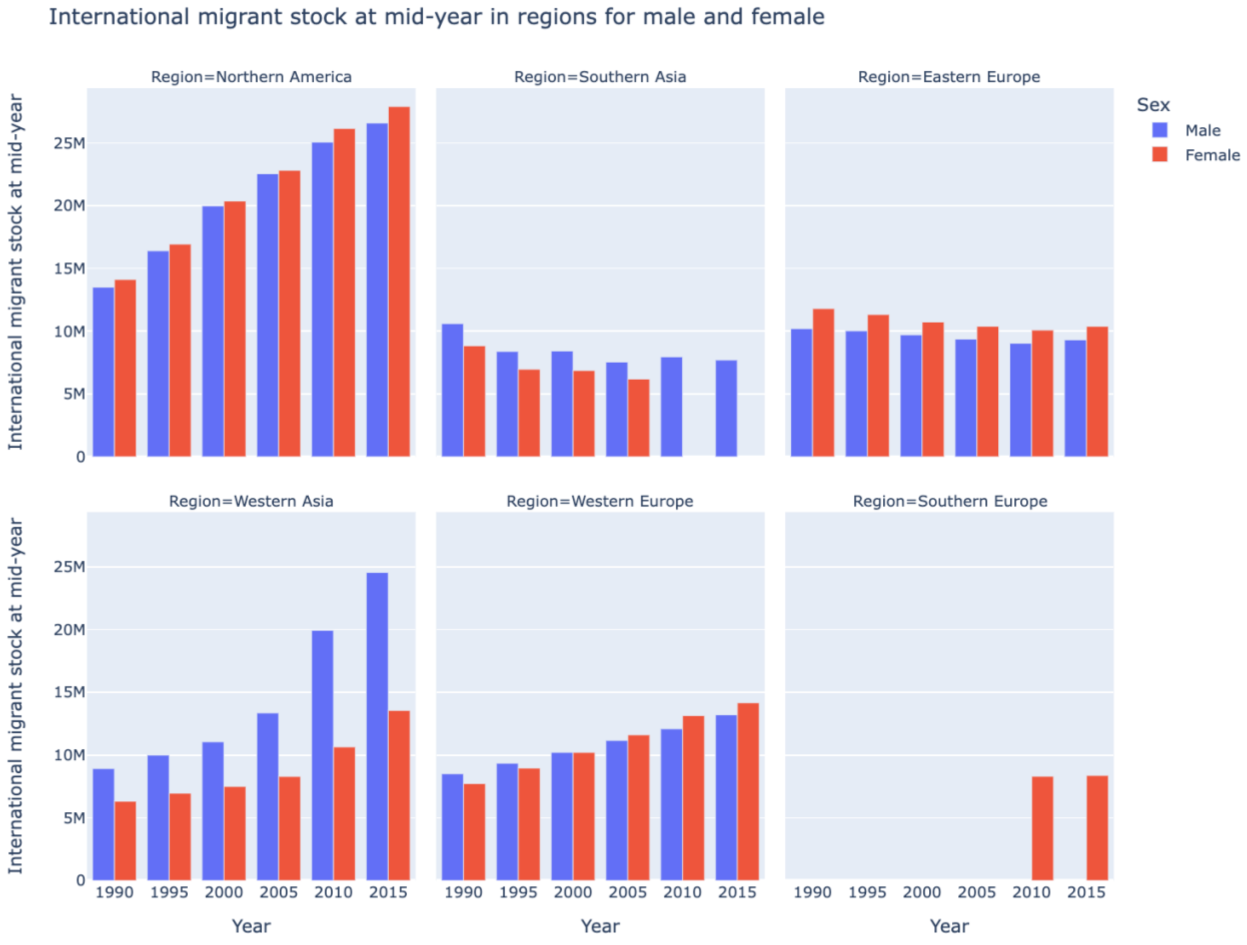
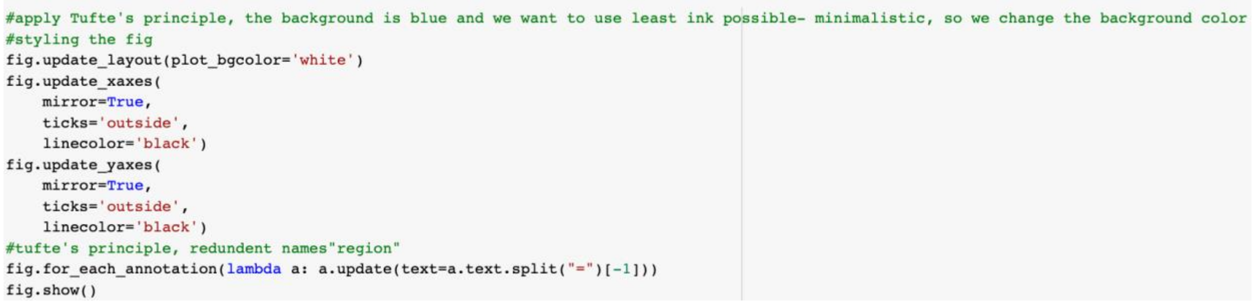


Figure 6

Secondly, according to **Tufet, it is important to erase non-data-ink, remove redundant data-ink, and create simplified visuals[2].** Because figure 6 has a blue background with the redundant title "Region" before all the region names, I applied another of **Tufet's principles, "Revise and edit"[2]** made the changes as below, and created figure 7.



```
#apply Tufte's principle, the background is blue and we want to use least ink possible- minimalistic, so we change the background color
#styling the fig
fig.update_layout(plot_bgcolor='white')
fig.update_xaxes(
    mirror=True,
    ticks='outside',
    linecolor='black')
fig.update_yaxes(
    mirror=True,
    ticks='outside',
    linecolor='black')
#tufte's principle, redundent names"region"
fig.for_each_annotation(lambda a: a.update(text=a.text.split("=")[-1]))
fig.show()
```

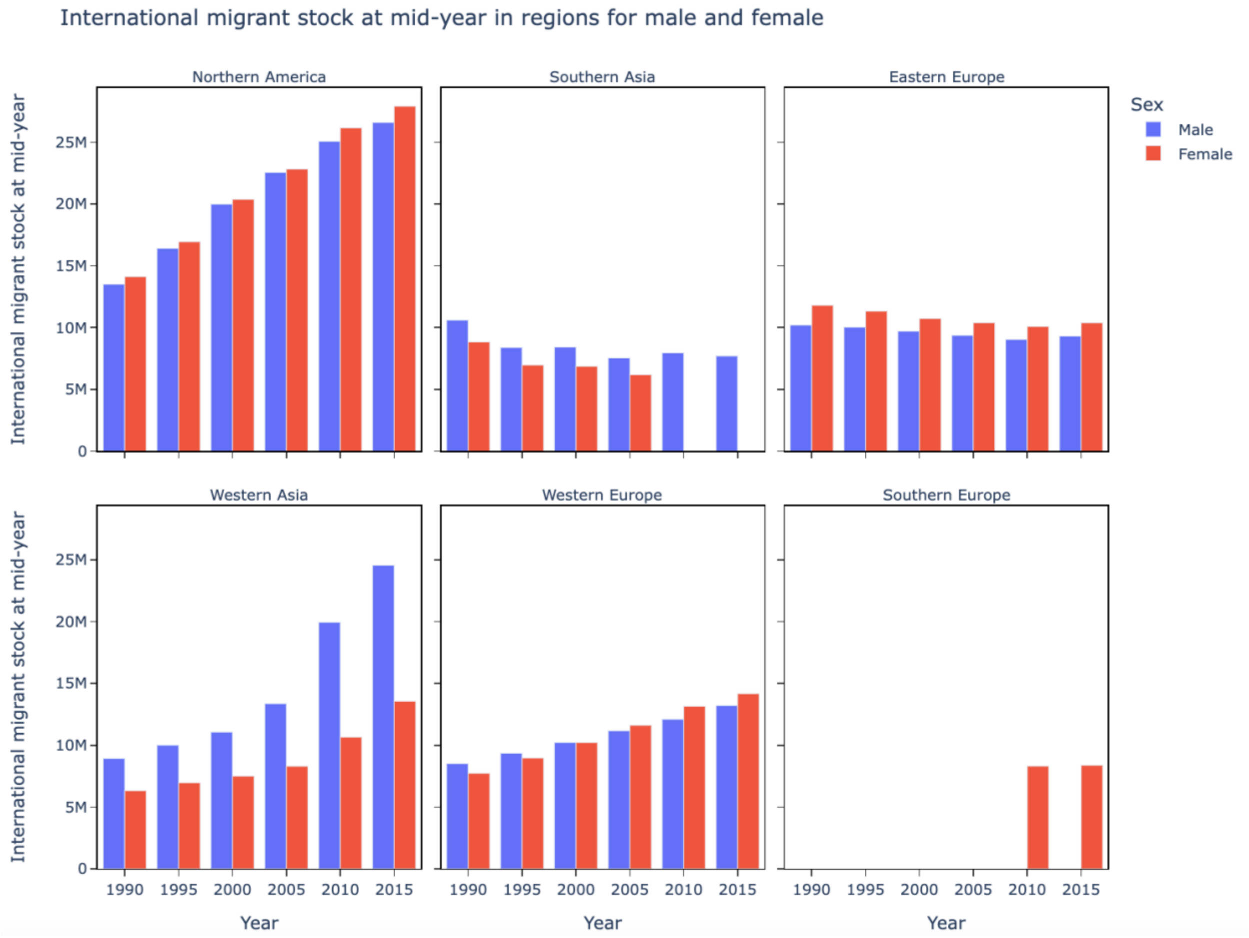
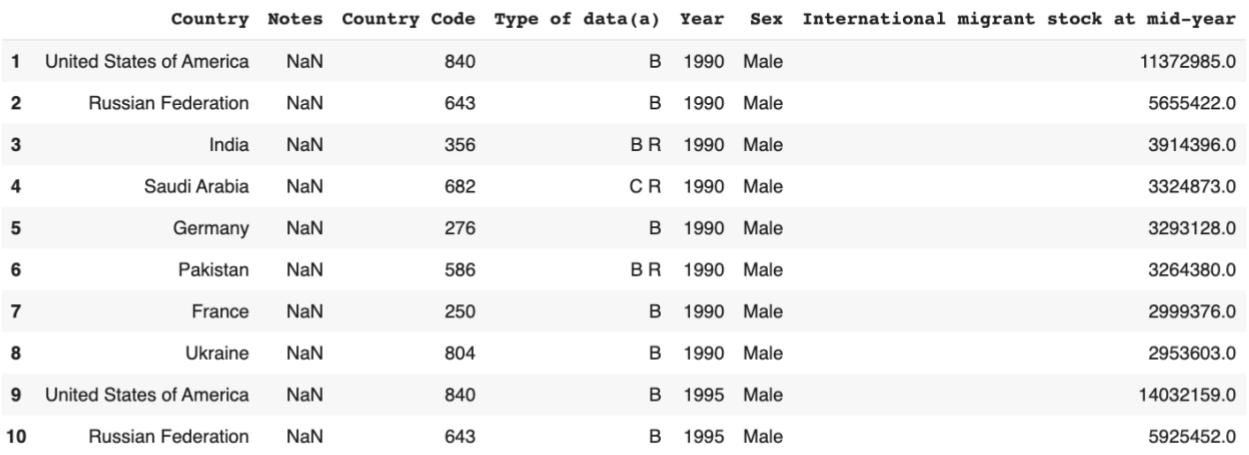


Figure 7

We can see from figure 7, Northern **America** was the region with the most international migrants and it is constantly growing over the years. For **Asia regions**, the migrant in **western Asia** is also increasing every year, especially the male population, I believe this also explains why we noticed the huge difference between the two genders in Figures 1, 2, and 5. Comparatively, in **Southern Asia**, the international migrant is getting stabler and has the tendency to decrease. Meanwhile, the migrant stock in both **Eastern and Western Europe** is relatively stable and I also find that there tend to be **more female migrants** in these regions compared to male migrants.

After exploration of various regions, now I would like to see the detailed migrant stock in different countries. Similarly, because there are hundreds of countries and it is hard to generate valuable insights if we put all of the data into one figure, so I selected the **top 8 countries with the most international migrant stock for each year for both males and females** and created the new table as below.



	Country	Notes	Country Code	Type of data(a)	Year	Sex	International migrant stock at mid-year
1	United States of America	NaN	840	B	1990	Male	11372985.0
2	Russian Federation	NaN	643	B	1990	Male	5655422.0
3	India	NaN	356	B R	1990	Male	3914396.0
4	Saudi Arabia	NaN	682	C R	1990	Male	3324873.0
5	Germany	NaN	276	B	1990	Male	3293128.0
6	Pakistan	NaN	586	B R	1990	Male	3264380.0
7	France	NaN	250	B	1990	Male	2999376.0
8	Ukraine	NaN	804	B	1990	Male	2953603.0
9	United States of America	NaN	840	B	1995	Male	14032159.0
10	Russian Federation	NaN	643	B	1995	Male	5925452.0

Table 4, International Migrant Stock for both sexes(female/male) in top 8 countries each year

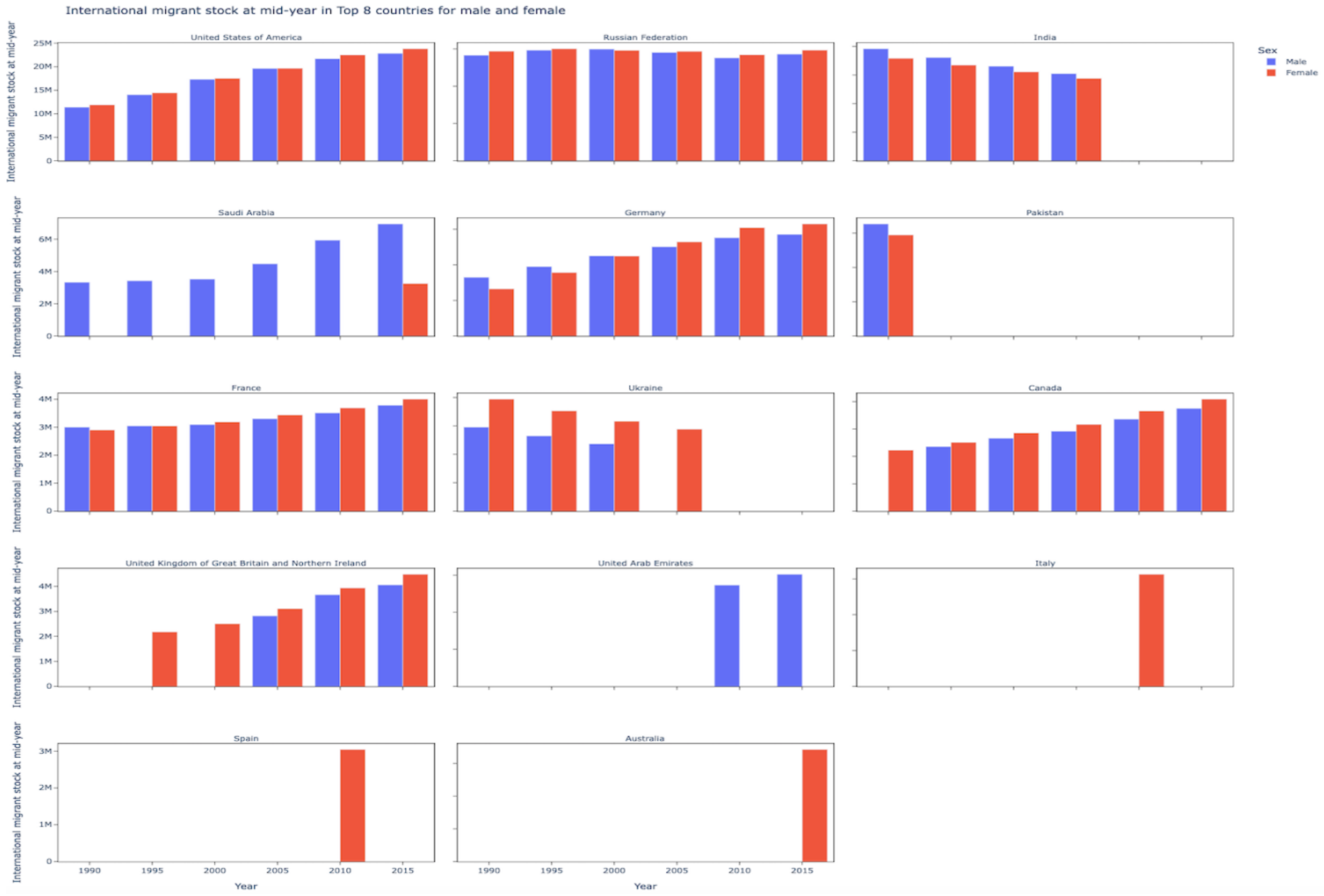


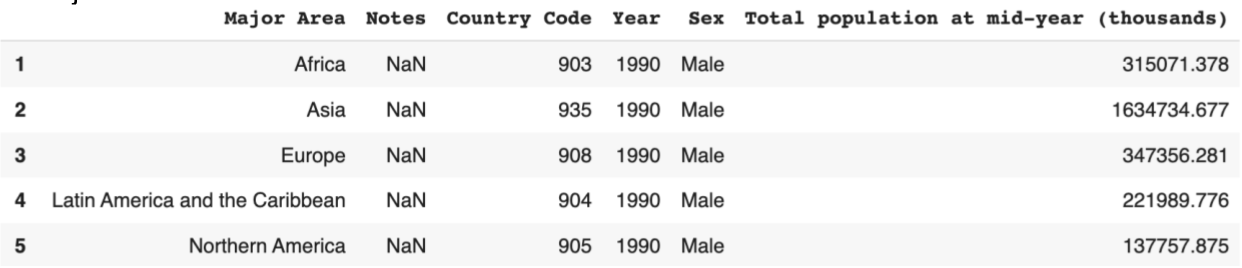
Figure 8

Then, I used table 4 and created figure 8 above. I followed **Tufte's principle, "Use appropriate scale"[2]** when drafting this graph and assigned different countries with different scale bars(i.e. scale for the U.S. is 0-25M and the scale for Spain is 0-3M). Even though this graph is relatively large, we can get lots of valuable insights from it. For example, **U.S. and Russia** are the two countries with the most migrant stock over the years, the migrant stock for the **U.S., and Canada**, is gradually increasing over the years, this may also explain why the stock for Nothern America was also increasing as we observed in figure 7. We also noticed that the migrant stock in countries such as **Germany and France** is relatively stable. Moreover, by observing the migrant changes in **Saudi Arabia**, it is obvious that the male stock is much higher than the female stock, thus this observation could clarify the question we had when observing figure 5 and figure7: the reason why there was a huge gender difference in **Western Asia** was mostly due to the wide gender difference in the countries such as **Saudi Arabia.**

Table 2

By observing the UN dataset table 2, we noticed that the only difference between table 1 and table 2 is that table 2 shows the total population for males, female and summation of both sexes by major area, region, country, or area from 1990-2015 and the UN-table1 only covers the international migrant stock. Therefore, we will follow the same logic and conduct data visualization analyses for UN-table 2.

Starts with the first tidy dataset listed below, the table shows the total population for both sexes in major areas from 1990-2015.



	Major Area	Notes	Country Code	Year	Sex	Total population at mid-year (thousands)
1	Africa	NaN	903	1990	Male	315071.378
2	Asia	NaN	935	1990	Male	1634734.677
3	Europe	NaN	908	1990	Male	347356.281
4	Latin America and the Caribbean	NaN	904	1990	Male	221989.776
5	Northern America	NaN	905	1990	Male	137757.875

Table 5, Total population for both sexes(female/male) in Major Areas from 1990-2015

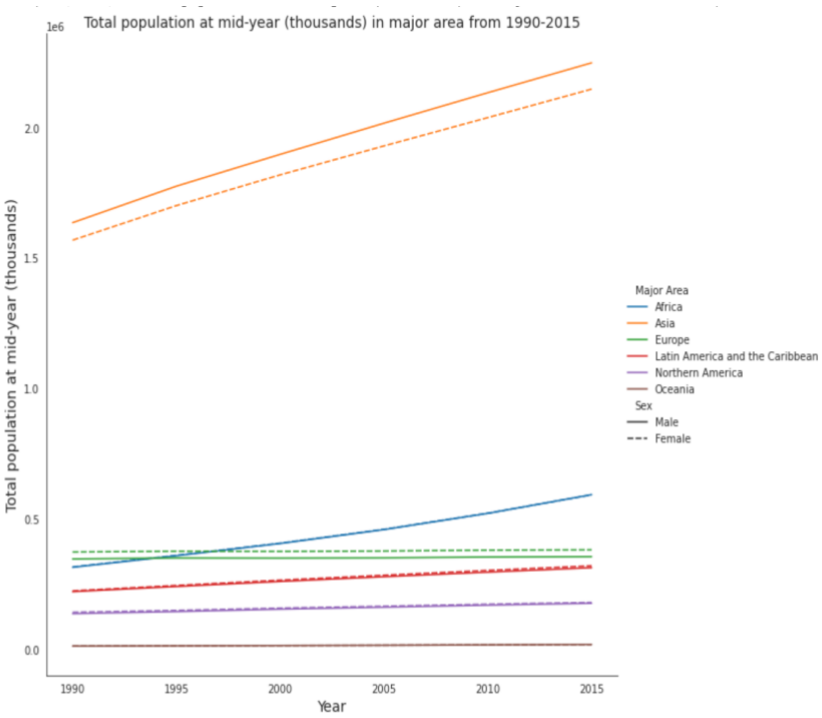


Figure 9

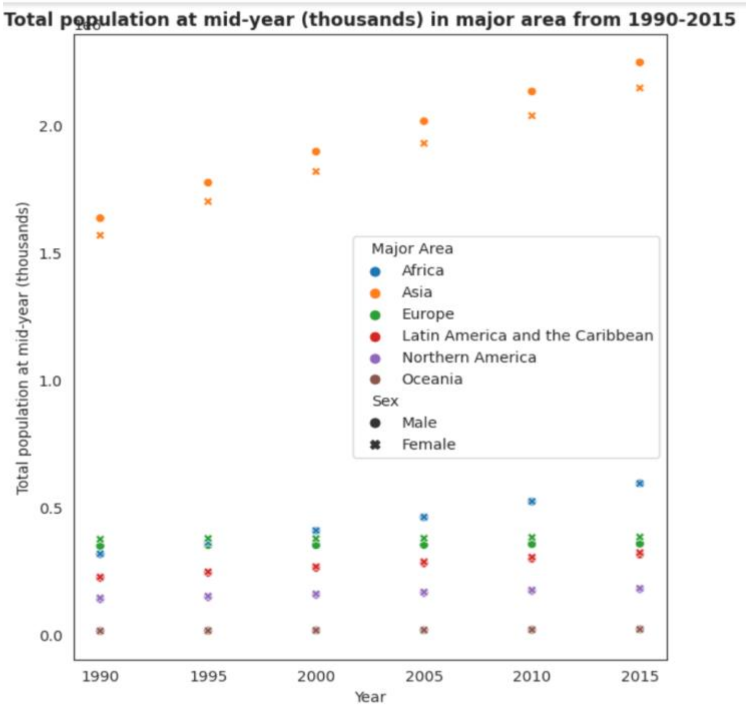


Figure 10

Then, I created a line chart and scatterplot using table 5. From figure 9 and figure 10 and figure 11, we can see that overall, the total population of female and male is increasing over the years from 1990-2015 for most of the major areas. **Asia** is the area with the highest increase among all other regions. By observing the outlier dot in Figure 11, the box plot emphasized the statement that the total population in Asia is significantly higher than all the other five regions. The total population in **Northern America, Eruope and Oceania** are relatively stable but **Europe** has a tendency of slight decrease in total population. For most regions, the change in population is similar for both female and male, but Asia has a higher increase in male population. In order to find what caused this higher increase, I will look deeper into specific regions, countries and discover data with more details.

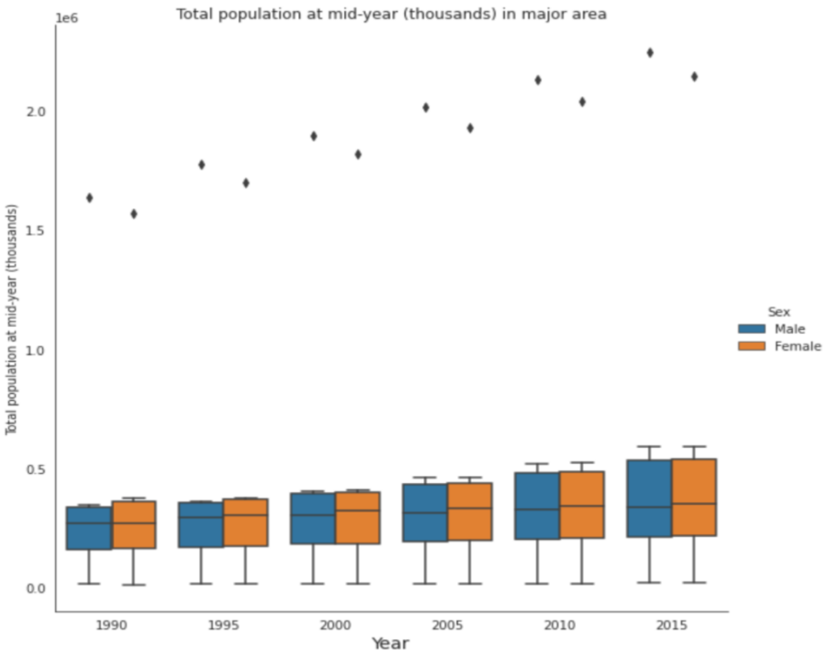


Figure 11

Then, I followed the same EDA logic from UN-table1 and made a new table with the Top 5 regions with the highest total population for males/females from 1990-2015. From figure 12, we can see that the top three regions with the highest total population are all in **Asia**, they are **Eastern Asia, Southern Asia, and South-Eastern Asia**. I also noticed that for both **Eastern Asia and Southern Asia**, the male population is higher than the female, which probably explained why there was a gender difference in previous figures 9,10, and 11. There are not many gender differences for **South-Estern Asia and South America**, but for **Eastern Europe, Northern America, and Eastern Africa**, the total population of females is comparatively higher than males.

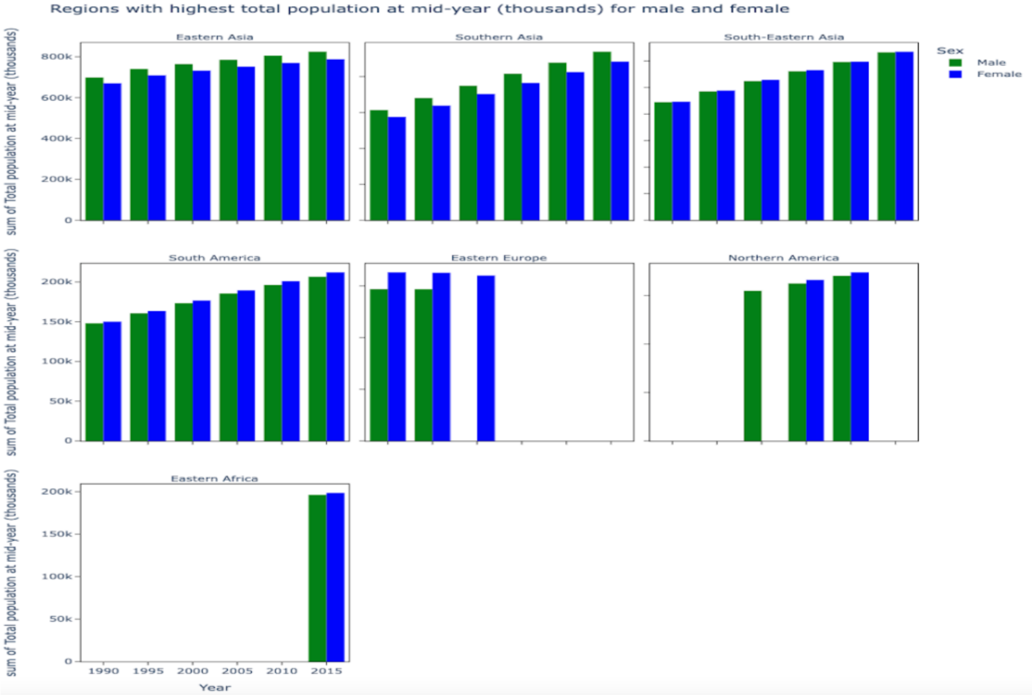


Figure 12

After looking into the top 5 regions with the highest total population, I created another table of the top 8 countries with the highest total population for males and females from 1990-2015. Instead of grouping them by “countries”, I grouped them by “year” this time and generated small a multiple figure13. From the histogram, we can see that **China and India** are the two countries with the highest total population and the number is significantly higher than all other countries. Also, both China and India have a higher male population than female. This could help to explain why Asia’s total population has a significant gender difference. Moreover, all other countries have a relatively even population in terms of males and females.

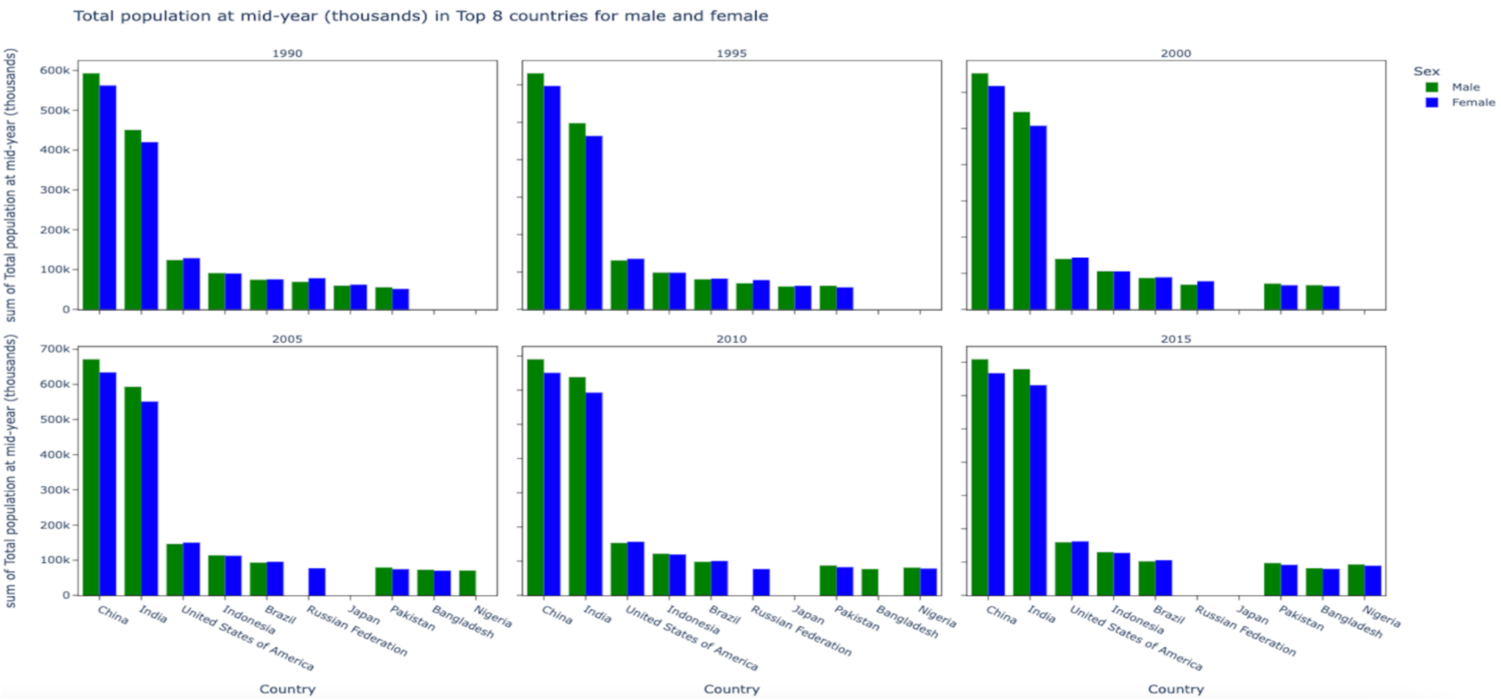


Figure 13

I also created another table using the summation of both sexes' populations in the top 8 countries and generated Figure 14. This box plot can help to reiterate the statement that we made earlier that **China and India** are the two countries with the highest total population. However, there are still differences between the two countries, we can see that the box plot for china is wider and shorter for China than for India, meaning that the population data in China is more compact and it is more spread out in India over the years between 1990-2015.

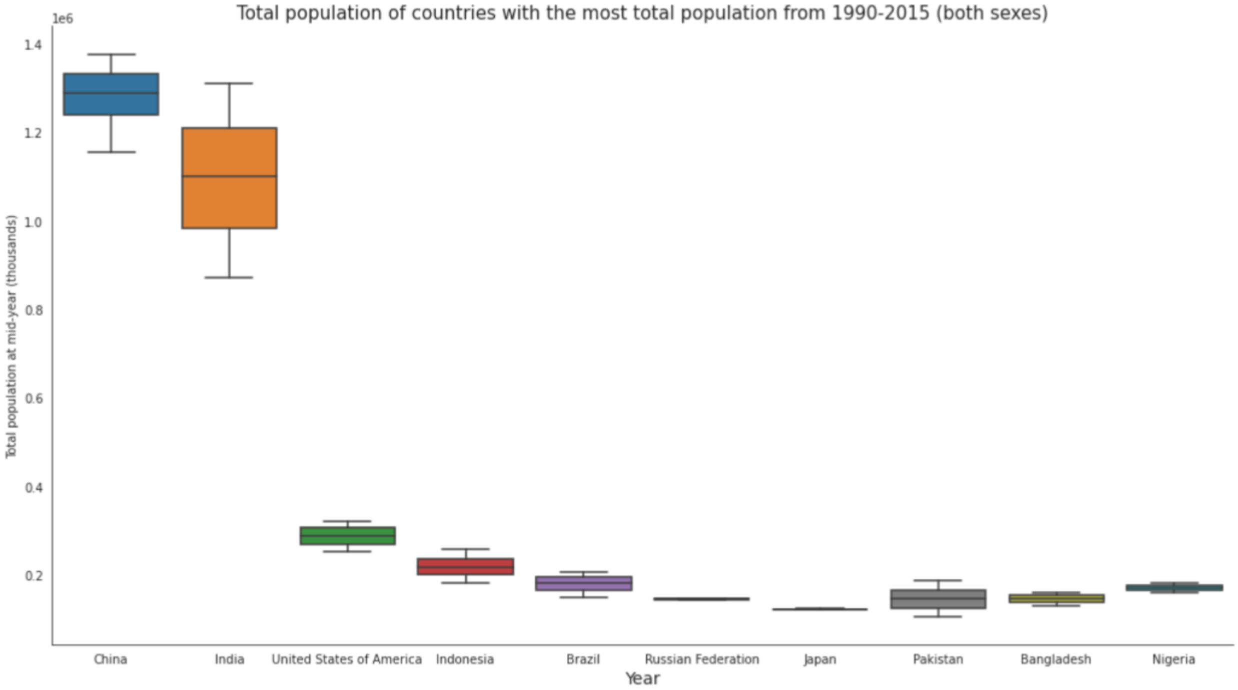


Figure 14

Table 3

By observing table 3, we noticed that table 3 represents the International migrant stock as a percentage of the total population by major area, region, country, or area from 1990-2015. It is the result of dividing table 1 to table 2. However, the nature of the datasets is still the same as in table 1 and table 2, so we will follow the same analysis logic and execute similar EDA steps for table 3.

Firstly, I used the tidy dataset and created a line chart that shows the International migrant stock as a percentage of the total population for males and females in Major areas from 1990-2015. We can see from figure 15 below that **Oceania, Northern America, and Europe** are the three major areas with the highest percentage of International migrant stock as a percentage of the total population, and the percentage is increasing over the years. However, we can see that the percentage is decreasing in **Africa** over the years and the percentages remain relatively stable for **Latin America and the Caribbean and Asia**. I did make another scatterplot for this dataset but the scatter dots looked very dense and confusing in the chart for the three major areas with a lower percentage, according to **Tufet's principle**, **“A good visual needs to illustrate complex ideas and communicates with clarity, precision, and efficiency”[2]**, but that figure failed to accomplish the purpose, so I decided to only keep the line chart in figure 15 and removed the scatterplot.

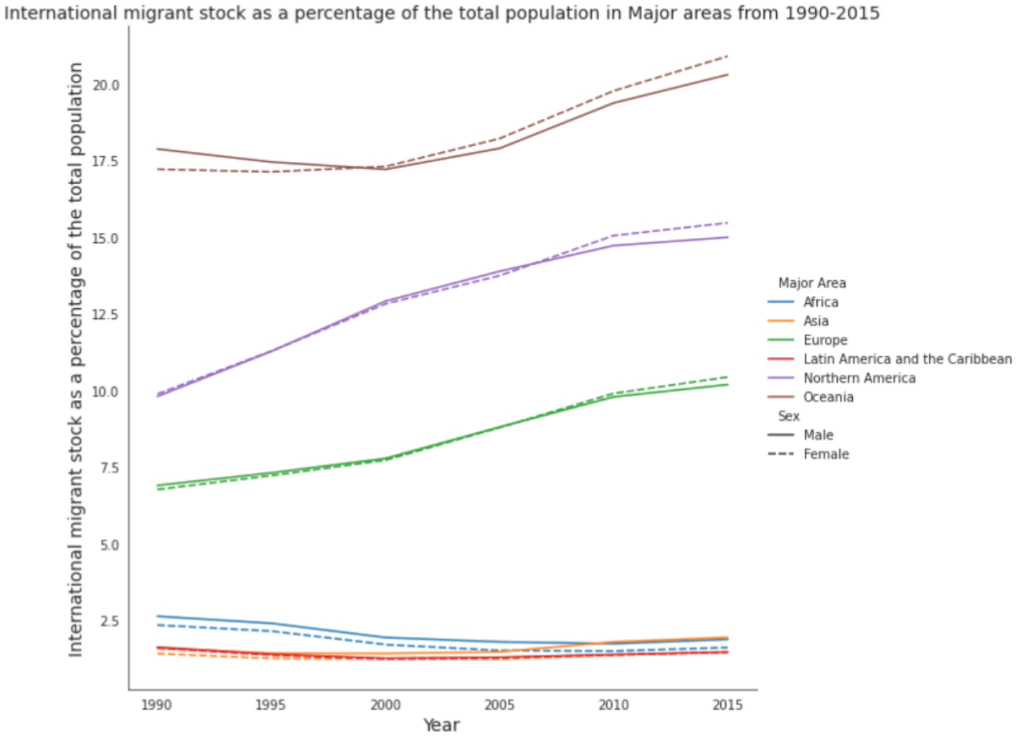


Figure 15

I also generated a box plot using the same dataset, I found that the **median** of the overall percentage didn't change much over the 25 years, but the box is getting **longer** as time goes by, meaning that there is **more diversity** in the dataset and the data is getting more and **more positively skewed**. Meanwhile, I also noticed that there is a **long upper whisker,** meaning that the percentages are varied amongst the upper quartile group. The figure can be used to guide exploring further into regions and countries.

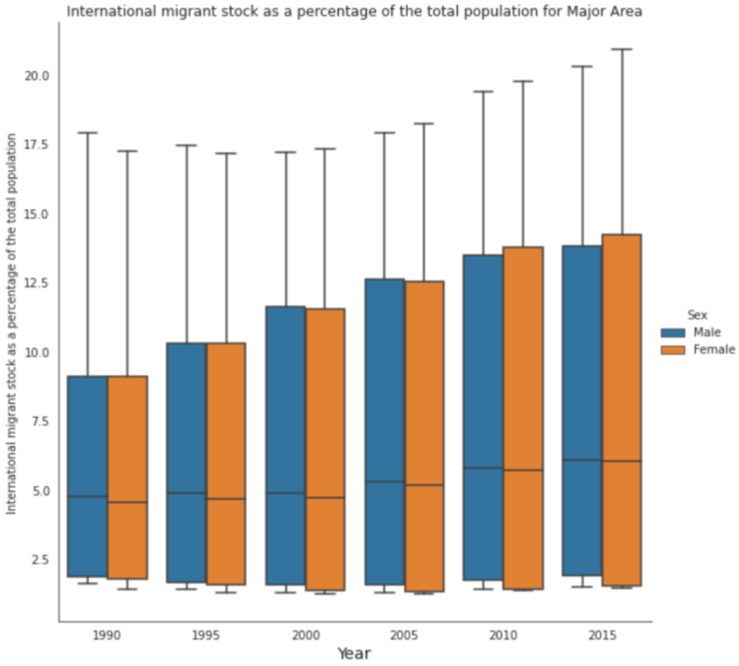


Figure 16

Then, I followed the same EDA logic and made a new table that shows the Top 5 regions with the highest percentage of migrants as a percentage of the total population for males/females from 1990-2015. Looking into the Top 5 regions, we can see that **Micronesia, Australia, and New Zealand** are the two regions with the highest percentage among all other regions. I believe it would be due to the reason that they also have a smaller population base than other regions with bigger population bases, such as Asia regions or North America; therefore, this means that the same increase in the migrant will impact the percentage more significantly for regions such as Micronesia. Noticeably, we also found that in **Western Asia**, there were only male data, and in **Northern Europe**, there were only female data. It could potentially be an increase in migrant stock as we discussed earlier in table 1, figure 6, or it could also be an increase in total population as we discussed in table 2, figure 12. Then, let's look further into various countries.

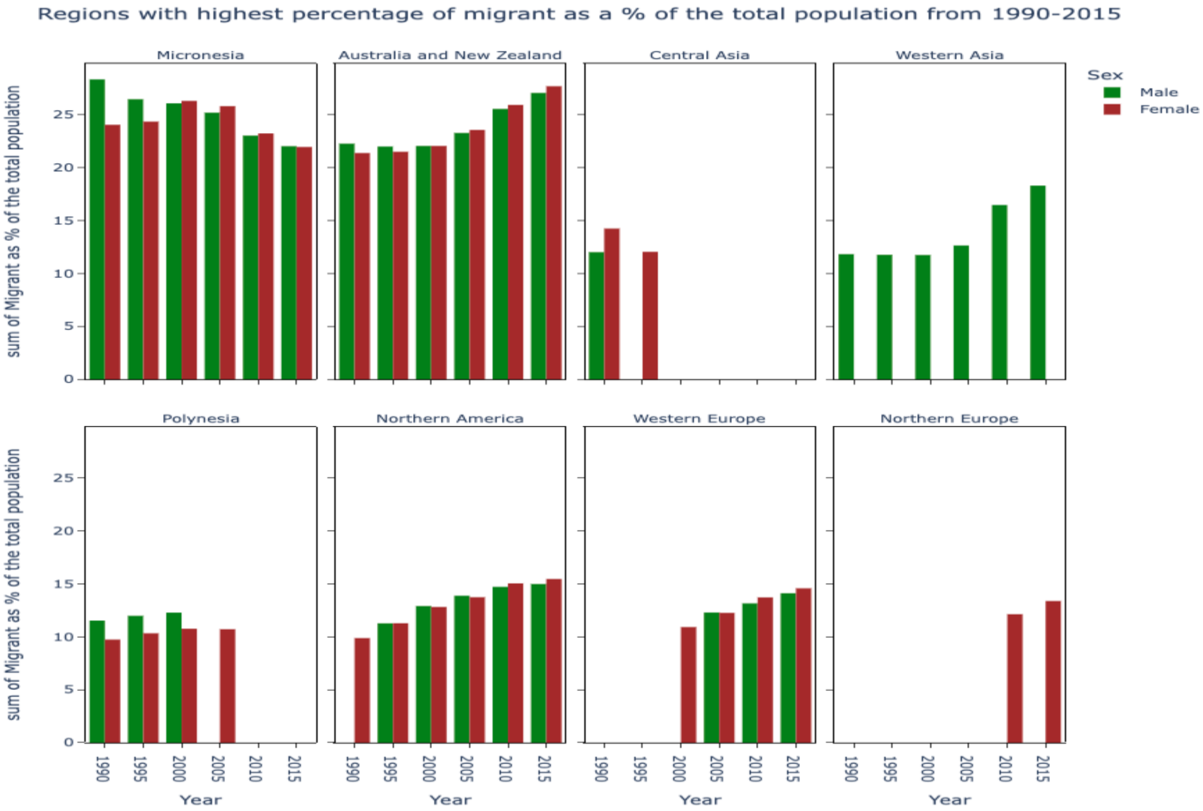


Figure 17

After looking into the top 5 regions with the highest migrant percentages, I created another table of the top 8 countries with the highest migrant percentages for males and females from 1990-2015. I grouped them by "year" this time and generated the small multiple histograms in figure 18. As shown in figure 18, we can see that **United Arab Emirates, Qatar, and Kuwait** are the three countries with the highest migrant percentages over total population and all of them are countries located in **Western Asia**, we also noticed that there the **male percentages are also higher than female percentages** for all three countries. This finding also aligns with our analysis made in figure 17. And the higher percentages of these countries could be due to the unstable political environment over the years between 1990-2015.

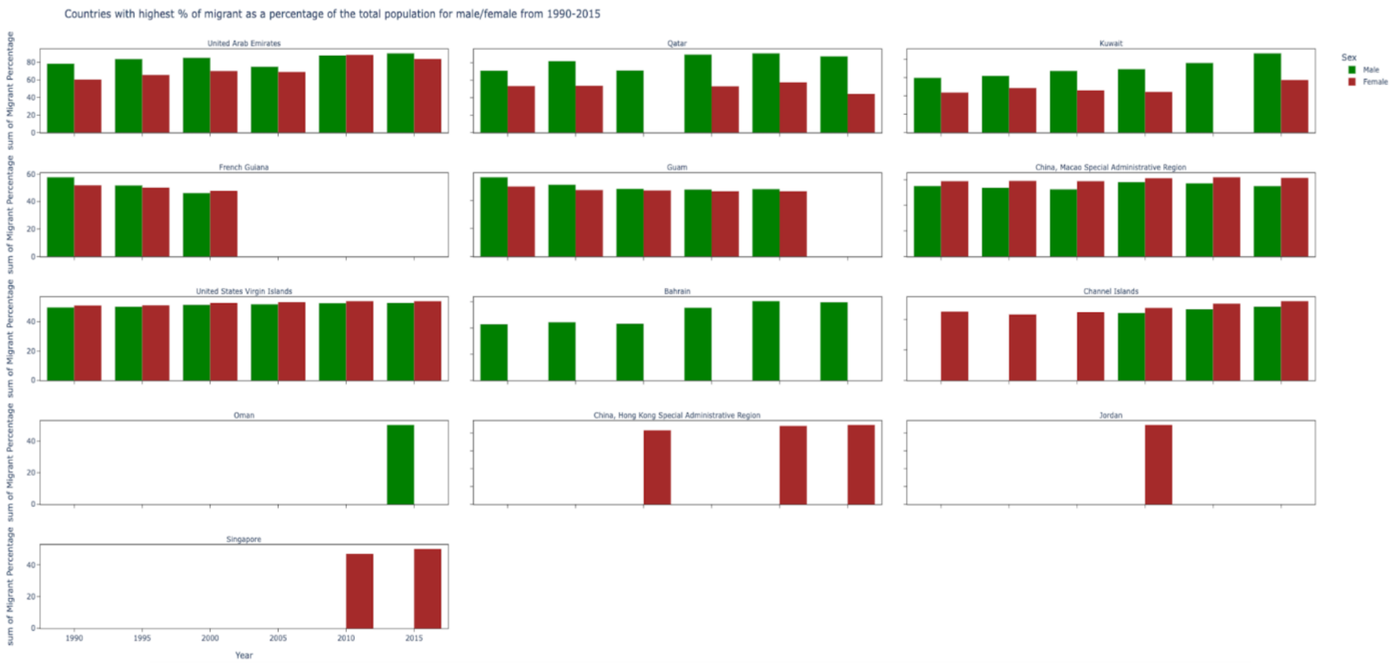


Figure 18

Table 4

By observing table 4, we can see that it is the dataset that represents female migrants as a percentage of the international migrant stock from 1990-2015. This is the result of dividing the table - International migrant stock for females, by the table - the sum of International migrant stock for both sexes. Still, the nature of this dataset is the same as in previous tables, and as some of the information and findings are already discussed in the previous sections for UN table 1 to UN table 3, I will use the same logic to proceed with a simplified EDA analysis for Table 4.

Firstly, by using the tidy dataset, I generated the following line chart that shows the female migrant stock as a percentage of the international migrant stock in Major areas from 1990-2015. We can see from figure 19 below that **Europe** is the major area with the highest percentage of female migrant stock as a percentage of the international migrant stock, and the percentage is increasing over the years. The percentages of **North America, Oceania, Latin America, and the Caribbean** are also increasing from 1990-2005, but starting to have a tendency to remain stable from 2005-2015. Noticeably, there is a significant percentage drop in **Asia** from 2005-2010. This finding also aligns with the **long lower whisker** as shown in figure 20 in the year of 2005-2010. We can observe from the box plot that even though the medians of the data didn't change significantly over the years, the plots are getting more **negatively skewed**, meaning that the medians are becoming closer to the top and the data constitute a higher frequency of lower end of values.

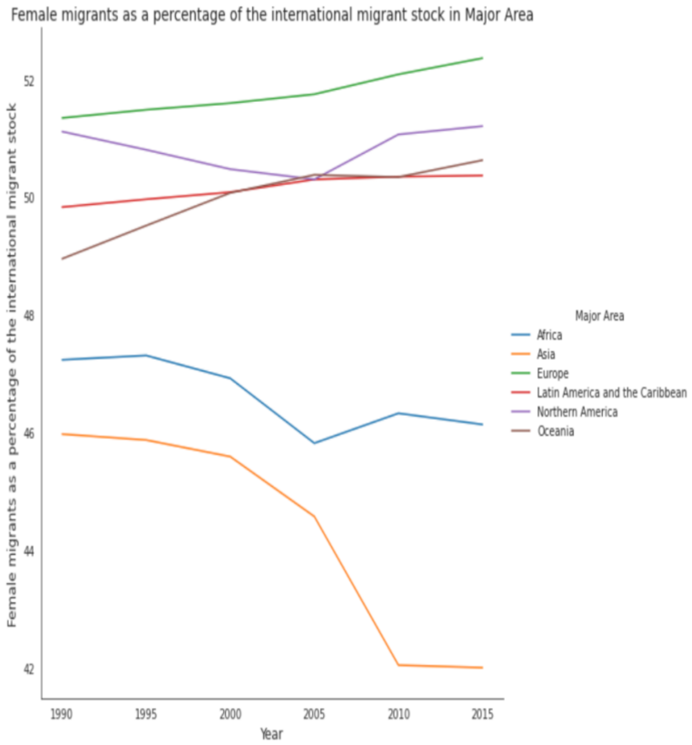


Figure 19

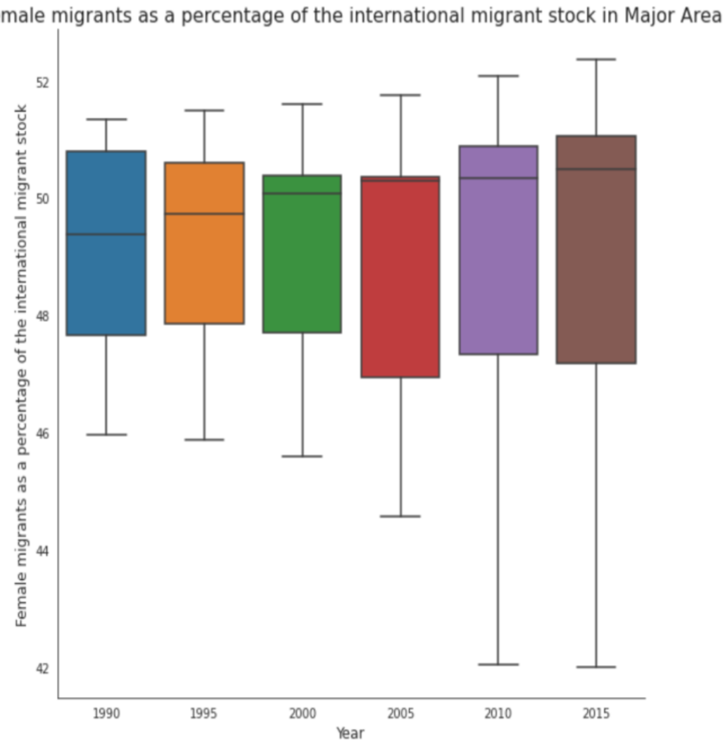


Figure 20

Table 5

Table 5 documented the annual rate of change of the migrant stock from 1990-2015. Because some of the information and findings are already discovered and discussed in the previous sections for UN table 1 to UN table 4, we will use the same logic and proceed with a simplified EDA analysis for Table 5.

Firstly, I created the following line chart for the annual rate of change of the migrant stock in the major areas for males and females as shown in figure 21. We can see that there was a significant negative drop in **Africa** between **1995-2000**. We can also observe that there was a positive rate of change in **Asia** between **2015-2010**, especially for the **male migrant population**, which aligns very well with our discussions and findings in the previous sections. The box plot in figure 22 can also help to provide more details of what we discovered in earlier figures as well as in figure 21, we can see that there is a **long upper whisker in 1990-1995 for female migrant stock,** meaning that the annual rates of change are varied significantly amongst the upper quartile group, Similar in 1995-2000, there is a **long lower whisker in 1995-2000 for male migrant stock**, meaning that datasets are very diverse amongst the lower quartile group.

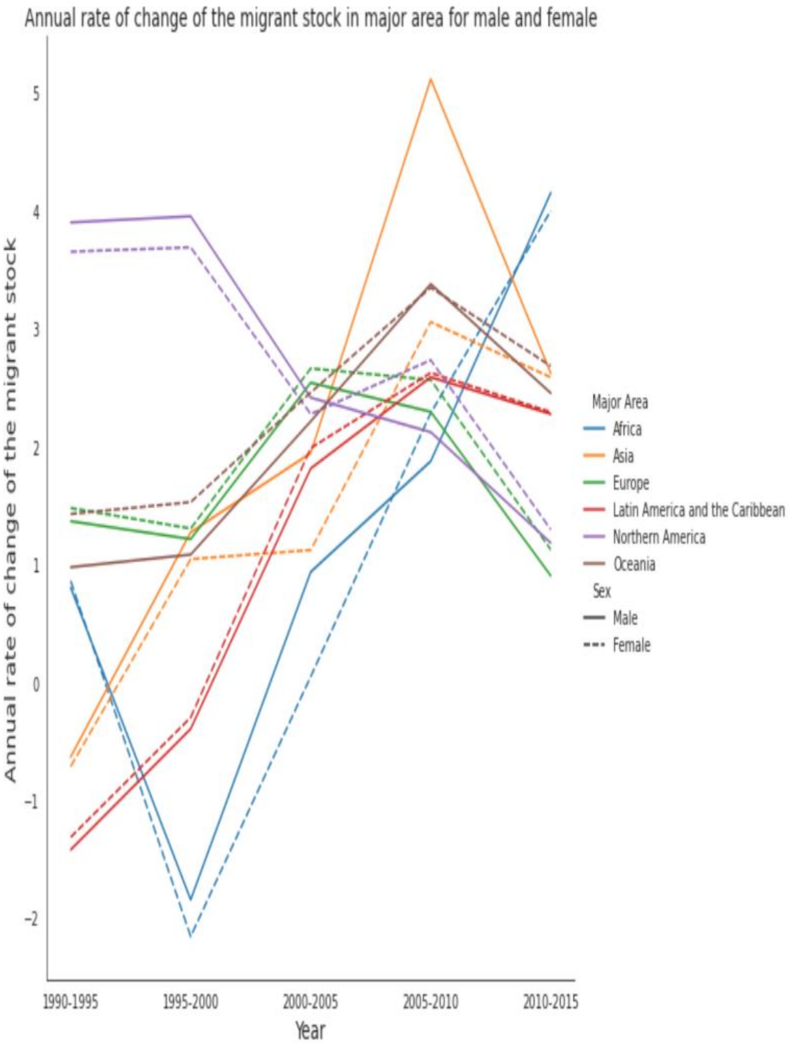


Figure 21

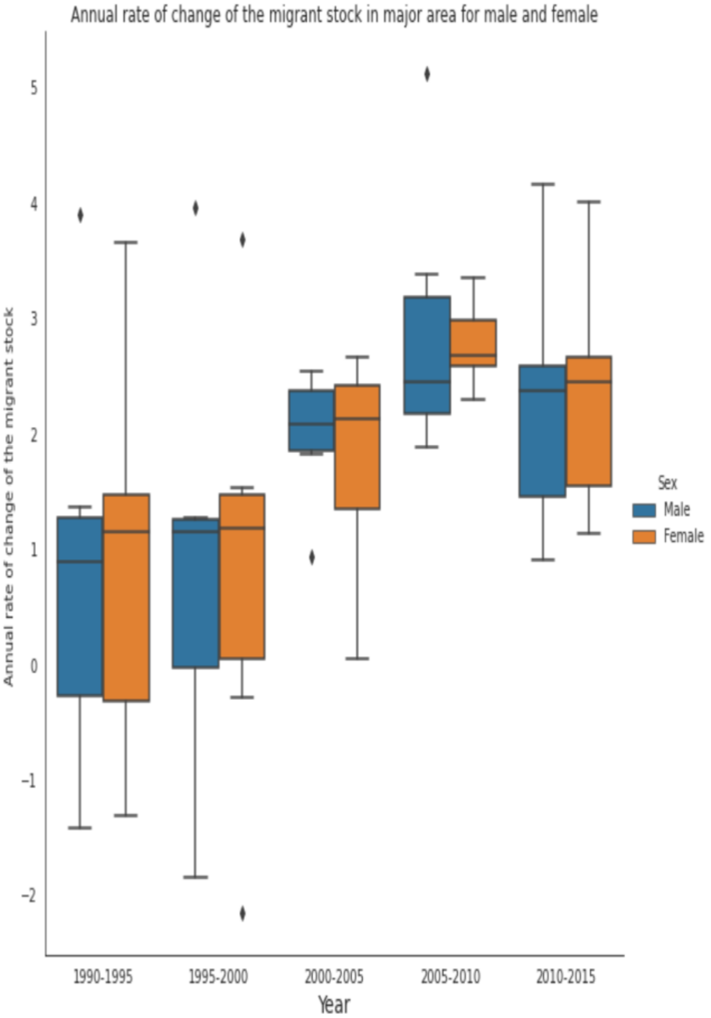


Figure 22

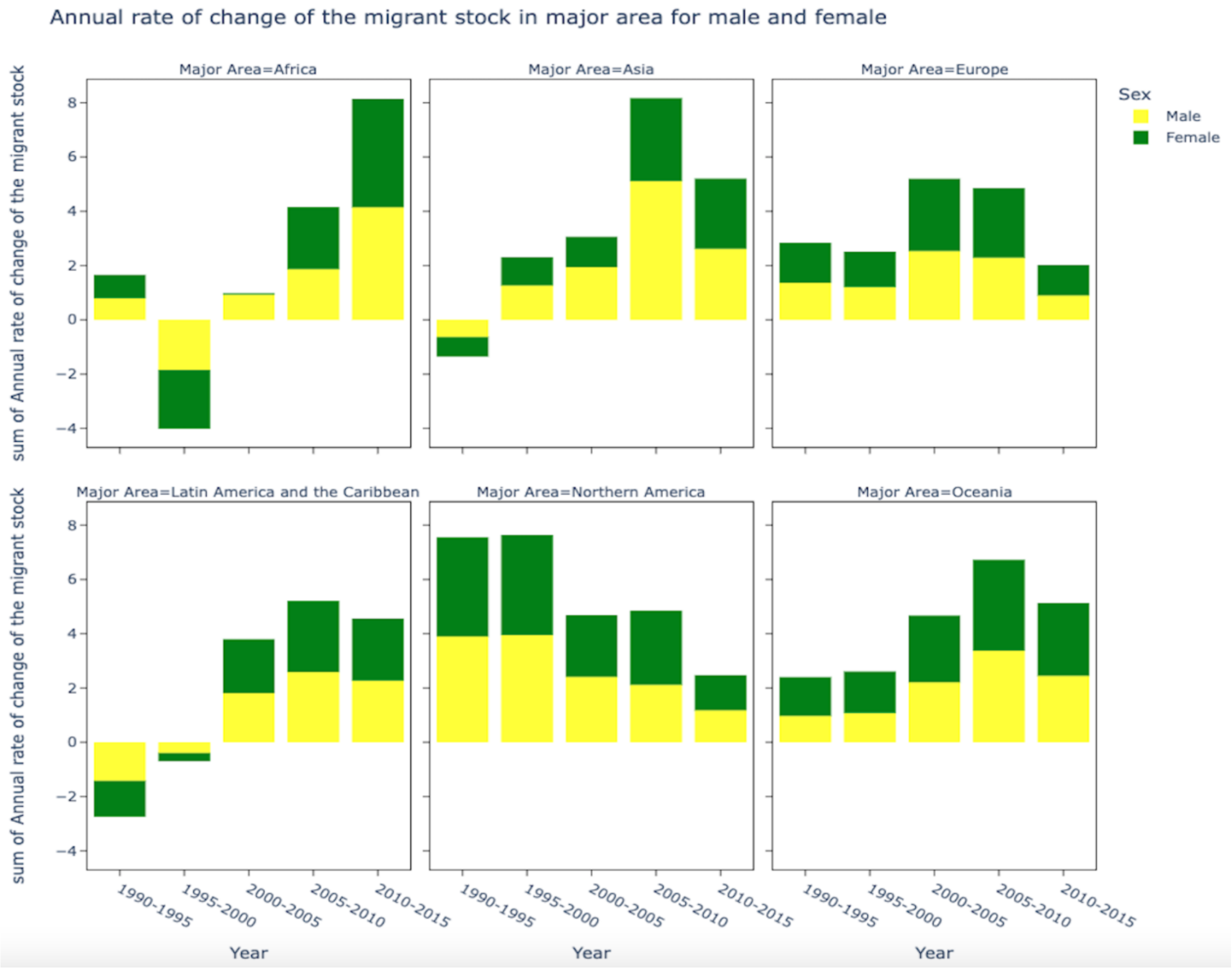


Figure 23

Because the lines shown in figure 21 are mixed together and hard to distinguish, I created small multiples staked histograms that show the annual rate of change in major areas for males and females from 1990-2015. Figure 23 provides a clearer view of the annual rate of change in various areas and we can see that most regions have a positive annual rate of change over the years, especially for **Northern America, Oceania, and Europe**. But there are also negative rates of change for **Latin America, the Caribbean, and Africa** between 1990-2000.

I also drafted the figures for the top 5 regions with the highest positive rate of change in migrant stock, and the top 5 regions with the highest negative rate of change in migrant stock, because the sizes of the figures are relatively large and the findings also align with our findings discussed earlier, these two figures can be found in the appendix as well as in the google collab notebook.

Table 6

By observing table 6, I found that table 6 has three types of data, which are 1. Estimated refugee stock at mid-year (both sexes), 2. Refugees as a percentage of the international migrant stock and 3. The annual rate of change of the refugee stock from 1990-2015. Because they are different data types, I will separate them into two parts and conduct EDA analyses for them.

Part 1: Estimated refugee stock at mid-year (both sexes)
As we can see from the below figures that show the Estimated refugee stock at mid-year (both sexes) for major areas from 1990-2015, **Asia** has the highest number of estimated refugee stock and **Africa** has the second highest refugee stock among other regions. We noticed that there was a significant increase in refugee stock from **2005-2015 in Asia** and a significant drop in refugee stock from **1995-2015 in Africa**.

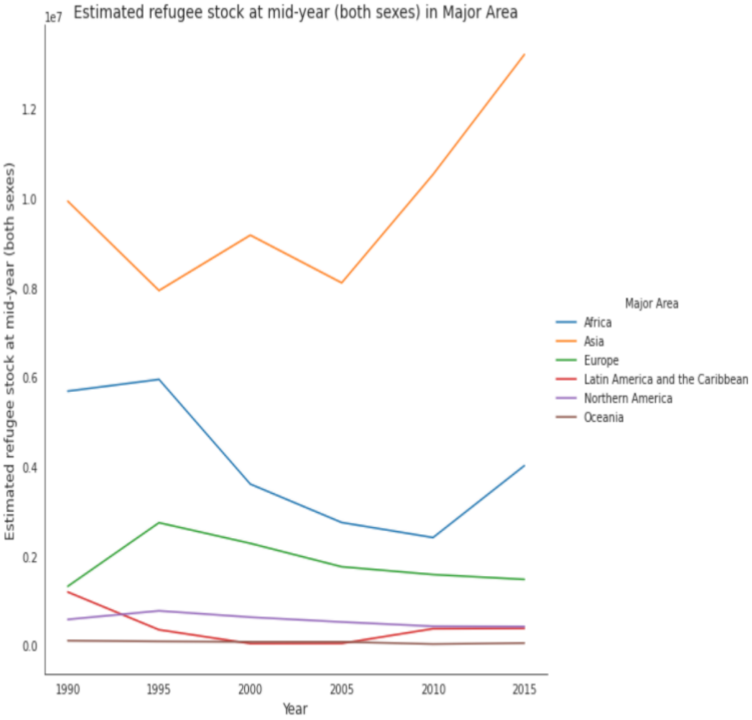


Figure 24

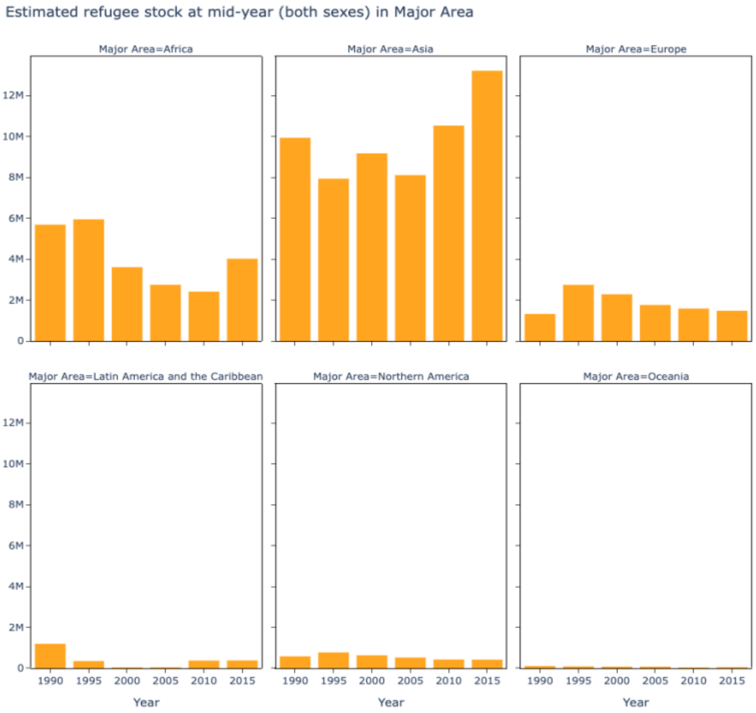


Figure 25

By zooming in and looking and the top 5 regions with the most estimated refugee stock from 1990-2015, we found that the three regions with the most refugee stock were **Southern Asia, Western Asia, and Eastern Africa**, which matched perfectly with our findings discussed above. Among them, there was a significant yearly increase in **Western Asia**, and this could be the potential reason why there was an **increase in Asia from 2005-2015**.

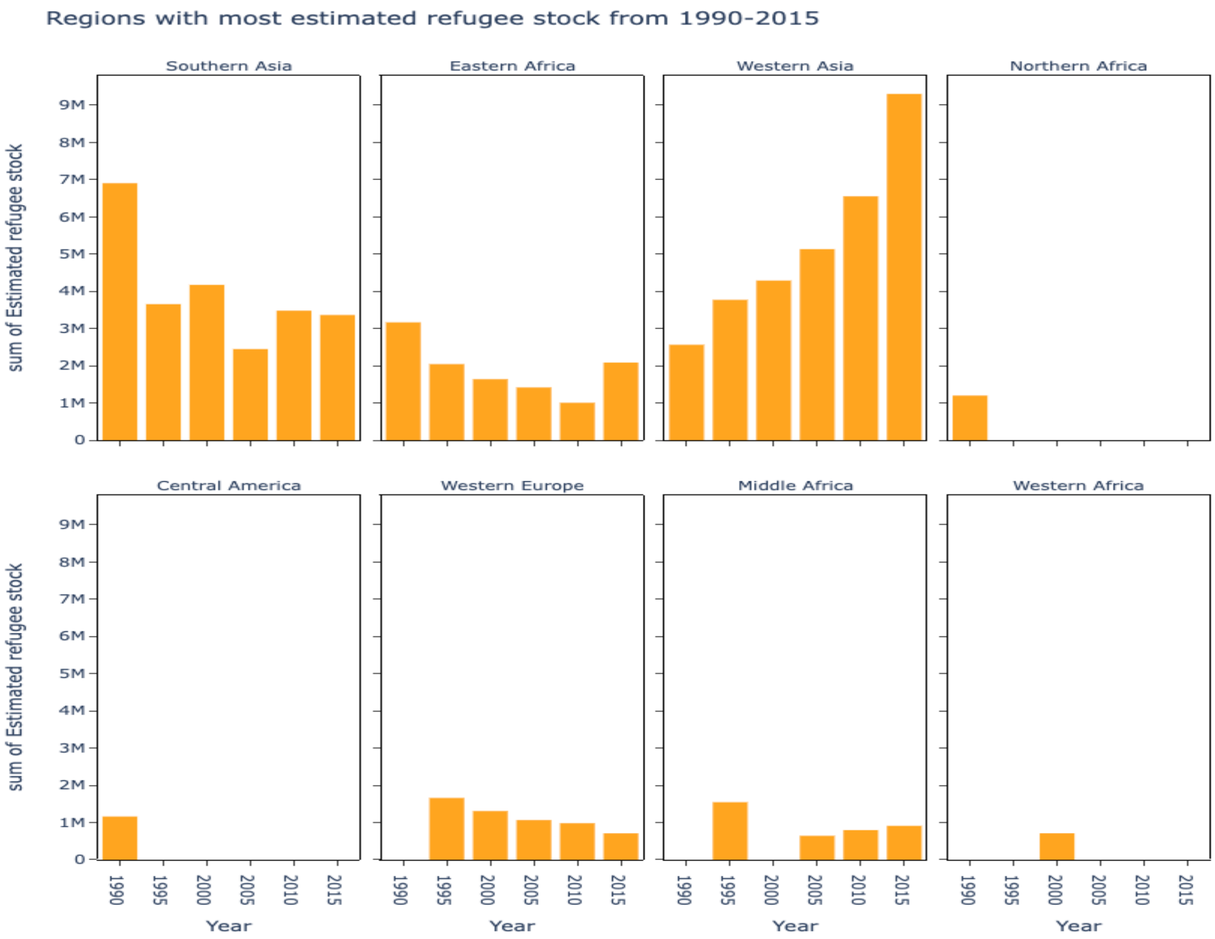


Figure 26

I also created the small-multiple line charts for the top 8 countries with the most estimated refugee stock from 1990-2015. As shown in Figure 28, most of the countries with the highest number of refugees are located in **Asia**, especially **Jordan and the State of Palestine**, the refugee is constantly growing from 1990-2015, and the increase in refugees could be due to the negative political and economical environments in these two countries over the years.

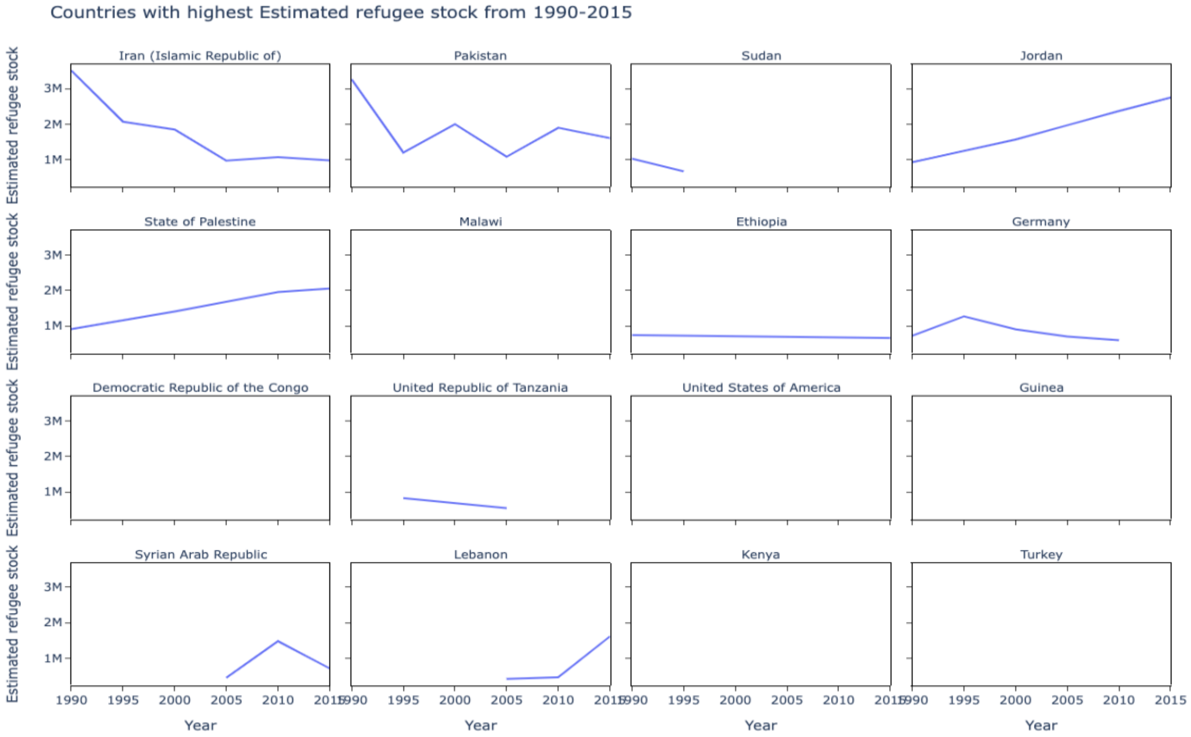


Figure 27

Part 2: Refugees as a percentage of the international migrant stock

According to the below figures that show the refugees as a percentage of the international migrant stock in major areas from 1990-2015, we can see that **Africa, Asia, and Latin America and the Caribbean** were the three major areas with the highest refugees percentage in 1990, the percentage dropped significantly during 1990-1995 for **Asia and Latin America and the Caribbean** but didn't drop until 1995 for **Africa**. It was potentially because the refugee stock started to decrease in 1995 in **Africa**, and the **reduction in refugees** caused a significant drop in Africa during 1995-2000. From figure 30, we can see that **Northern America and Oceania** are the two regions with the lowest refugee percentage of international migrant, it could be because these two areas have maintained relatively stable political and economical environments over the years from 1995-2015.

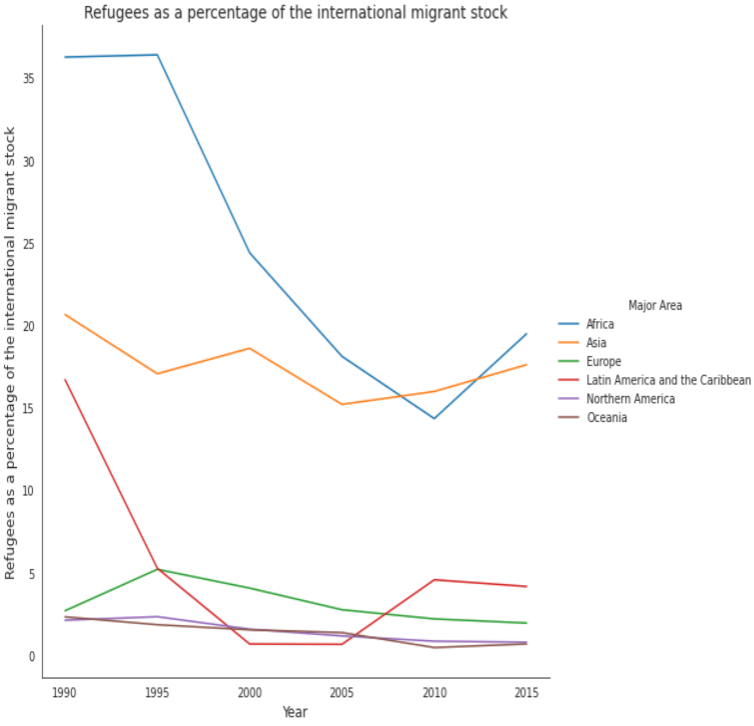


Figure 28

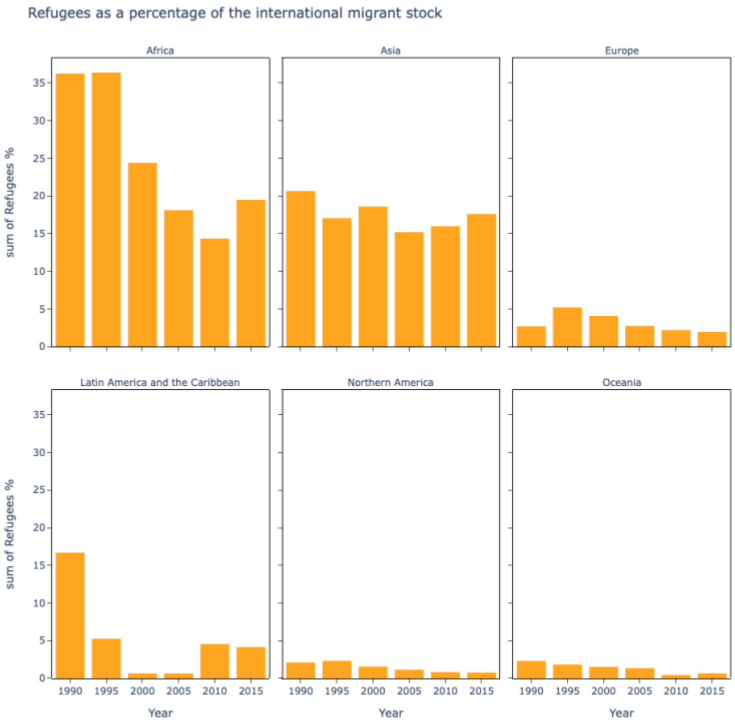


Figure 29

After analyzing the major areas, I created the histograms below that show the top 5 regions with the highest percentage of the international migrant stock in major areas from 1990-2015, we can see that most of the regions are either located in **Africa or Asia**, which matches our results from figure 30. It is also worth mentioning that **Central America** had the highest percentage in the year of **1990**, but the percentage dropped significantly since **1990**, which also alighs perfectly with our findings in figure 29, that the line of **Latin America** started at a relatively high percentage but started to drop since **1990**.

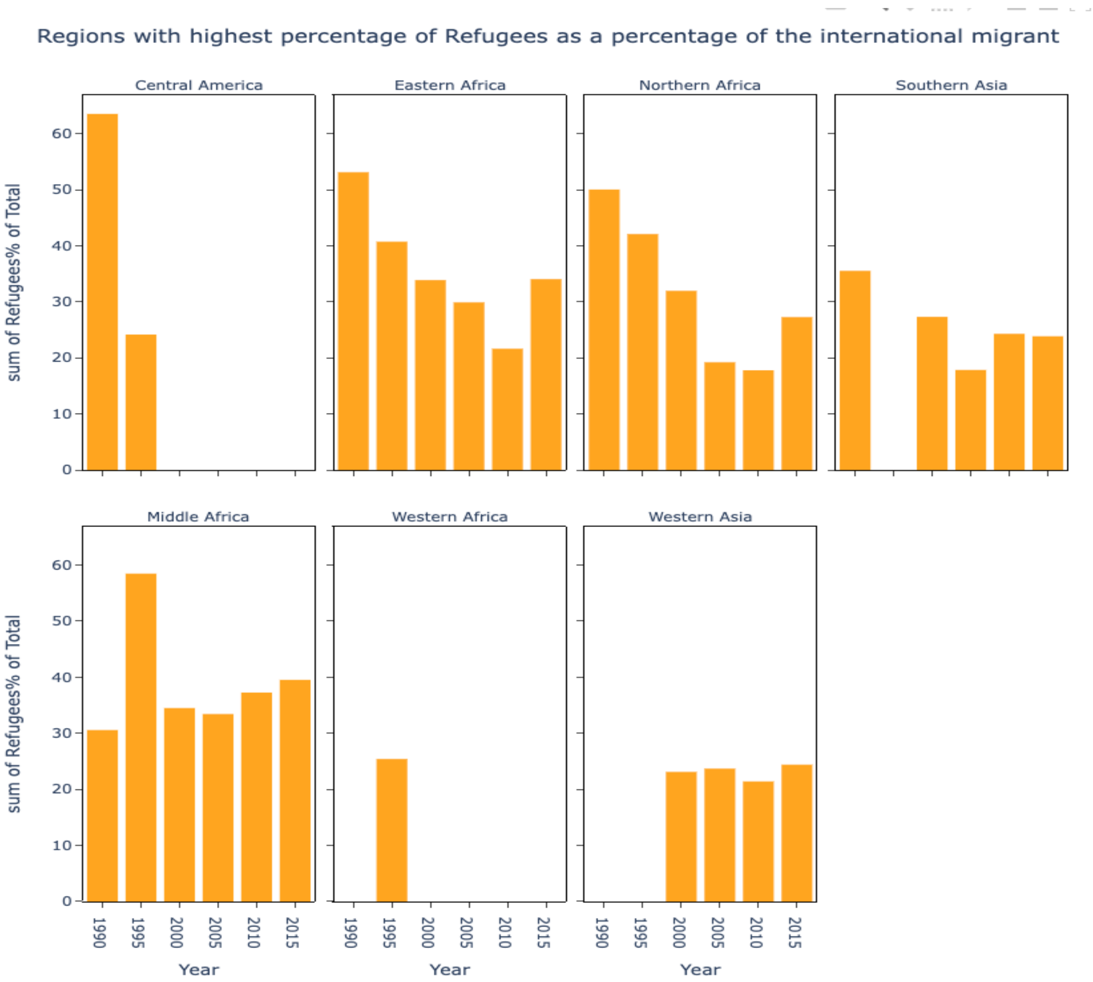


Figure 30

Discussion and Reflection

From the first half of this project, by applying the five tidy data principles to a real messy UN dataset, I began to understand the nature of tidy data. In the second portion of this project, by utilizing various exploratory data analysis(EDA) methods and using Edward Tufte's data visualization principles for the datasets that I cleaned and generated previously, I was able to convert the initial messy datasets into various story-telling figures that can help me further to generate valuable insights.

In detail, for exploratory data analysis methods, I used line charts, histograms, dot plots, violin plots, and small multiple graphs using Seaborn, Matplotlib, and Plotly data visualization libraries. When drafting various figures, I always refer to Tufet's principles.

Firstly, according to Tufet, data visualization needs to be data-intense, design-simple, and word-sized graphics, so when I was trying to make the comparison between the migrant stock among different regions, I chose to use small multiple histograms in Plotly Express. Because small multiples is a great tool to visualize large quantities of data with a high number of dimensions.

Secondly, I realized that the figure I created had a blue-white background, with redundant name labels, and according to Tufet’s principle, we need to erase the non-data ink and keep the visual in a more concise, simple way, so I changed the background into white color and removed the repeated labels.

Thirdly, the ultimate goal of Tufet’s principle is to present the greatest number of ideas in the shortest time with the least ink in the smallest space. Therefore, even though I created more than 80 figures as shown in the google collab notebook, I selected carefully and only choose to use the most effective figures to describe, explore, and summarize the datasets, allowing the audience to understand the ideas communicated with clarity, precision, and efficiency.

Limitations and potential issues

There are certain limitations and potential issues for the final project. Firstly, because I ended up with 36 tidy datasets and I started to generate visualization based on the 36 tidy datasets, I ended up with more than 80 figures as shown in the google collab notebook. Due to the page limitation, I was only able to select the ones that I think illustrate the datasets in the most efficient way; however, there are chances that I could potentially miss including some figures that are worth analyzing. And I may not choose the most suitable type of visualization due to my entry level of knowledge in python visualization.

Conclusions

In conclusion, coming from a marketing background, most of my previous experience with data cleaning and data analysis is either using Excel or SPSS, and this was my first time using Python to conduct data cleaning and data visualization. I felt lucky that utilizing the knowledge that I learned in class and with the help of the professor and the super awesome TAs, I was able to conduct thorough data cleaning and data visualization procedures for the messy UN datasets.

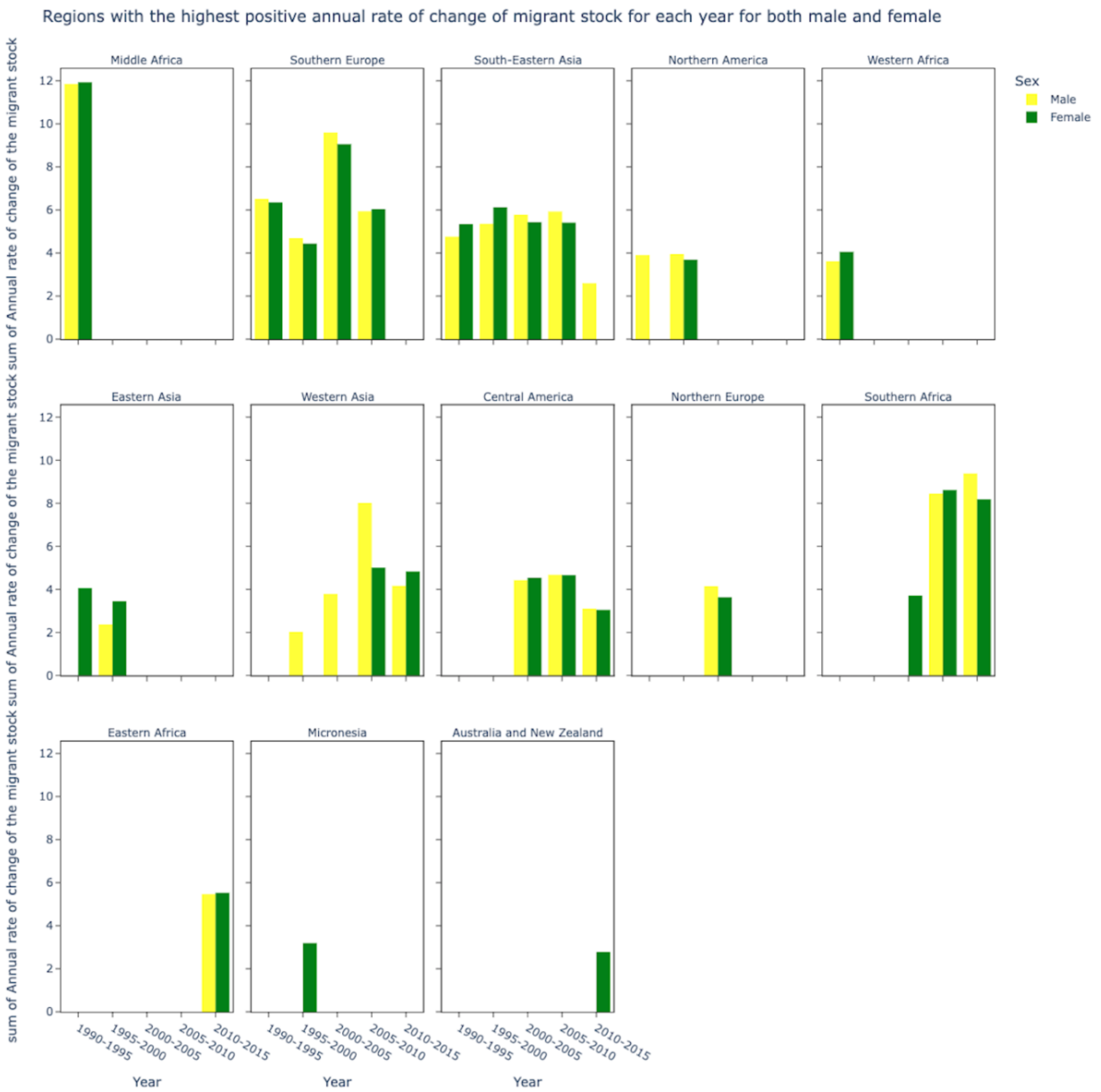
My key takeaway from this class and from this project is that data science is not an individual, simply a one-on-one process, instead, it is a collective and reflexive practice. Therefore, I am eager to keep learning and build my competency along the way.

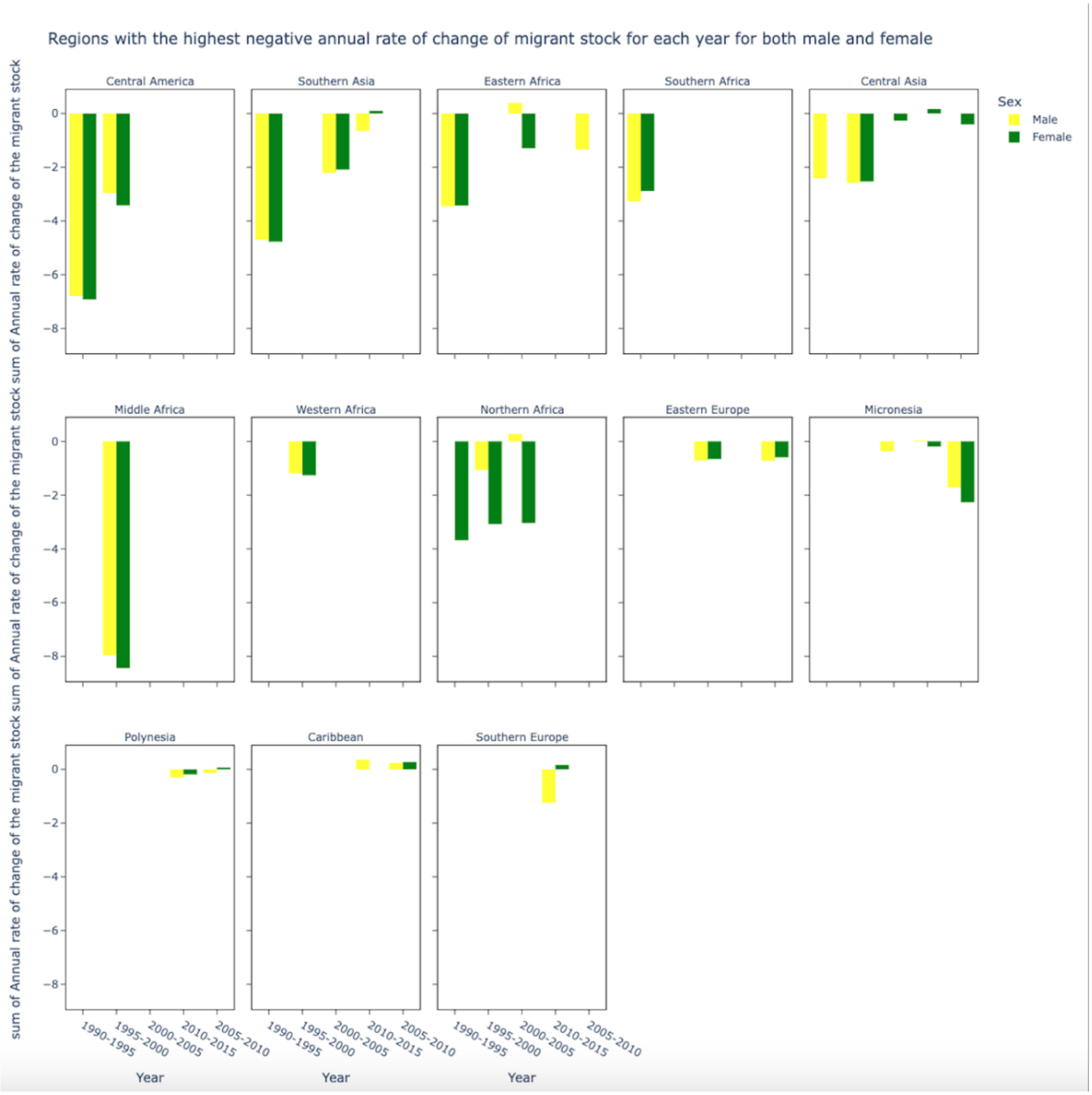
Reference

[1] “Graphical Excellence | Tips on Creating Charts.” https://www.qimacros.com/free-excel-tips/graphical-excellence/ (accessed Dec. 15, 2022).

[2] “What is Exploratory Data Analysis? | IBM.” https://www.ibm.com/cloud/learn/exploratory-data-analysis (accessed Dec. 15, 2022).

Appendix


Regions with the highest positive annual rate of change of migrant stock for each year for both male and female
Middle Africa
Southern Europe
South-Eastern Asia
Northern America
Western Africa
Eastern Asia
Western Asia
Central America
Northern Europe
Southern Africa
Eastern Africa
Micronesia
Australia and New Zealand
Sex
Male
Female
sum of Annual rate of change of the migrant stock
0
2
4
6
8
10
12
1990-1995
1995-2000
2000-2005
2005-2010
2010-2015
Year


Regions with the highest negative annual rate of change of migrant stock for each year for both male and female
Central America
Southern Asia
Eastern Africa
Southern Africa
Central Asia
Middle Africa
Western Africa
Northern Africa
Eastern Europe
Micronesia
Polynesia
Caribbean
Southern Europe
Sex
Male
Female
sum of Annual rate of change of the migrant stock
0
−2
−4
−6
−8
1990-1995
1995-2000
2000-2005
2010-2015
2005-2010
Year