INF 1340 Final Project

Bohao Cui
1009314467
10 December 2022

Introduction

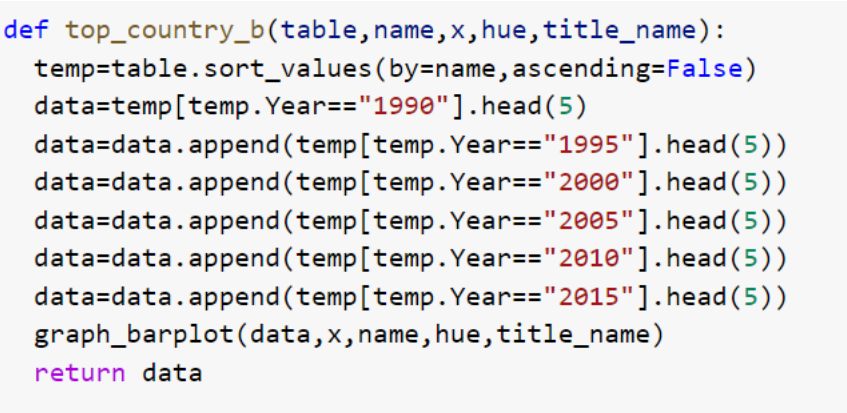
In this project, we will continue to work with the UN_MigrantStockTotal 2015.xlsx dataset. This dataset records the immigrant population and total population changes of 232 countries from 1990-2015. And according to geographical location, it is divided into 6 major areas, and each major area is further divided according to the direction. So, there are 22 region areas in total. In previous projects, we have cleaned data based on tidy data principle and got 36 new tables. So, the main showcase of this project is to explore these 36 tables using EDA (Exploratory Data Analysis).

In this report, the following questions are mainly explored:

- What has the migration of the global population looked like over the past 25 years?
- How has global population changed over the past 25 years?
- How has the proportion of the immigrant population in the local population changed over the past 25 years?
- Where are the refugees concentrated? How many refugees are included in the immigrant stock?

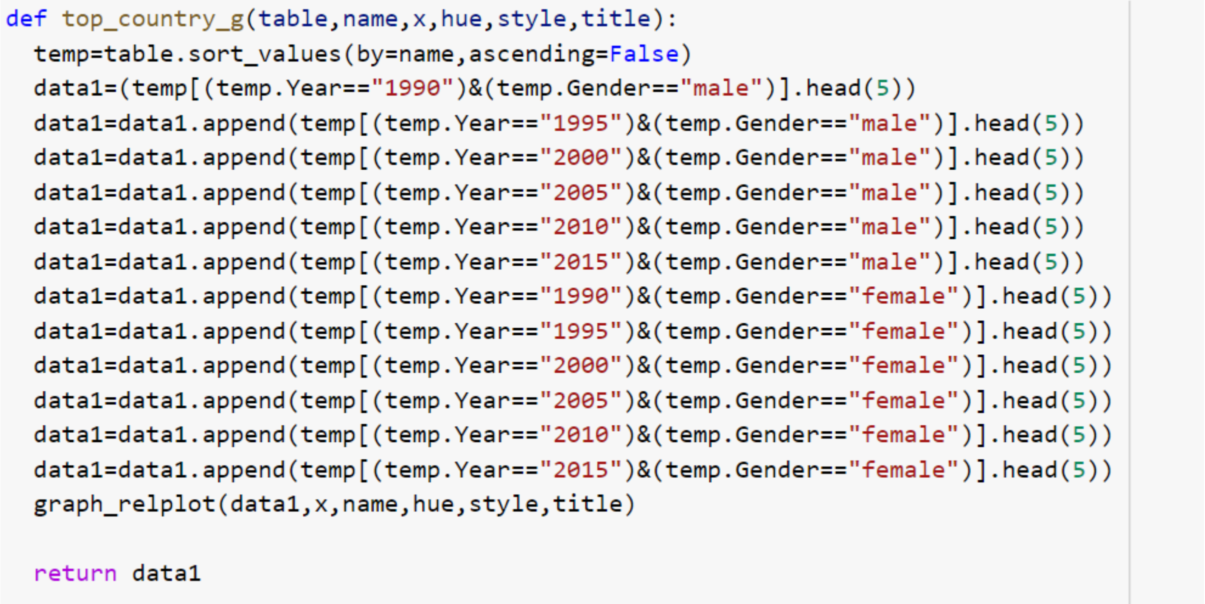
Python Method

The top_country_b function will combine the top five country data based on each year's data. Then call the graph_barplot function to make a histogram.



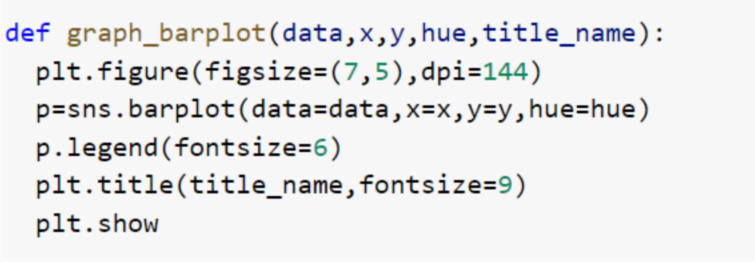
```
def top_country_b(table,name,x,hue,title_name):
  temp=table.sort_values(by=name,ascending=False)
  data=temp[temp.Year=="1990"].head(5)
  data=data.append(temp[temp.Year=="1995"].head(5))
  data=data.append(temp[temp.Year=="2000"].head(5))
  data=data.append(temp[temp.Year=="2005"].head(5))
  data=data.append(temp[temp.Year=="2010"].head(5))
  data=data.append(temp[temp.Year=="2015"].head(5))
  graph_barplot(data,x,name,hue,title_name)
  return data
```

The top_country_g function is an added part based on top_country_b, which can filter out the top five countries for men and women according to the year.



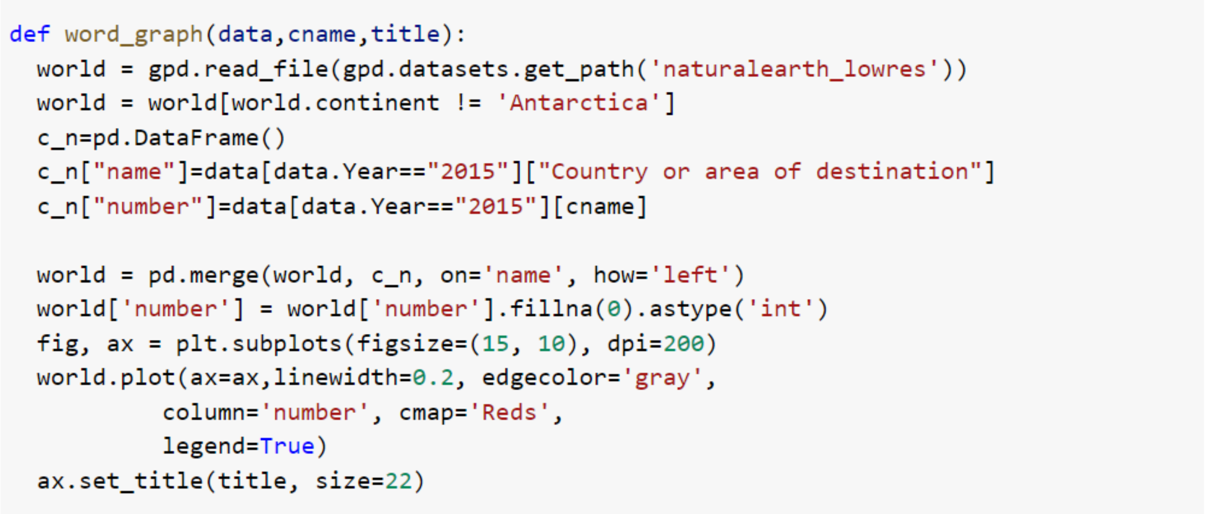
```
def top_country_g(table,name,x,hue,style,title):
  temp=table.sort_values(by=name,ascending=False)
  data1=(temp[(temp.Year=="1990")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="1995")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="2000")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="2005")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="2010")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="2015")&(temp.Gender=="male")].head(5))
  data1=data1.append(temp[(temp.Year=="1990")&(temp.Gender=="female")].head(5))
  data1=data1.append(temp[(temp.Year=="1995")&(temp.Gender=="female")].head(5))
  data1=data1.append(temp[(temp.Year=="2000")&(temp.Gender=="female")].head(5))
  data1=data1.append(temp[(temp.Year=="2005")&(temp.Gender=="female")].head(5))
  data1=data1.append(temp[(temp.Year=="2010")&(temp.Gender=="female")].head(5))
  data1=data1.append(temp[(temp.Year=="2015")&(temp.Gender=="female")].head(5))
  graph_relplot(data1,x,name,hue,style,title)

  return data1
```

The graph_barplot function makes a histogram based on the imported data table, not just the x and y axes. We can also use a different color to represent a new variable.



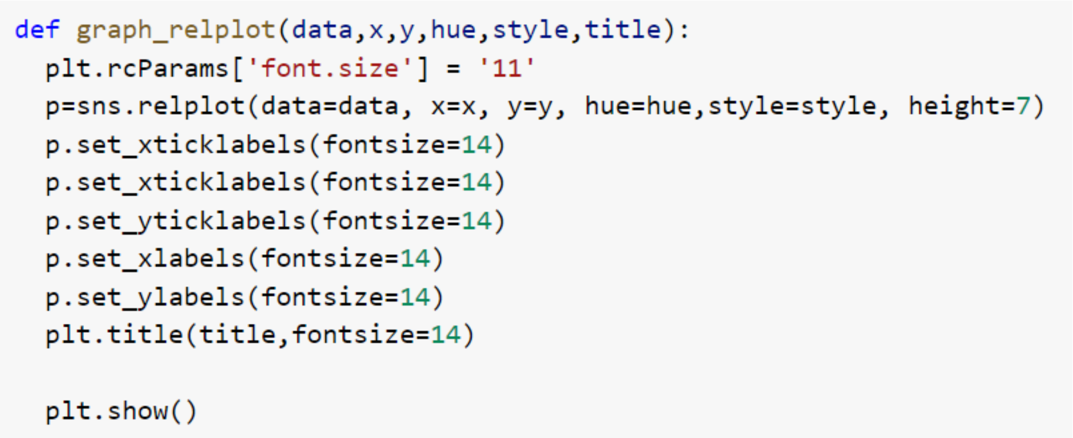
```
def graph_barplot(data,x,y,hue,title_name):
  plt.figure(figsize=(7,5),dpi=144)
  p=sns.barplot(data=data,x=x,y=y,hue=hue)
  p.legend(fontsize=6)
  plt.title(title_name,fontsize=9)
  plt.show
```

The word_graph function will generate a world map, and different countries will be displayed in different colors in the map according to the inserted data. Antarctica has no countries, so the map will not include Antarctica.



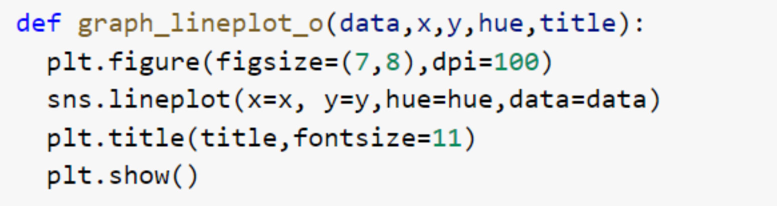
```
def word_graph(data,cname,title):
  world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
  world = world[world.continent != 'Antarctica']
  c_n=pd.DataFrame()
  c_n["name"]=data[data.Year=="2015"]["Country or area of destination"]
  c_n["number"]=data[data.Year=="2015"][cname]

  world = pd.merge(world, c_n, on='name', how='left')
  world['number'] = world['number'].fillna(0).astype('int')
  fig, ax = plt.subplots(figsize=(15, 10), dpi=200)
  world.plot(ax=ax,linewidth=0.2, edgecolor='gray',
          column='number', cmap='Reds',
          legend=True)
  ax.set_title(title, size=22)
```

The graph_replot function is to make a dotted graph, which can display four different variables, first the x-axis and y-axis, and then display the other two variables according to different colors and shapes of different points.

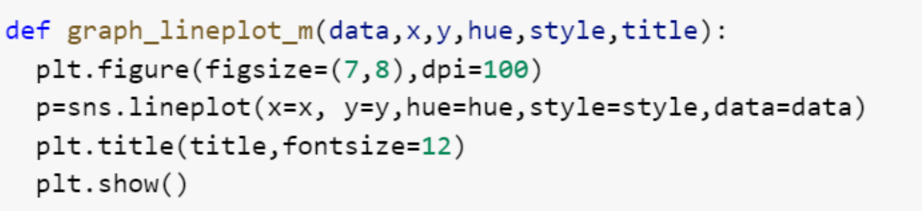


```
def graph_relplot(data,x,y,hue,style,title):
  plt.rcParams['font.size'] = '11'
  p=sns.relplot(data=data, x=x, y=y, hue=hue,style=style, height=7)
  p.set_xticklabels(fontsize=14)
  p.set_xticklabels(fontsize=14)
  p.set_yticklabels(fontsize=14)
  p.set_xlabels(fontsize=14)
  p.set_ylabels(fontsize=14)
  plt.title(title,fontsize=14)

  plt.show()
```

The graph_lineplot function is to make a line graph, which can display three different variables. For example, the x-axis represents the year, the y-axis represents the quantity, and the lines of different colors represent countries. The only difference between the graph_lineplot_m function and graph_lineplot_o is that it can also use dashed lines to represent new variables. A gender variable can be added on the basis of the previous example.



```
def graph_lineplot_o(data,x,y,hue,title):
  plt.figure(figsize=(7,8),dpi=100)
  sns.lineplot(x=x, y=y,hue=hue,data=data)
  plt.title(title,fontsize=11)
  plt.show()
```



```
def graph_lineplot_m(data,x,y,hue,style,title):
  plt.figure(figsize=(7,8),dpi=100)
  p=sns.lineplot(x=x, y=y,hue=hue,style=style,data=data)
  plt.title(title,fontsize=12)
  plt.show()
```

Since each table contains a large amount of data, full visualization is not feasible because the large amount of data would make the whole picture illegible. This also violates Tufte's visualization principles 1: Communicate complex ideas in a clear, accurate and efficient manner. So, for the country data, we will show and analyze the top five countries in order of data size.

We will follow Principle 2: Presentation of Well-Designed Matter, Statistics and Design Data. Use different charts for different data, such as using a line chart so that the audience can see the data changes and trends of 25 years, and with a world map, the audience can have a clear understanding of the geographical location of the data. We also followed principle 3: generate the most ideas in the smallest space, we used not only the x and y variables on the 2D number line, but also increased the variables through different colors and shapes. While ensuring principle 1, it is also possible to allow the audience to compare the overall data as much as possible.

What has the migration of the global population looked like over the past 25 years?

From Figure 1, we can analyze that in the past 25 years, the main areas of immigration were Europe, Asia and North America. Europe has the highest stock of international migrants, and it is still rising. Asia is ranked second, but the rise is faster than Europe. As a result, the gap in international migrant stocks between the two countries continued to narrow and was close in 2015. North America is also maintaining an upward trend, but the growth rate has been slow in the past 10 years. The other three continents have not changed much and have remained stable. Of the three main continents, Europe and North America have a larger international migrant stock of women than men, and the gap is growing. Asia is the exact opposite, with the steepest growth slope for the male international migrant stock. From Figure 2, we can see the 5 countries with the largest number of international migrants. The highest is the United States, even surpassing the sum of the other four countries with the highest stocks in 2015. And it has remained number one for the past 25 years. The second place is Russia, but it has been overtaken by Germany in the last ten years. Figure 3 shows the stock of international migrants for male and female, and there is no obvious difference in the proportion of men and women. Figure 4 shows the distribution of international immigrant populations in 2015. The darker the color, the greater the number of international immigrant populations.

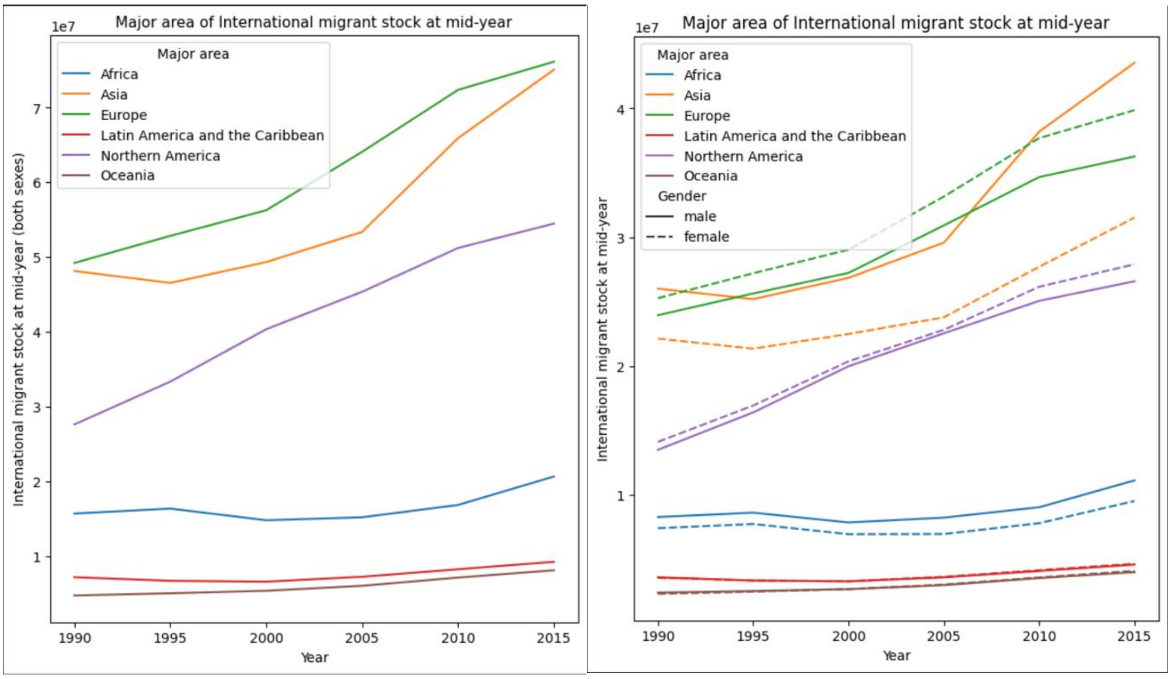


Figure 1: Major area of international migrant stock at mid-year

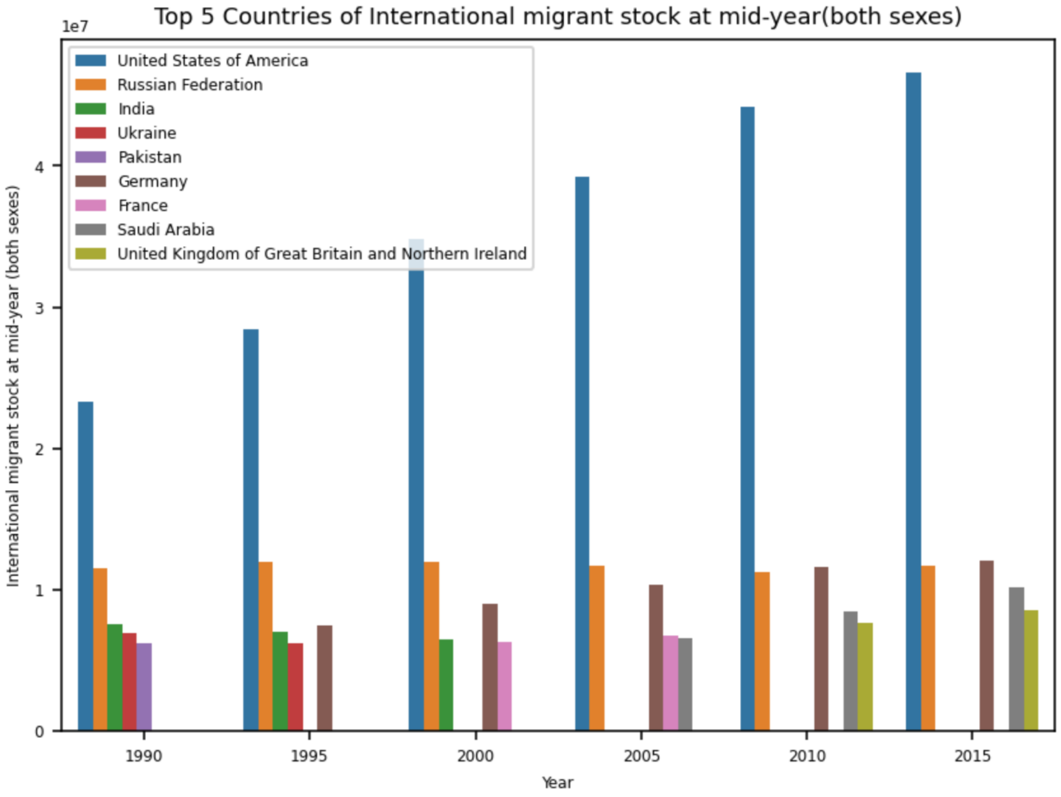


Figure 2: Top 5 Countries of international migrant stock at mid-year (both sexes)

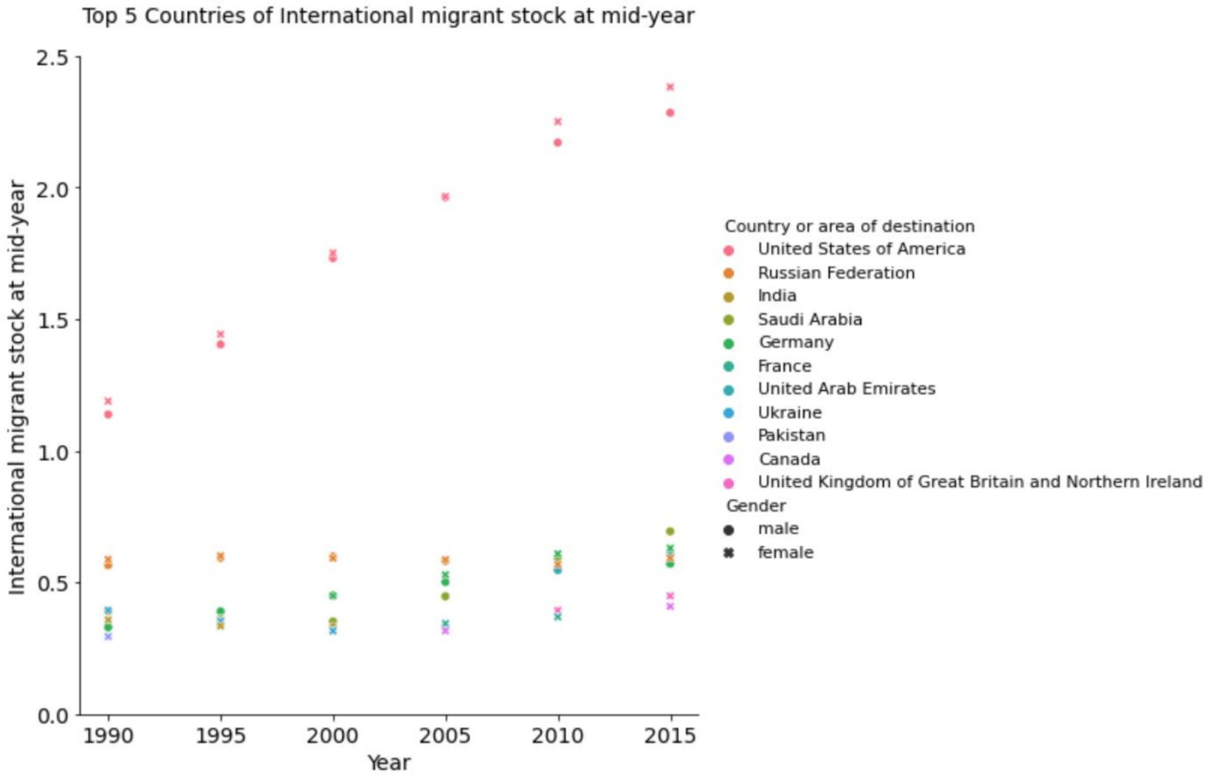


Figure 3: Top 5 Countries of international migrant stock at mid-year1

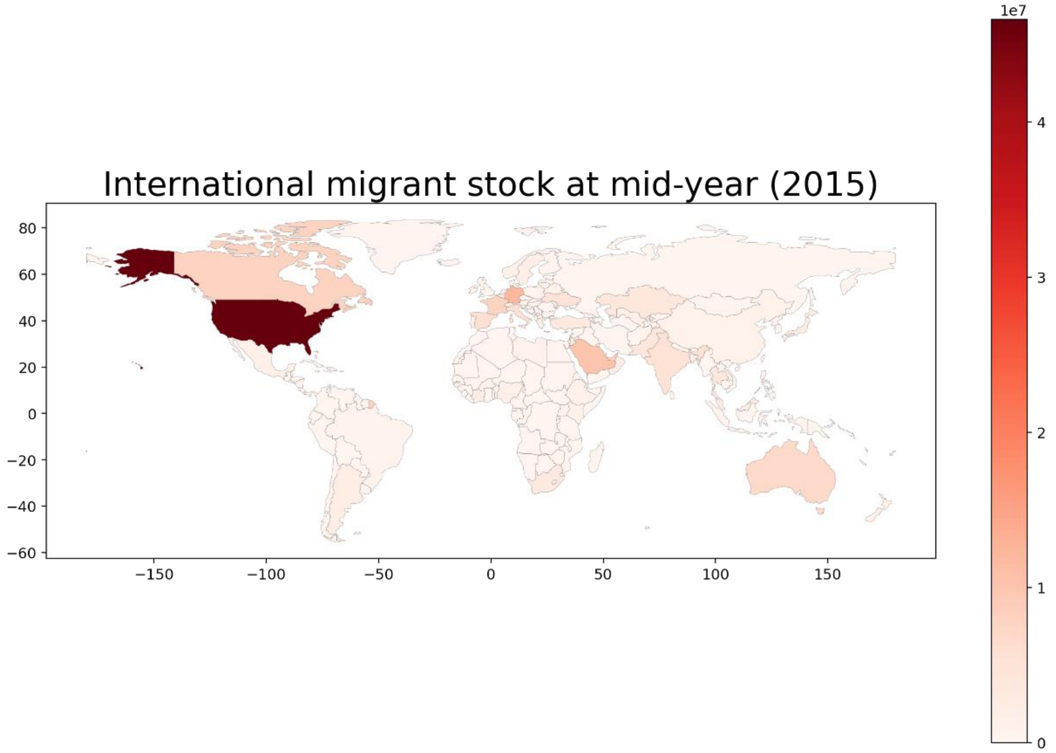


Figure 4: international migrant stock at mid-year (2015)

How has global population changed over the past 25 years?

From Figure 5, we can analyze that in the past 25 years, the total population of Asia has always been the largest and has maintained the largest increase. And only Asia has always had more males than females, and there is no significant difference between males and females in other regions. The populations of the other five major regions are all below one billion. Only Africa has shown a clear upward trend in the past ten years, and the total population of the remaining regions has not changed significantly. Figure 6 shows the five countries with the largest populations over the past 25 years, and the rankings have remained constant. The most populous country is China, followed by India. It can be seen from the figure that the population gap between China and India is constantly narrowing. Comparing Figure 7, it can be seen that both China and India have an imbalance in the ratio of men to women, and the male population has always been larger than the female population. And the gap has widened over the years. The total population of the remaining three countries has remained basically unchanged over the years, and the ratio of male to female is balanced. Figure 8 uses a world map to show the population distribution map. It can be seen more clearly that the world's population is mainly concentrated in Asia.

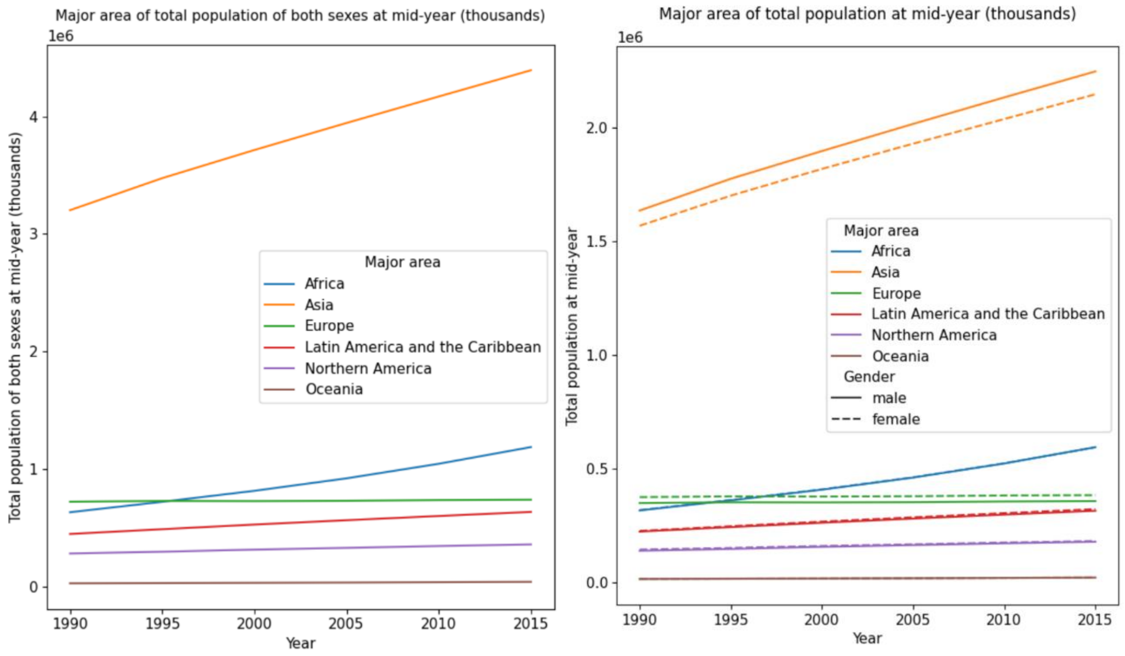


Figure5: Major area of total population at mid-year (thousands)

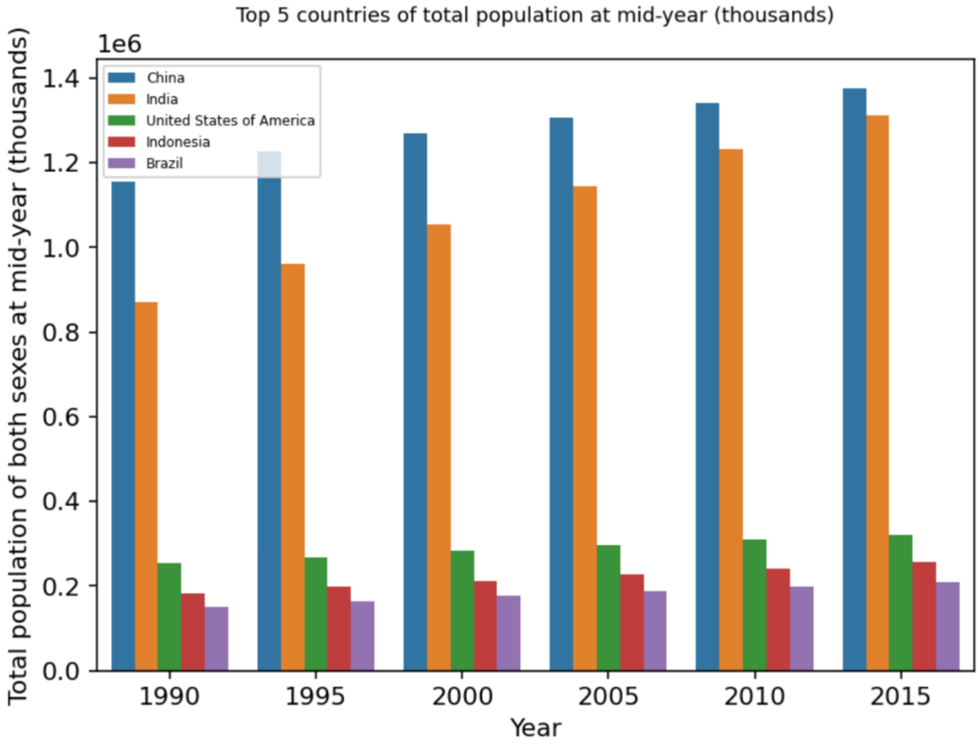


Figure6: Top 5 countries of total population at mid-year (thousands)

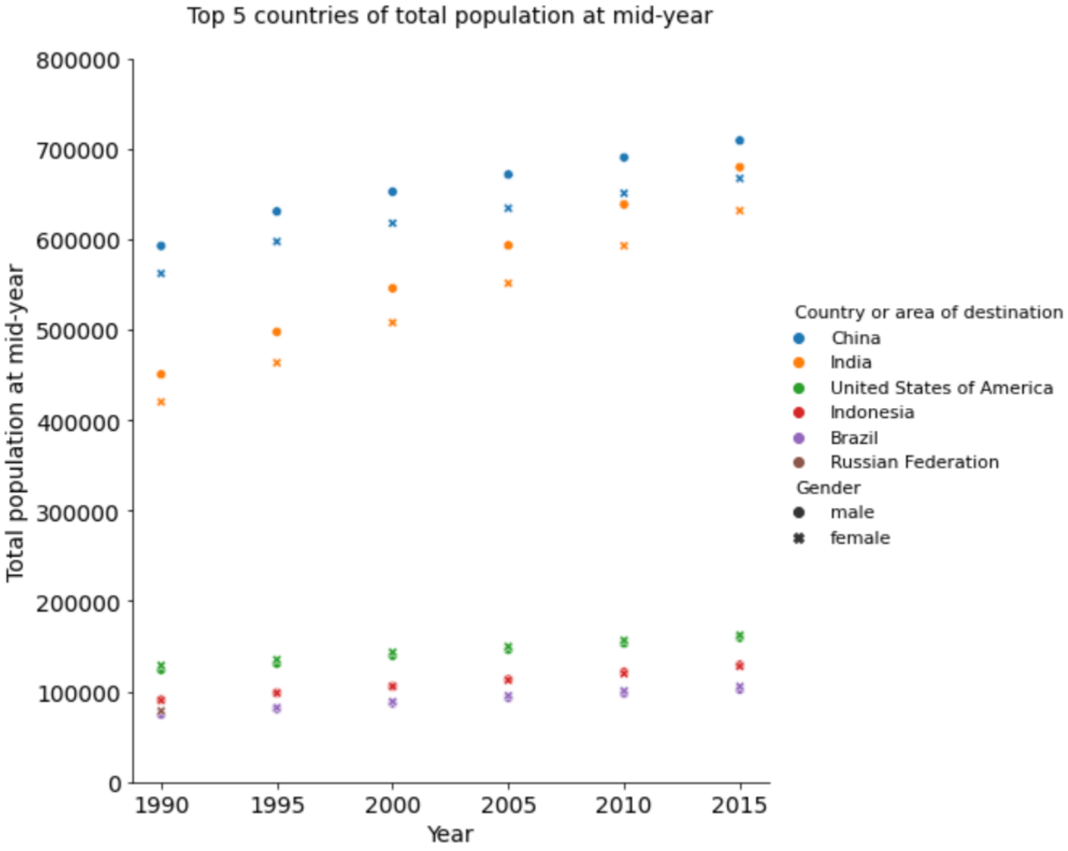


Figure7: Top 5 countries of total population at mid-year (thousands)

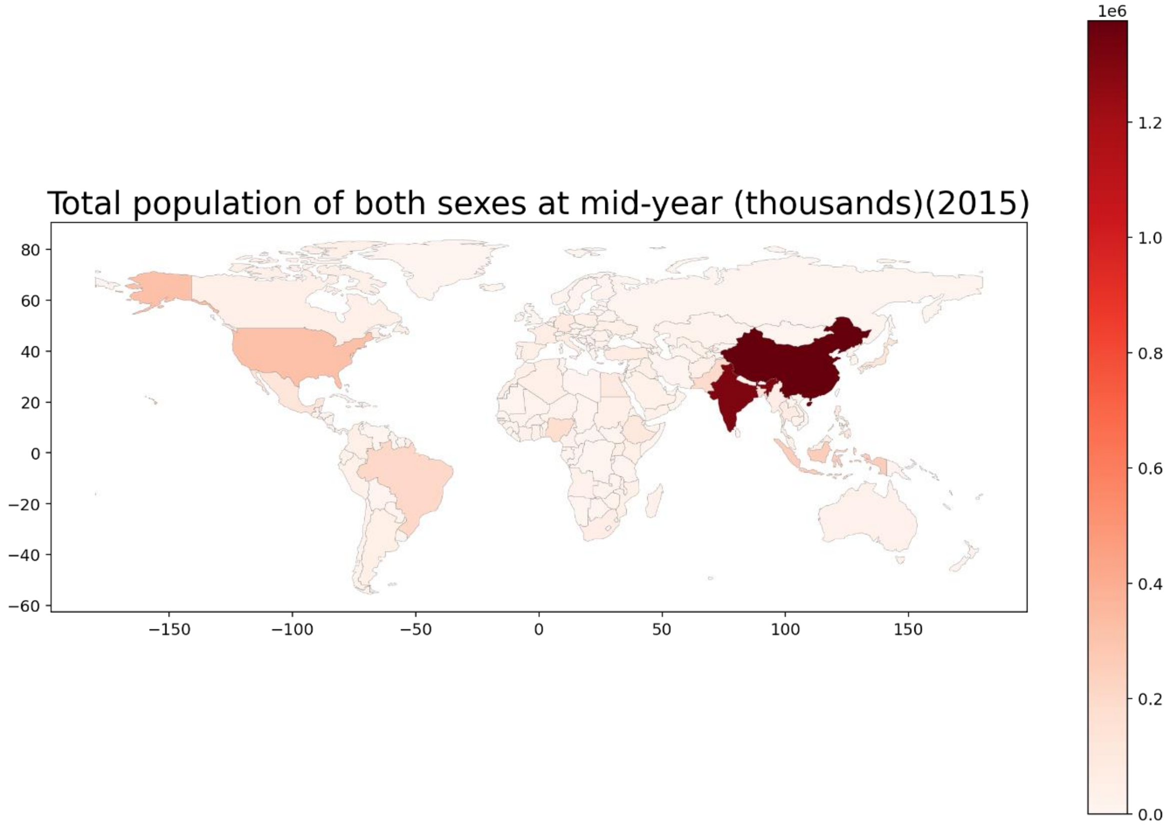


Figure8: Total population at mid-year in 2015 (thousands)

How has the proportion of the immigrant population in the local population changed over the past 25 years?

From Figure 9, we can see that in the past 25 years, the stock of international immigrants in Asia accounted for the largest percentage of the total population, and it rose from 17.5% in 1990 to 20% in 2005. North America and Europe are ranked second and third and also maintain an upward trend. The stock of international immigrants in the remaining three major regions is less than 2.5% of the total population and maintains a downward trend. Figure 10 presents the five countries with the highest immigrant stocks as a percentage of total population. The Holy See is basically an immigrant population with almost no locals. In the United Arab Emirates, 70% of the population is from immigrants. The remaining countries also have at least 50% of their populations as international immigrants rather than locals.

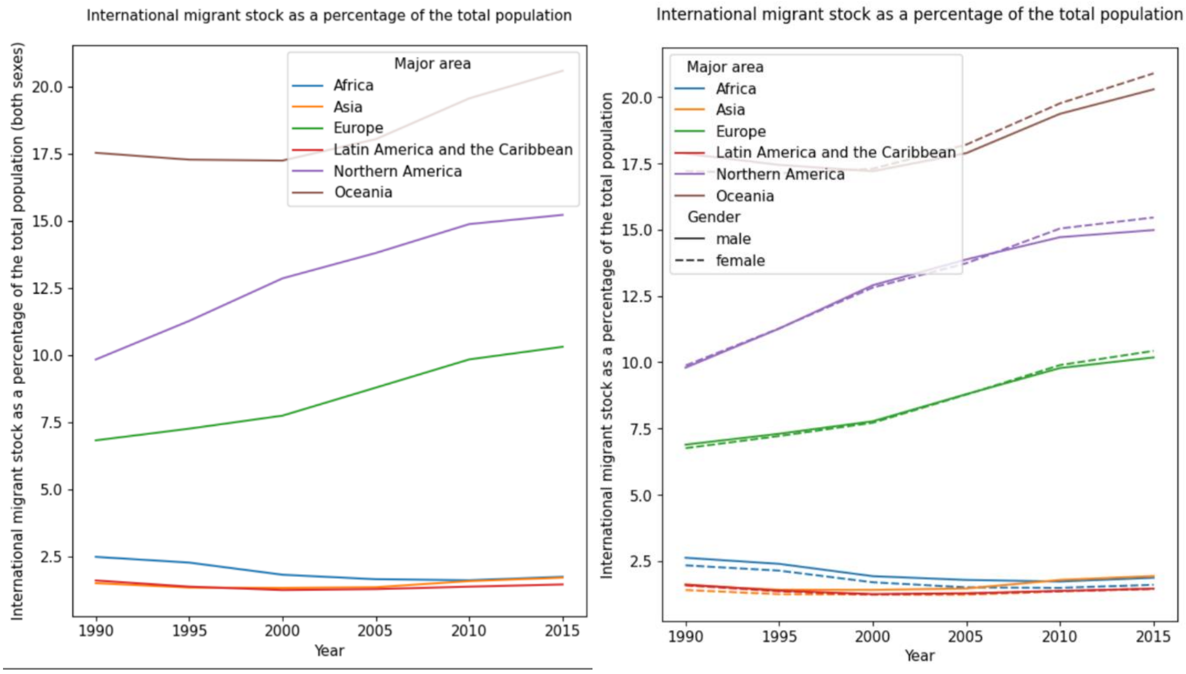


Figure9: International migrant stock as a percentage of the total population

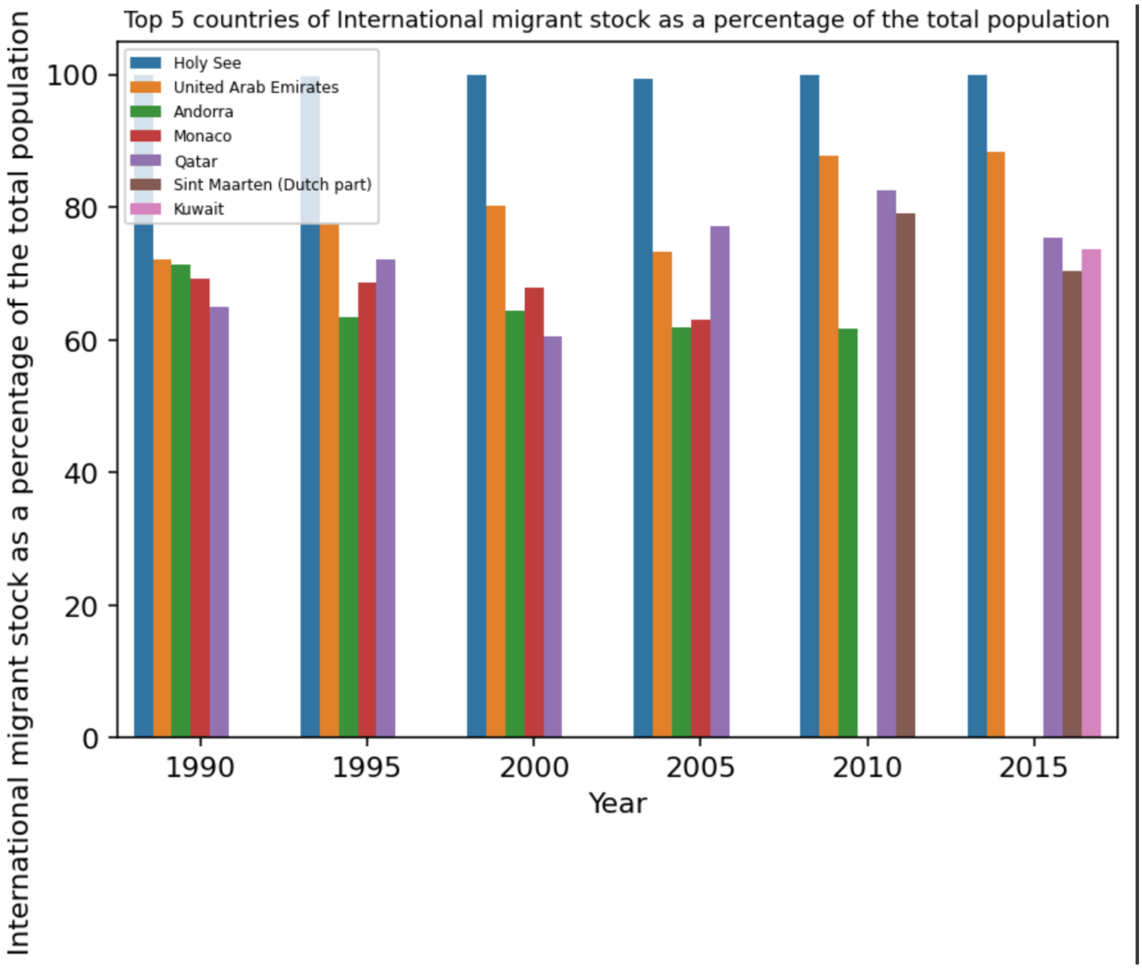


Figure10: Top 5 countries of International migrant stock as a percentage of the total population

Where are the refugees concentrated? How many refugees are included in the immigrant stock?

Figure 11 shows the changes in refugee stocks by continent from 1990 to 2015. From the figure, we can see that Asia has always been the continent with the largest refugee stock, especially after 2005, the refugee stock has risen sharply. In contrast, refugee stocks in the other five major regions all maintained a downward trend. Only in Africa has the refugee stock increased since 2010. Combined with Figure 12, it can be found that the curve of refugee stock is similar to the percentage ratio of immigrants. In 1990, 40% of Africa's immigrant stock was refugees, and by 2015 it was nearly 20%, the highest compared to other continents. Asia's immigrant stock has remained around 20% refugees. The remaining continents are only about 5 percent. Figure 13 shows the five countries with the largest refugee stock. The refugee stock has gradually declined since 1990, but it is still concentrated in three countries: Iran, Jordan and Pakistan. The boxplot in Figure 14 confirms this, with the three countries having the highest medians. And we can also see that the data of several countries in the chart are represented by a line, which means that in the past 25 years, their refugee stock has only been ranked in the top 5 once. The red area in Figure 15 shows the regions with the largest refugee stock in 2015.

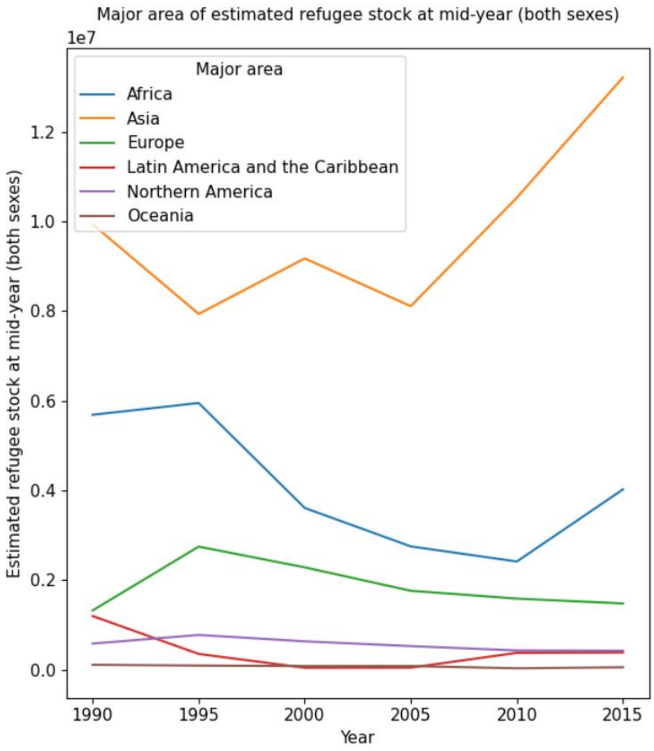


Figure11: Major area of estimated refugee stock at mid-year

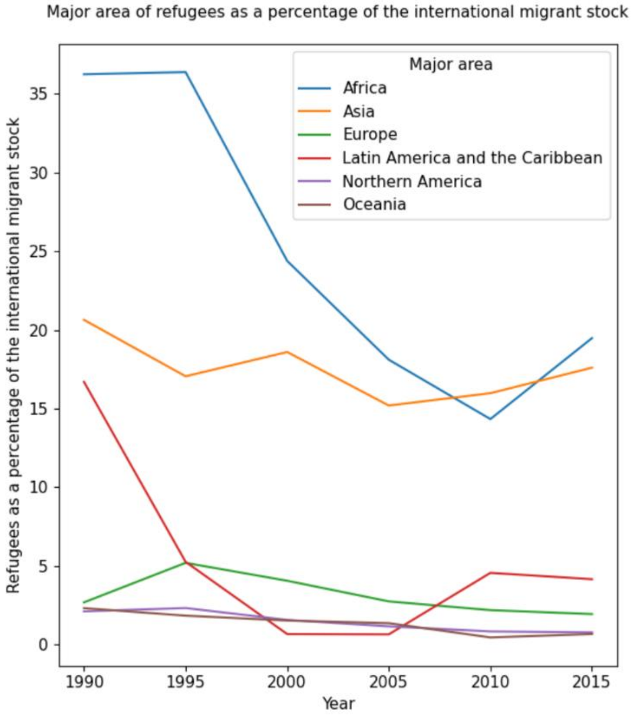


Figure12: Major area of refugees as a percentage of the international migrant stock

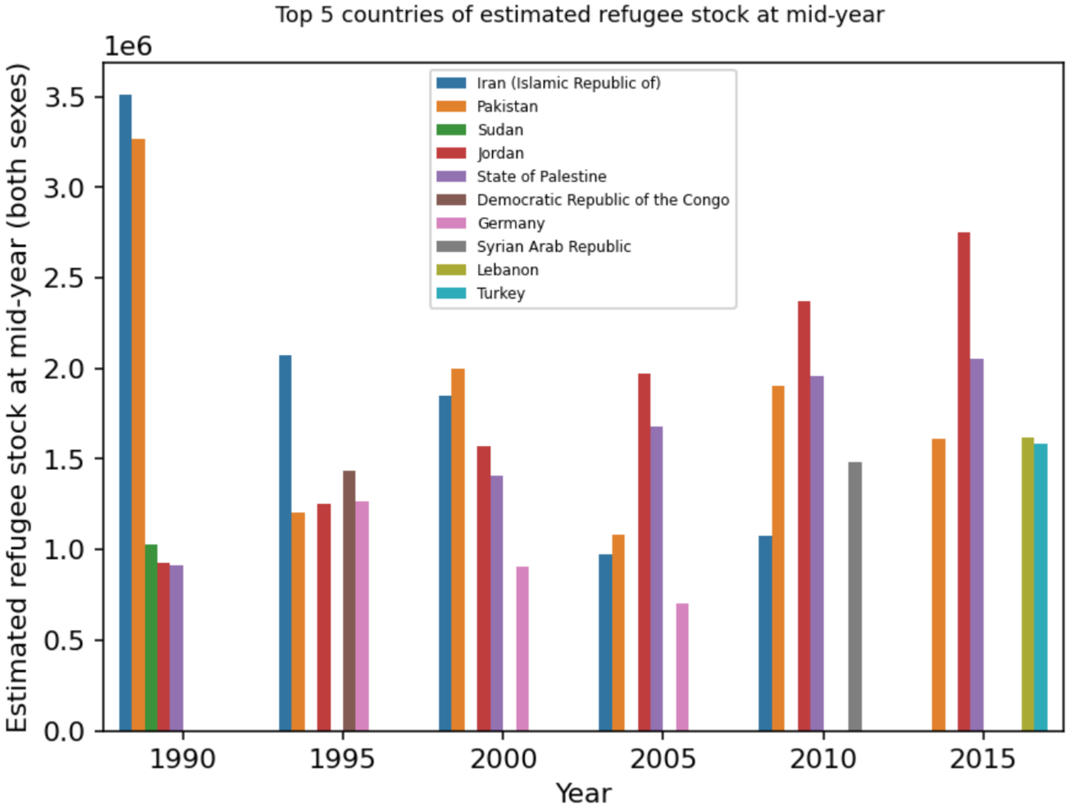


Figure13: Top 5 countries of estimated refugee stock at mid-year

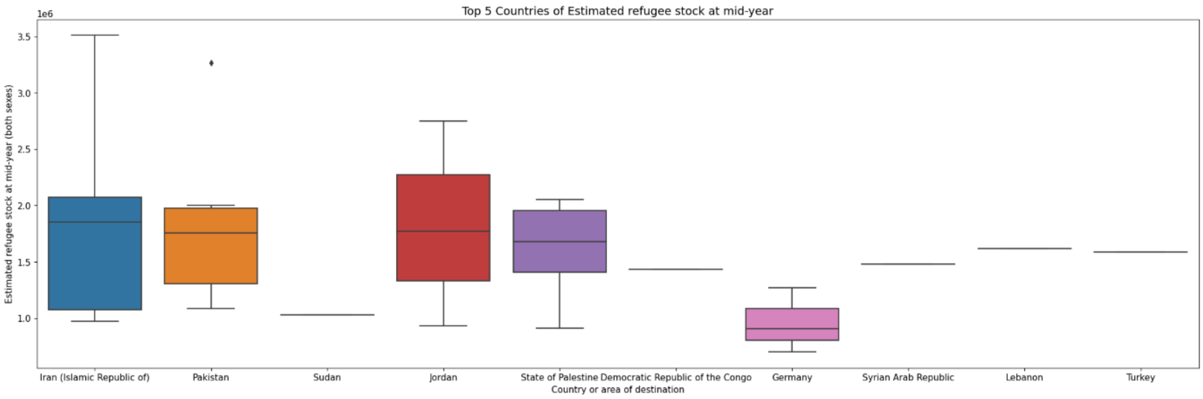


Figure14: Top 5 countries of estimated refugee stock at mid-year

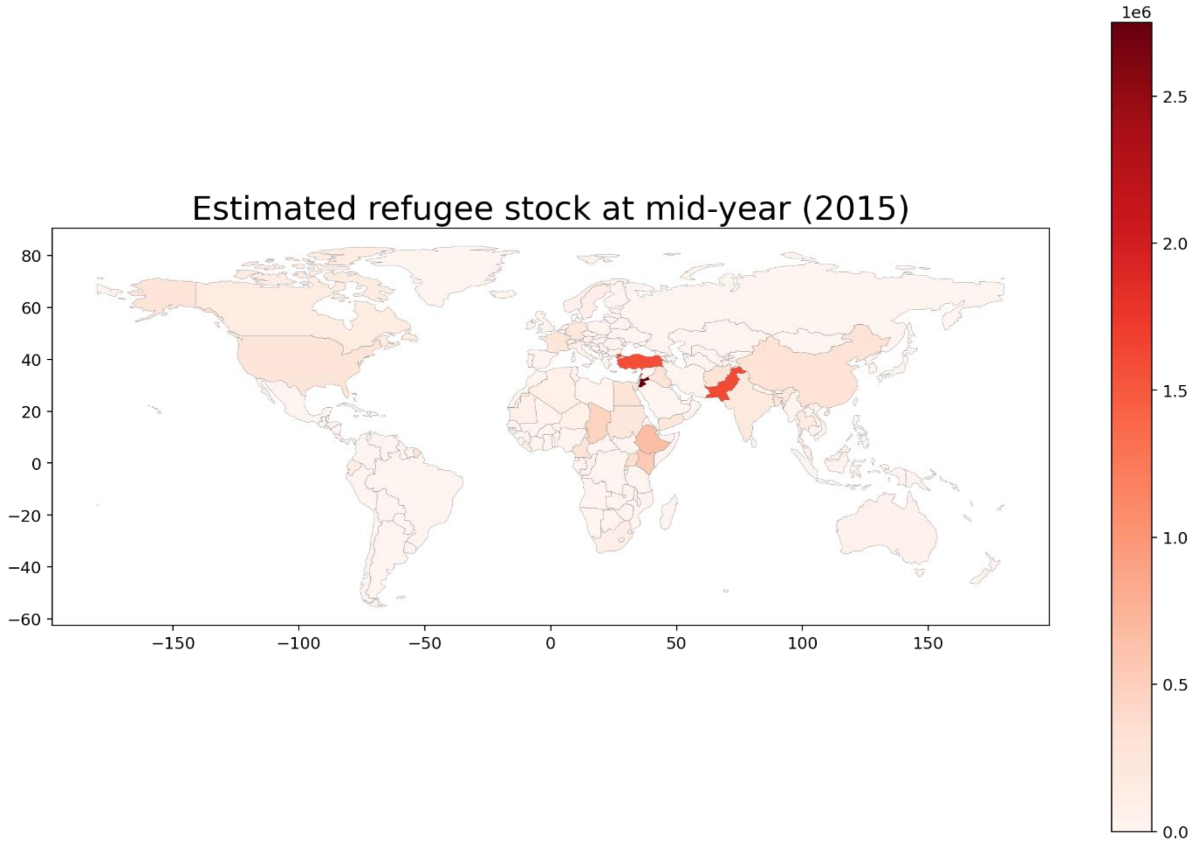


Figure15: Estimated refugee stock at mid-year (2015)

Conclusion & Discussion

Using the EDA process to process UN_MigrantStockTotal data, we can clearly observe the structure and characteristics of the data. For example, use a line chart to clearly show the increase and decrease trend of the annual immigrant stock in each major area. Using the histogram, we can clearly see the size of the total population of the country according to the length, and it is easy to compare the differences between the populations of countries.

According to the data visualization, we can see from the picture which countries have the largest number of immigrants and which countries have the largest number of refugees. If we only search from the table, it will take a lot of time and it is difficult to display it to the user.

Data visualization frees users from focusing on specific data and text. Users can know what data that the picture will display according to the title, and then compare the changes in the graph and coordinate axes to understand the changes in the data.