Yang Xuan (Student ID: 1003907427)
INF 1340
2022/12/09
Final Data Visualization Project Writeup

1. Introduction

In the digital age, where almost everything is data-driven, human-centered data science encourages data analysts to incorporate more socially responsible and ethical aspects into data analysis. Empirically, data analysis consists of two essential elements, including data cleaning and data visualization. Mathematician John Tukey brought up the concept of Exploratory Data Analysis (EDA) in 1977. Tukey has argued that EDA is a collection of approaches to analyzing datasets to explore their characteristics by using data visualization methods (Leinhardt, 1980).

Data visualization could capture the information and trends that may be hard to notice in sheets and forms. American statistician Edward Tufte's 6 principles for Graphic Integrity have served as a guide for data visualization in statistics. He mentioned that accurate data visualization should include different elements to compare and demonstrate variables' differences and potential causality. Data scientists should provide documentation to graphs for the credibility and context of any project. Tufte also argues that "a well-designed presentation of data should consist of complex ideas but communicate with clarity, precision, and efficiency (Tufte, 2001)."

Based on the data cleaning and some preliminary exploration of the dataset, I learned that the data in each sheet are interconnected. Hence, I plan to merge Table 1 and Table 2 to examine the correlation between international migrant stock and total population. I also plan to merge Table 6A and Table 6B to see if I can derive a trend of refugee flow. The rest tables will be explored as a single sheet. I will conduct further data visualization to examine the potential trends, correlations, and other significant information by incorporating Tufte's principles of Graphic Integrity and Graphic Excellence principles. The following content includes Methodology & Results, Discussion & Interpretation, and Conclusion.

2. Methods & Results

2.1 Package Importing and Data Preparation

Before starting the visualization process, I imported the packages I would need to use, including *datascience* package, *matplotlib, plotlyexpress, seaborn*, and *pandas*. Then I imported Table 1 as a data frame to check the tables' columns and content. Since only numeric variables can be visualized in python, I changed the datatype of 'International Migrant Stock at Mid-year' from string to float data. In this way, all statistical methods could perform on the dataset.

```
# Load packages
from datascience import *
import matplotlib
matplotlib.use('Agg', warn=False)
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
import warnings
warnings.simplefilter(action="ignore", category=FutureWarning)
import plotly.express as px
```

```
<ipython-input-215-4d034789aae3>:4: MatplotlibDeprecationWarning: The 'warn' pa
  matplotlib.use('Agg', warn=False)
```

2.2 Table 1 & 2 EDA: International Migrant Stock / Total Population at Mid-year

2.2.0 Table 1 Preliminary Sorting & Exploration

Table 1 contains international migrant stock at mid-year (by sex) for all the observations. I intend to create a bar chart to visualize which regions have the most international migrant stock by using *matplotlib*. After I changed the datatype for Table 1, I saved Table 1 again as a *.csv* file and read in the table by using the *datascience.tables* function. I used the *table.sort* function to operate on *.csv* files instead of a data frame. I sorted Table 1 by the value of international migrant stock in descending order. The results showed that the United States had had the most international migrant stock for the past two decades, which seemed very accurate.

```
[224] # Save a new table of data sorted by international migrant stock at mid-year
      sorted_by_migrantstock = table1_EDA.sort('International Migrant Stock at Mid-year', descending=True)
```

```
[226] sorted_by_migrantstock
```

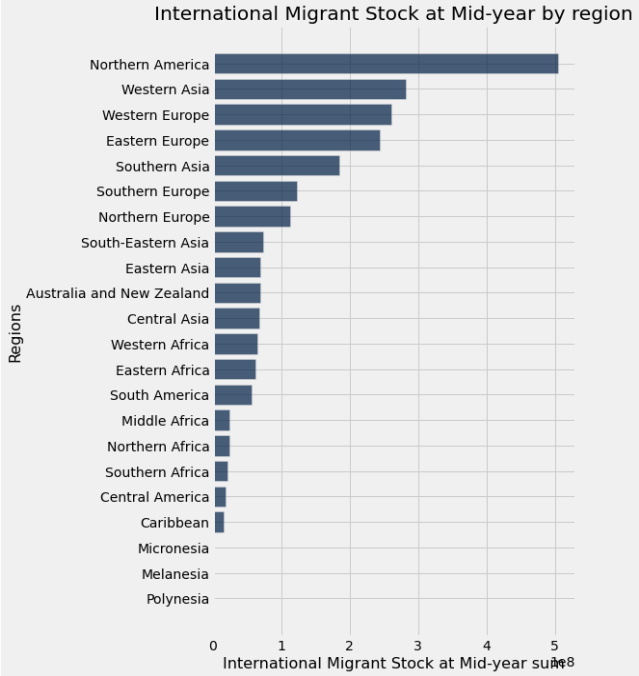
Unnamed: 0	Unnamed: 0.1	Regions	Country of destination	Country Code	Year	International Migrant Stock at Mid-year	Sex
1368	1561	Northern America	United States of America	840	2015	4.66271e+07	Both
1136	1296	Northern America	United States of America	840	2010	4.41836e+07	Both
904	1031	Northern America	United States of America	840	2005	3.92583e+07	Both
672	766	Northern America	United States of America	840	2000	3.48141e+07	Both

2.2.1 Table 1 Matplotlib: Bar Chart of International Migrant Stock at Mid-year by region

Following Tufte's first principle of data comparison, I created a bar chart to visualize a descending bar chart of international migrant stock at mid-year by region. I grouped the sorted data by region and added the value together for each region. A new column of 'International Migrant Stock at Mid-year sum' was generated. Next, I used the *table.barh* function to visualize the 'Regions' and ''International Migrant Stock at Mid-year sum' columns.

Visually, Figure 1 corresponded to the sorted data, with Northern America having the highest international migrant stock.

Figure 1 International Migrant Stock at Mid-year by Region



2.2.2 Table Processing: Merge Table 1 and Table 2

I merged Table 1 and Table 2. While keeping the variables ("Country of destination," "Regions," "Year," "Country Code," "Sex") which were included in both tables, I merged the unique column 'Total population of both sexes at mid-year (thousands)' onto Table 1. I merged the tables at a relatively early visualization stage to process different data in one table. At the same time, I used the *df.groupby.sum()* function to group three variables that I would use as the parameters for visualization: Regions, Sex, and Year.

2.2.3 Table 1 Plotly: Bar Chart of International Migrant Stock at Mid-year

To better visualize the international migrant stock by region, I followed Graphic Excellence's principle to fit more information into one graph. I created a new bar chart using *px.bar()* in Plotly. Figure 2 could display each region with a clearer x-axis and y-axis. Next, I plotted Figure 3, which was color-coded by sex, but the rest looked similar to Figure 2.

```
[304] # Use plotly to plot the international migrant stock
      # ...
      import plotly.express as px
      fig1 = px.bar(group1, x='Regions', y='International Migrant Stock at Mid-year',
                    title='International Migrant Stock at Mid-year by Region')
```

```
[241] fig2 = px.bar(group1, x='Regions', y='International Migrant Stock at Mid-year',
                    color = 'Sex',
                    title='International Migrant Stock at Mid-year by Region (color coded by sex)')
```

Figure 2 International Migrant Stock at Mid-year by Region

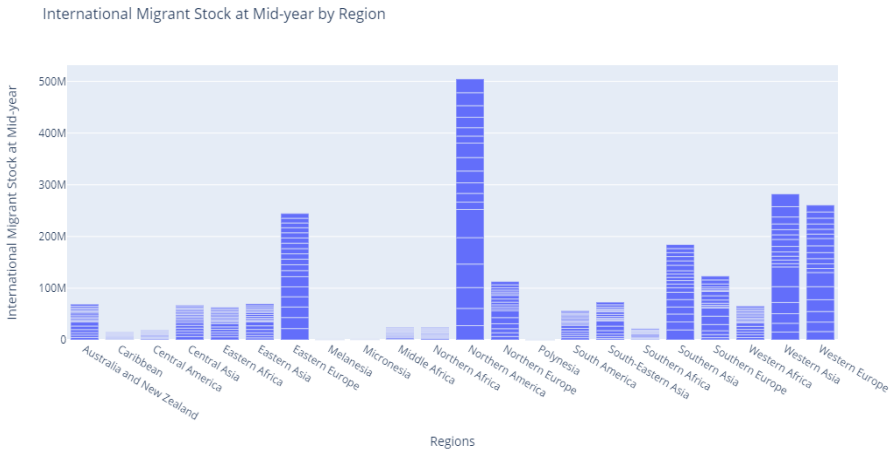
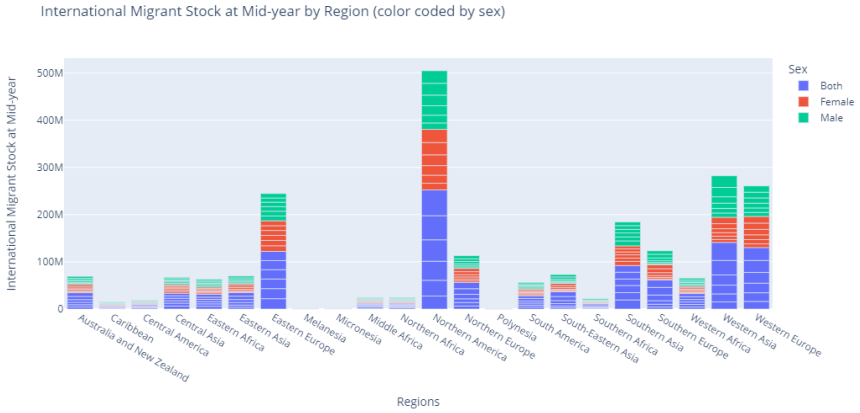


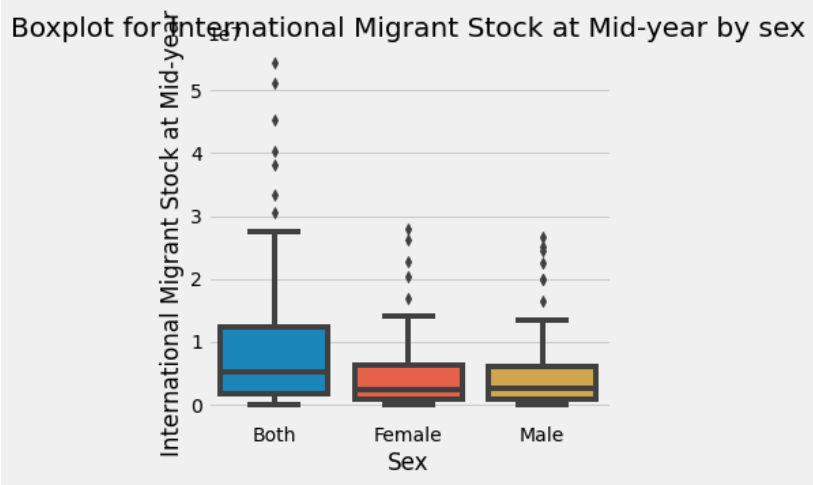
Figure 3 International Migrant Stock at Mid-year by region (color-coded by sex)



2.2.4 Table 1 Seaborn: Boxplot of International Migrant Stock at Mid-year

From a statistical point of view, a box plot is essential in analyzing data because it displays the basic descriptive statistics to depict data distribution. Hence, I generated a boxplot for the international migrant stock by sex. The sex variable included 'Female,' 'Male,' and 'Both.' I first plotted the boxplot by using *sns.catplot*. The final product (Figure 4) was informative but seemed clustered. Therefore, I chose to use *Plotly* to create another set of box plots.

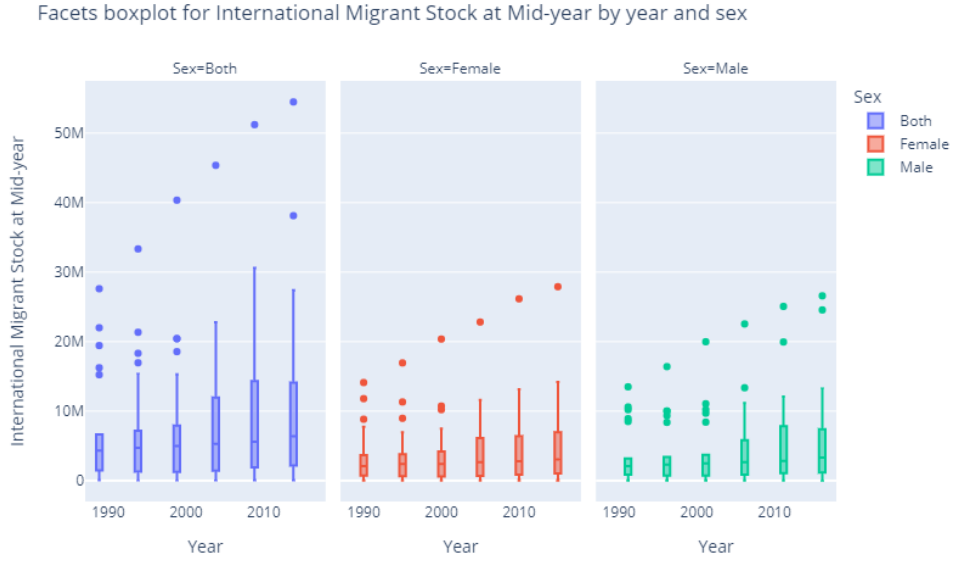
Figure 4 Boxplot for International Migrant Stock at Mid-year



2.2.5 Table 1 Small Multiples: Facets of Boxplots of International Migrant Stock at Mid-year

Instead of a single graph of boxplots, I used *px.box()* to create a collection of facets of boxplots for the International migrant stock by year and sex. **This step also follows Tufte's Principle of Comparison and Integration of information.** Also, **this is the first graph that contains 'small multiples' I created following Tufte's graphic excellence principle.** Figure 5 is the facet boxplot for international migrant stock at mid-year by year and sex, in which we can see a clear increasing pattern for both sexes. Here, I want to compare the international migrant stock by sex in the 25 years.

Figure 5 Facet boxplots for international migrant stock at mid-year by year and sex



2.2.6 Table 1 Small multiples: Facets bar plot and line charts of International Migrant Stock at Mid-year

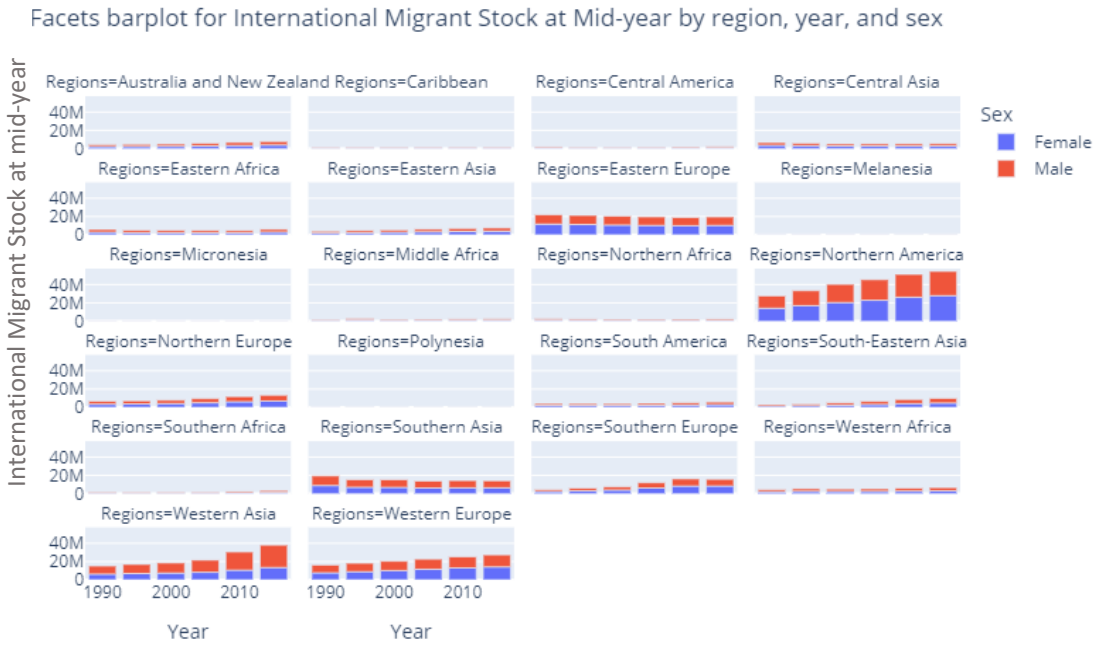
Continuing to follow Tufte's principle, the purpose of creating this group of small multiples was to only compare the male and female international migrant stock in different

regions by year and sex. This will allow me to get a closer look at the regional differences in the international migrant stock in the 25 years range. Figure 6 is a collection of facet bar plots for the international migrant stock by sex. With the visualization, we can see that there is an increasing pattern in international migrants in general (except in Western Europe), while some of the regions with more female migrants than male migrants.

```
[246] group1[group1["Sex"] != "Both"]
```

```
[344] barfig = px.bar(group1[group1["Sex"] != "Both"], x='Year',
                      y='International Migrant Stock at Mid-year', color='Sex',
                      facet_col='Regions', facet_col_wrap=4,
                      title='Facets barplot for International Migrant Stock at Mid-year by region, year, and sex')
```

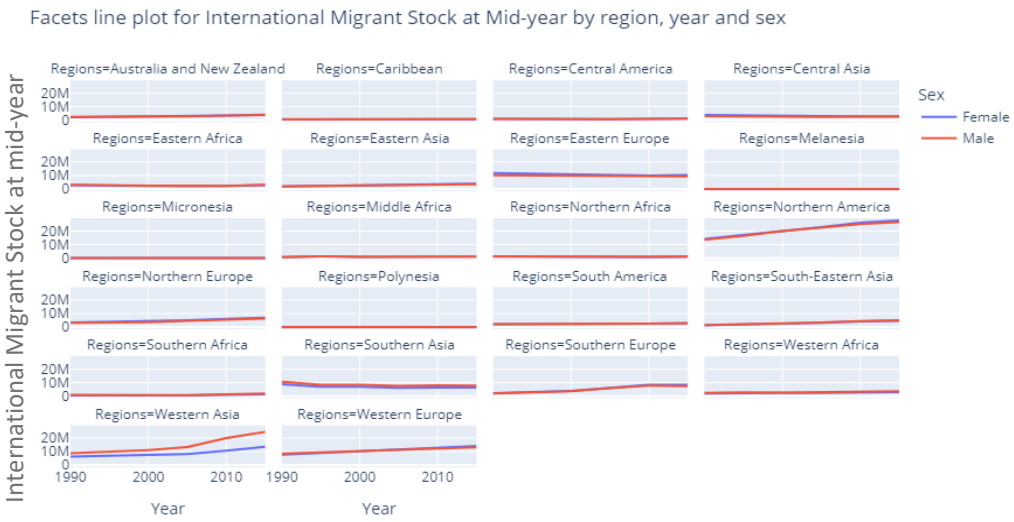
Figure 6 Facet bar plots for international migrant stock at mid-year by region, year, and sex



To better facilitate comparing international migrant stock by sex, I generate another facet collection of line charts to compare female and male migrants in different regions by year and sex. It is proper to use line charts here as I was plotting time series data. While the line charts did not show significant differences between female and male migrants in most regions, we could still read from Figure 7 that Western Asia had a significant increase in male migrants from 2000 to 2015 and South-Eastern Asia's migrants showed a positive pattern over the 25-year range.

```
[351] linefig = px.line(group1[group1["Sex"] != "Both"], x='Year', y='International Migrant Stock at Mid-year', color='Sex
                        facet_col='Regions',
                        facet_col_wrap=4,
                        title='Facets line plot for International Migrant Stock at Mid-year by region, year and sex')
```

Figure 7 Facet line charts for international migrant stock at mid-year by region, year, and sex

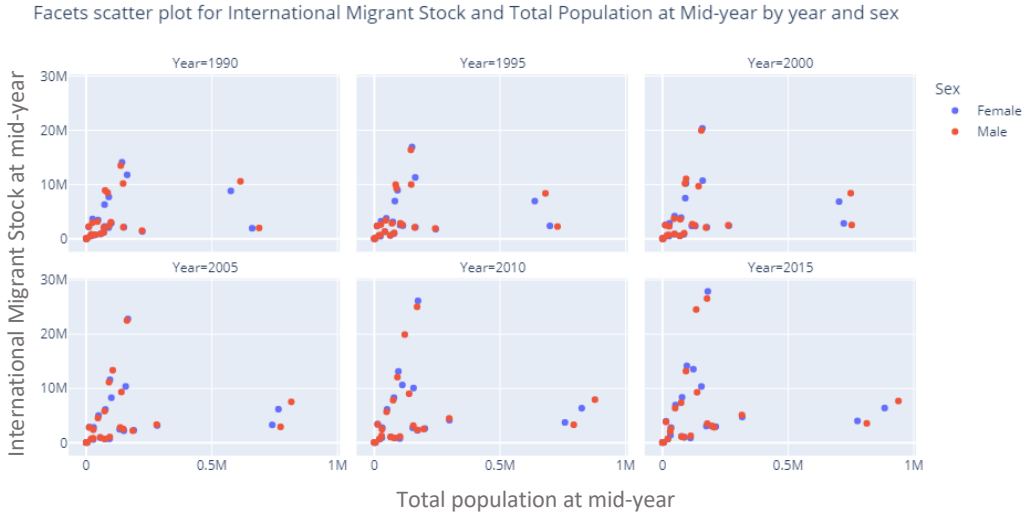


2.2.7 Table 1 & 2 Small Multiples: Facets scatter plot of International Migrant Stock/ Total Population

By following Tufte's Principle of Comparison and Multivariate, I created a facet of scatter plots that visualized the data points of international migrant stock and total population on the same x-y axis. A scatter plot could visualize the correlation of two variables in an obvious way, in which we could see that Figure 8 displays the dots spread out on each graph. From 1990 to 2015, both international migrants and the total population increased over the past 25 years.

```
[349] # Since we already merged table 1 and table 2, I want to plot the international migrant stock and total population in the same graph
      # when year increases, a tendency of total population increases and
      scatfig = px.scatter(group1[group1["Sex"] != "Both"], x='Total population of both sexes at mid-year (thousands)',
                           y='International Migrant Stock at Mid-year', color='Sex',
                     facet_col='Year', facet_col_wrap=3,
                     title = 'Facets scatter plot for International Migrant Stock and Total Population at Mid-year by year and sex')
```

Figure 8 Facets scatter plot for International Migrant Stock and Total Population at Mid-year by year and sex

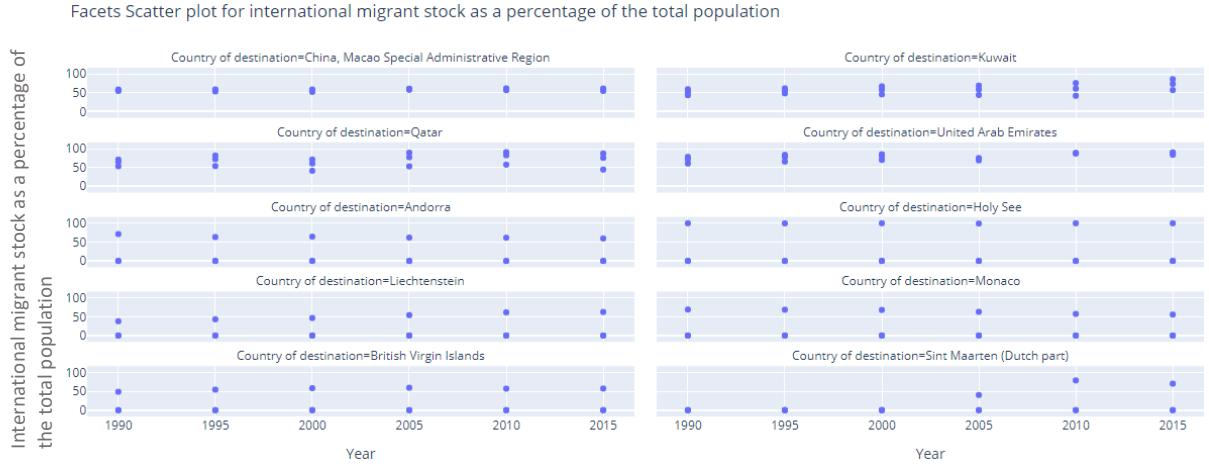


2.3 Table 3 Small multiples: Facet scatter plots for International Migrant Stock as a percentage of the Total Population

Table 3 contains the percentage of international migrant stock as to the total population. I selected the top 10 countries of 2015 and split the table containing the full-length data for these ten countries to be a new data frame. I then used *px.scatter()* to plot the facet of scatter plots to identify the change trend over the years. Figure 9 shows a slight increase in the pattern of international migrant stock as a percentage of the total population for these states.

```
fig_3333 = px.scatter(df3_2015_bothtop10_timeseries, x = 'Year', y = 'International Migrant Stock as a percentage of the total population',
                 facet_col = 'Country of destination',
                 title = "Facets Barplot for Female migrants as a percentage of the international migrant stock",
                 facet_col_wrap = 2,)
```

Figure 9 Facet Scatter plot for international migrant stock as a percentage of the total population

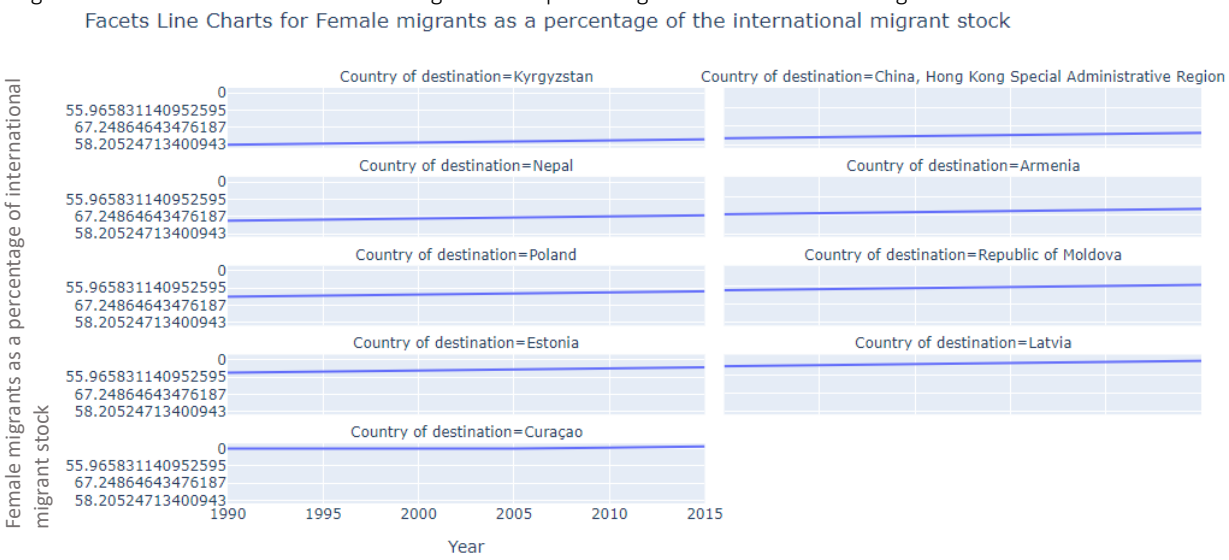


2.4 Table 4: Small multiples: Line charts for Female migrants as a percentage of the international migrant stock

Table 4 contains the percentage data of female migrants as a percentage of the international migrant stock. I used px.line() to generate the graph to compare and dig for more information from the table. Figure 10 displays the change in female migrant rates for the top 10 countries from 1990 to 2015. The figure shows a relatively stable straight line for each year because the original data clustered closely but with a slightly increasing pattern.

```
[413] fig_444 = px.line(df4_top10timeseries, x = 'Year', y = 'Female migrants as a percentage of the international migrant stock',
                     facet_col = 'Country of destination',
                     title = "Facets Barplot for Female migrants as a percentage of the international migrant stock",
                     facet_col_wrap = 2,)
```

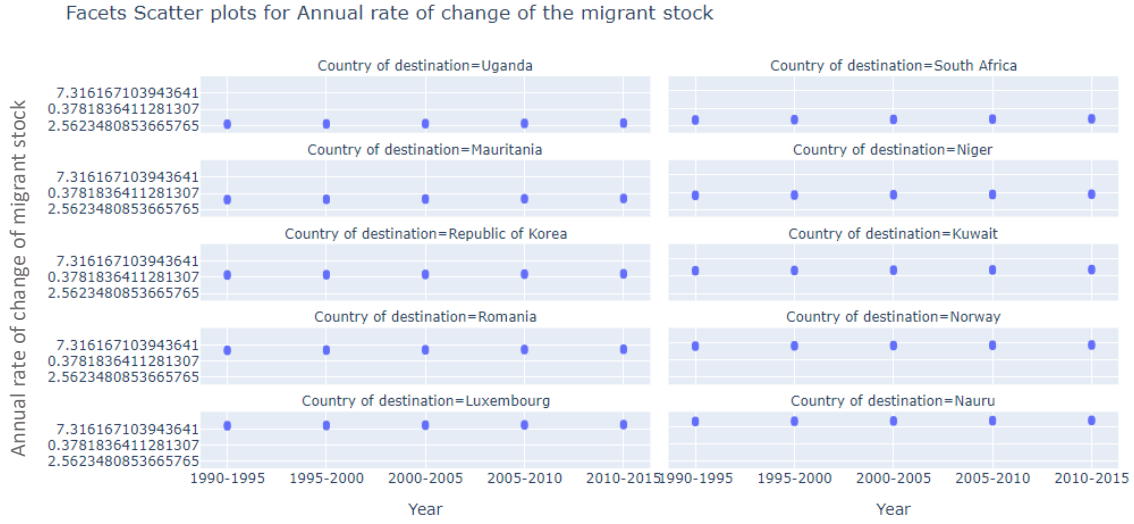
Figure 10 Facet Line Charts for female migrants as a percentage of the international migrant stock



2.5 Table 5: Small multiples: Facet Scatter Plots for Annual rate of change of the migrant stock

I used *px.scatter()* to generate the graph. Figure 11 displays the annual rate of change of the migrant stock. I used similar manual selection steps and picked the top 10 countries using 2015 as a reference level. The figure shows the scatter distribution representing the percentage change for each country.

Figure 11 Facets scatter plot for the annual rate of change of the migrant stock



2.6 Table 6A & 6B EDA: Estimated refugee stock at mid-year & Refugees as a percentage of the international migrant stock

2.6.1 Table 6A Preliminary Sorting & Exploration

For Table 6A and 6B, I intend to create a facet of bar charts to visualize the refugee stock and another facet of line charts to visualize the refugee as a percentage of the international migrant stock. The steps replicated what I did for Table 1 and Table 2. After I read Table 6A, I

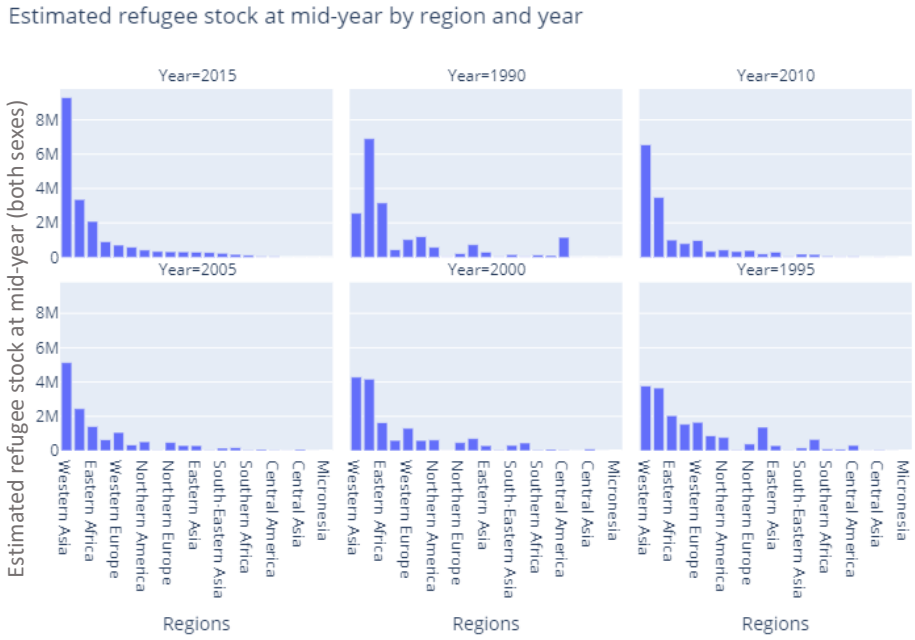
sorted the table by the estimated refugee stock in descending order and saved it into a new table. Then I grouped the data by region and year using the *tables.group()* function. Next, I summed the values for each variable and re-sorted the data in descending order for easy graphing.

2.6.2 Table 6A Small Multiples: Facet bar charts for estimated refugee stock at mid-year

Unlike Table 1, this time I created a collection of bar charts instead of a single graph for all regions. By using the *px.bar()* function, I created a facet bar chart for estimated refugee stock at mid-year for each region by year. F**ollowing Tufte's Comparison principle, a facet could display the difference in refugee stock for each country by year. Figure 12 displays each region's refugee stock pattern for the 25-year range.**

```
[355] # Create a bar chart by Regions, with the value of International Migrant Stock at Mid-year
      fig3 = px.bar(refugeestockbyregion.to_df(), x='Regions',
                    y='Estimated refugee stock at mid-year (both sexes) sum',
                      facet_col='Year', facet_col_wrap=3,
                    title='Estimated refugee stock at mid-year by region and year')
```

Figure 12 Estimated refugee stock at mid-year by region and year



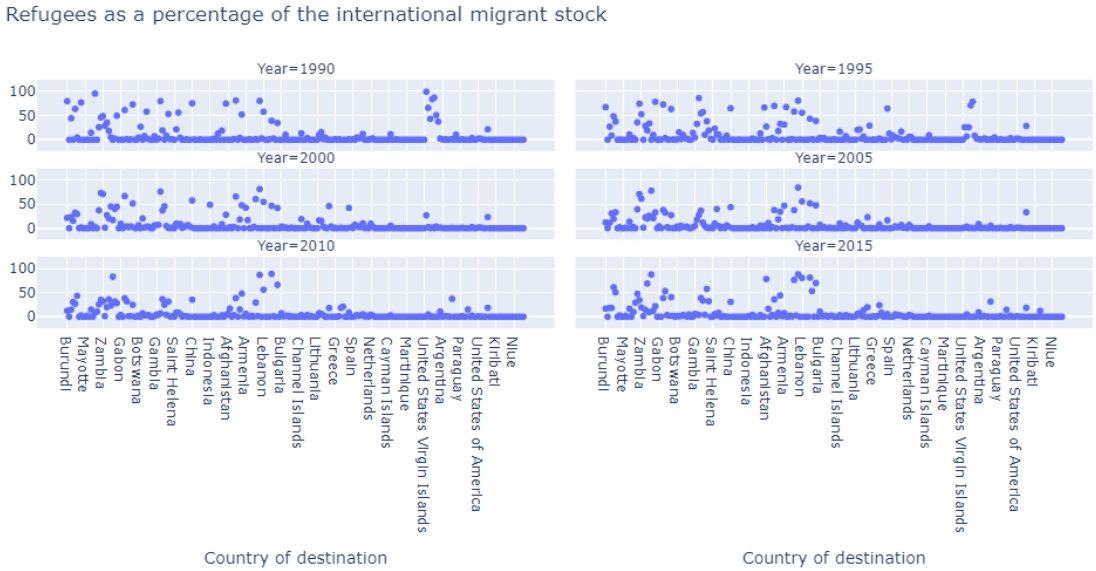
2.6.2 Table 6B Plotly: Facet scatter plots for refugee stock as a percentage of international migrant stock

By sorting the data, I realized that the State of Palestine has a percentage data over 100 percent for refugee stock as a percentage of mid-year. I considered these data points and excluded them from data visualization. This means there were only refugee overflows into Palestine instead of any other source of international immigrants. Therefore, I excluded the data for Palestine. Following the comparison principle, I created a new facet of scatter plots

(Figure 13) for the refugee percentage by using the function of *px.scatter()*.

```
# Create a line plot for table 6B to get the tendency of refugee as a % of international migrant stock
# Justify for why exclude the state of palestine in plotting the graph
# also justify the difficulty in choosing the x variable
fig6_3 = px.scatter(df_6B[df_6B["Country of destination"] != "State of Palestine"],
                    x='Country of destination',
                    y='Refugees as a percentage of the international migrant stock',
                     facet_col='Year',
                    title = "Refugees as a percentage of the international migrant stock", facet_col_wrap=2,)
```

Figure 13 Refugee as a percentage of the international migrant stock



3. Discussion & Interpretation

The UN dataset contains relevant information on the international migrant stock from 1990 to 2015. The stock of refugees was documented as part of the international migrant stock and was used as a calculation parameter. While I tried to plot the dataset with different graphic libraries, I primarily used Plotly Express, which genuinely provides functions that could produce beautiful and clear plots with extensive and easy-to-follow colors and legends. Since every table contains missing values, I replaced these missing values ('..') with zeros ('0'). Other than the tools and methods being used, several interesting facts can be derived from the visualization of the dataset. For ease of reading the discussion of the visualization, please refer to Appendix 1 for each figure.

By using Figure 5 (*Boxplot for International Migrant Stock at mid-year by year and sex*) as a motivation graph, we can see that there is an overall positive correlation between each year and international migrant stock. As time moves on, the total number of immigrant stock increases. Also, Figure 8 (*Facets scatter plot for International Migrant Stock and Total Population at Mid-year by year and sex*) shows that the dots spread out on both the x-axis and the y-axis, which means that both the total population and international migrants were increasing over the years. By observing Figure 6 (*Facet bar plots for international migrant stock at mid-year by region, year, and sex*) and Figure 7 *(Facet line charts for international migrant*

stock at mid-year by region, year, and sex), the distribution of female and male migrants remained relatively equal from 1990 to 2015, with the stock for female migrants being slightly higher than male in 2015. On the other hand, another motivation is Figure 1 (*International Migrant Stock at Mid-year by region*), which ranked international migrant stock by region. We can see that Northern America, Western Asia, and Western Europe are the top 3 regions with the most international migrant stocks over the years. These three regions are on top of the rank for very different reasons.

Northern America includes countries like the United States and Canada, whereas Western Europe includes countries like Germany, Netherlands, and Switzerland. These countries are commonly perceived as the best immigration destinations, with relatively open cultural environments, educational resources, and living situations.

Figure 12 (*Estimated refugee stock at mid-year by region and year*) shows the ranking change of estimated refugee stock from 1990 to 2015. Intriguingly, Western Asia is ranked at the top of international migrant and refugee stock charts. However, Northern America and Western Europe ranked relatively low on the bar charts for Estimated Refugee Stock by year. One primary reason is that these three regions' regional political situation was highly different. Western Asia includes countries like Iraq, Syria, and the State of Palestine. While Western Asia has the most international migrant stock for both men and women, the primary source of these migrants is refugee flows. From Figure 13 *(Refugee as a percentage of the international migrant stock*), we can see that most Western Asia countries have a constantly high refugee rate as a percentage of international migrant stock, as Jordan, Iraq, and the Syrian Arab Republic. The states where there was ongoing war or used to be a war zone tended to have more refugee flows. Due to regional instability and tension, a refugee may be the main and probably the only source of international migrants for most Western Asia states.

4. Conclusion

In conclusion, the project aims to visualize the "United Nations Migrant Stock" spreadsheet. In the visualizing process, I followed Tufte's 6 Principles of Data Visualization and Graphic Excellence principles to create meaningful charts to compare and demonstrate the underlying trends and correlations between various variables. Methodologically, I used boxplots, bar charts, line charts, and scatter plots to visualize the tendency and changes in international migrant stock from 1990 to 2015. There has been an increasing trend of international migration over 25 years. Male and female migration distribution has remained relatively even for each region. Northern America, Western Europe, and Western Asia are ranked top 3 of the region with the most international migrants. The first two regions are known for their popularity as migration destinations. For Western Asia, one reason for this phenomenon is that there used to be military activities or political crises in some Western Asia countries like Iraq and Syria. Political instability and unrest led to the overflow of refugees into other states in Western Asia and other regions. Unlike America and Europe, the primary source of international migrants for Western Asia is refugee flows.

As one of the most important parts of human-centered data science, analysts are trained to create comprehensive data packages in problem-solving, which require reproducible and understandable codes, reports, and graphics for not only professional data analysts but also potential audiences who do not have data science backgrounds. With Tufte's 6 Principles of Graphic Integrity and Tukey's guidelines in descriptive data analysis, I am better prepared for future complex and real-life data analytical projects.

Reference

Leinhardt, G., & Leinhardt, S. (1980). Exploratory Data Analysis: New Tools for the Analysis of Empirical Data. *Review of Research in Education*, *8*, 85–157. https://doi.org/10.2307/1167124

Tufte, E. (2001). *The visual display of quantitative information*. Graphics Press.

Appendix 1: Figures

Figure 14 International Migrant Stock at Mid-year by Region

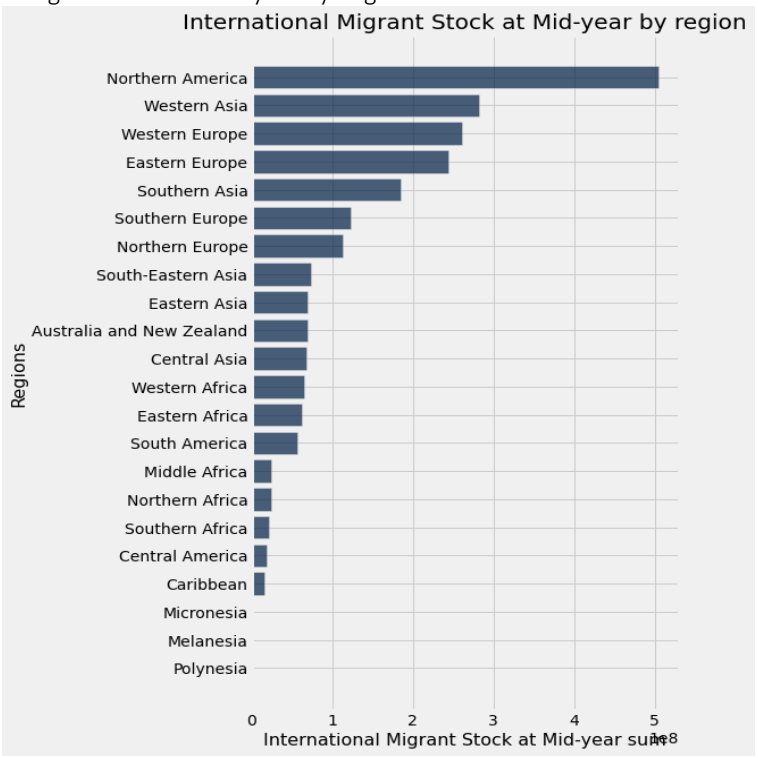


Figure 15 International Migrant Stock at Mid-year by Region

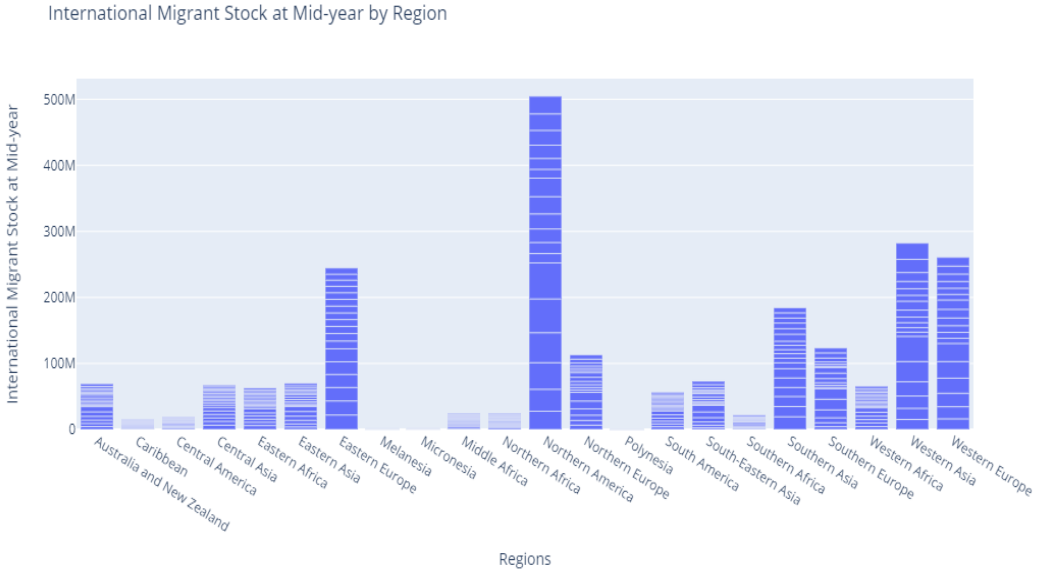


Figure 16 International Migrant Stock at Mid-year by region (color-coded by sex)

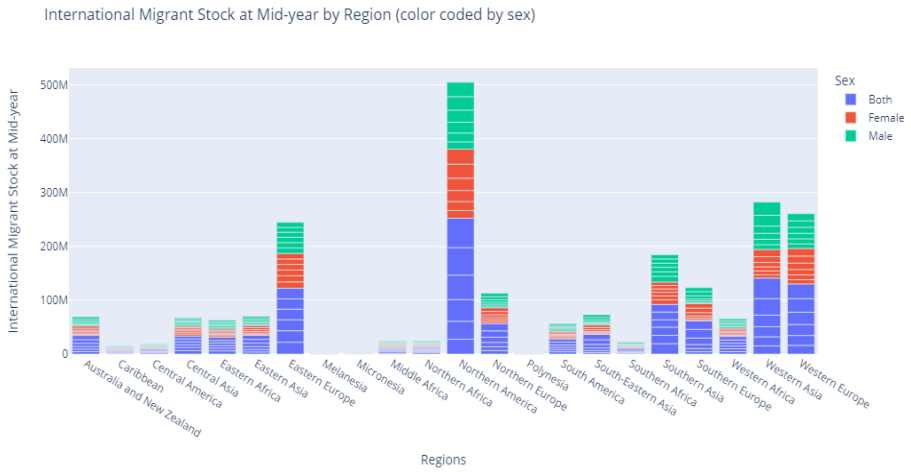


Figure 17 Boxplot for International Migrant Stock at Mid-year

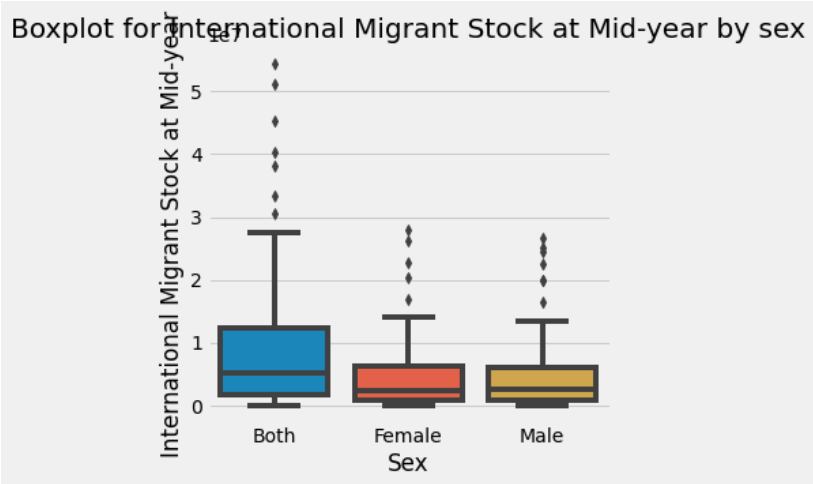


Figure 18 Facet boxplots for international migrant stock at mid-year by year and sex

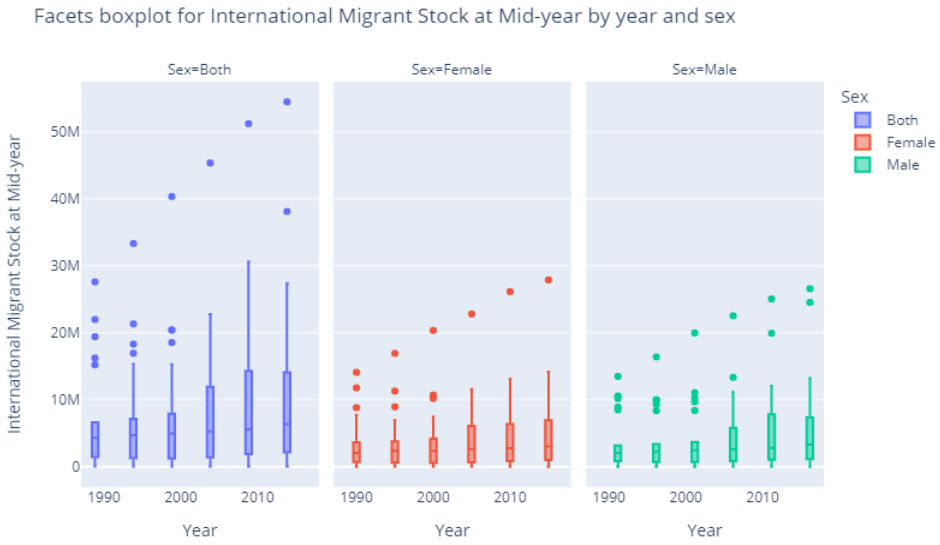


Figure 19 Facet bar plots for international migrant stock at mid-year by region, year, and sex

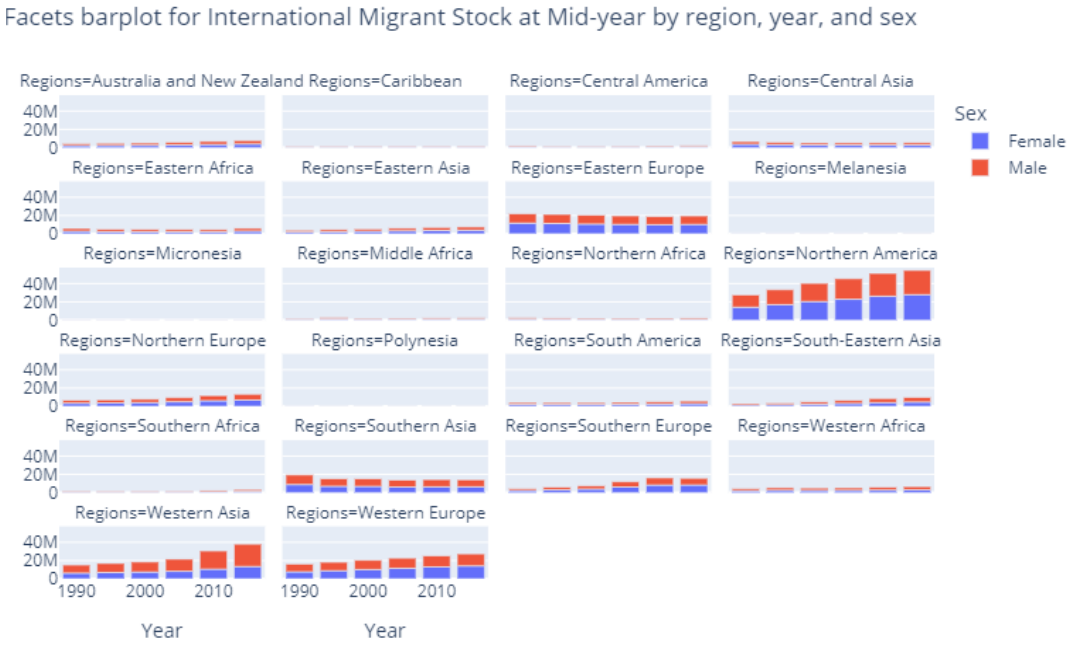


Figure 20 Facet line charts for international migrant stock at mid-year by region, year, and sex

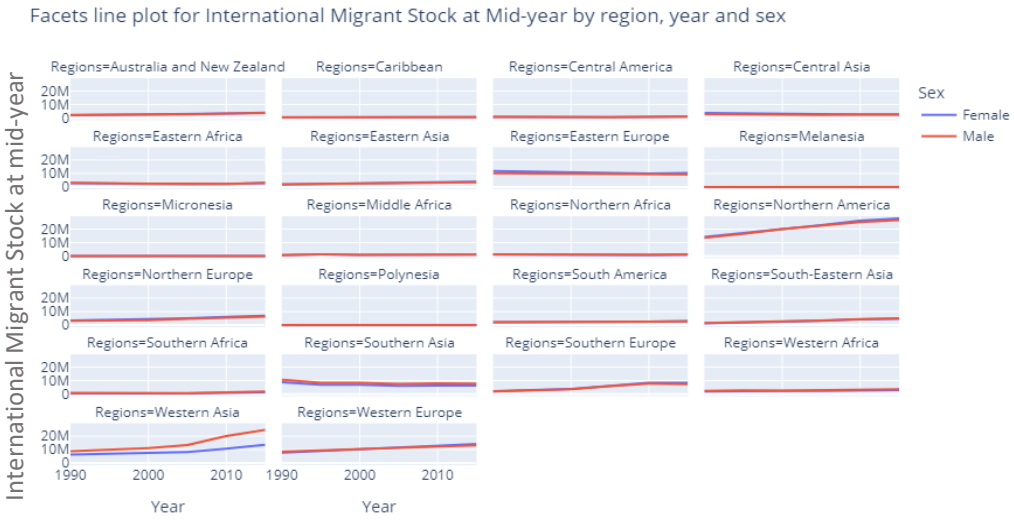


Figure 21 Facets scatter plot for International Migrant Stock and Total Population at Mid-year by year and sex

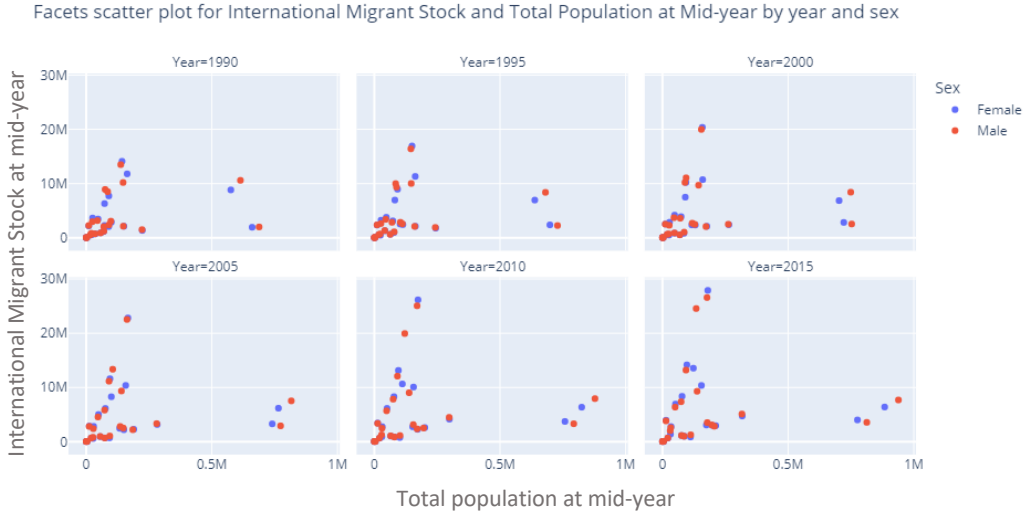


Figure 22 Facet Scatter plot for international migrant stock as a percentage of the total population

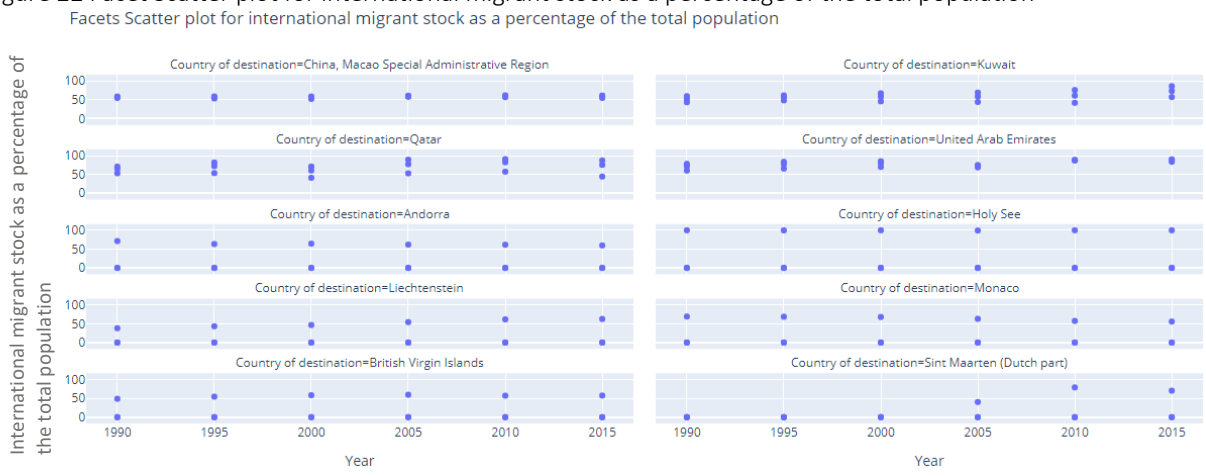


Figure 23 Facet Line Charts for female migrants as a percentage of the international migrant stock

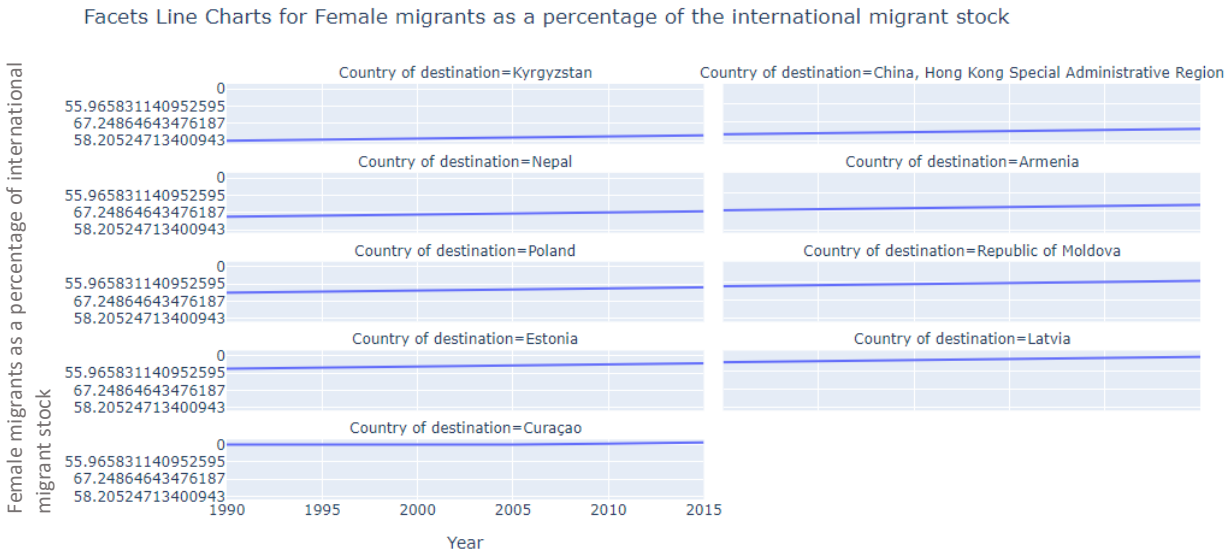


Figure 24 Facets scatter plot for the annual rate of change of the migrant stock

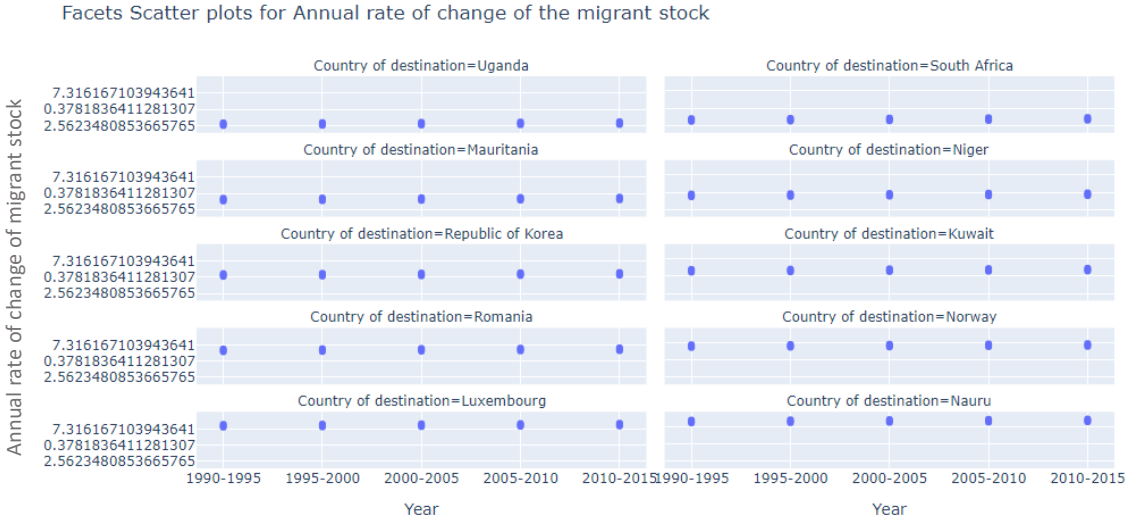


Figure 25 Estimated refugee stock at mid-year by region and year

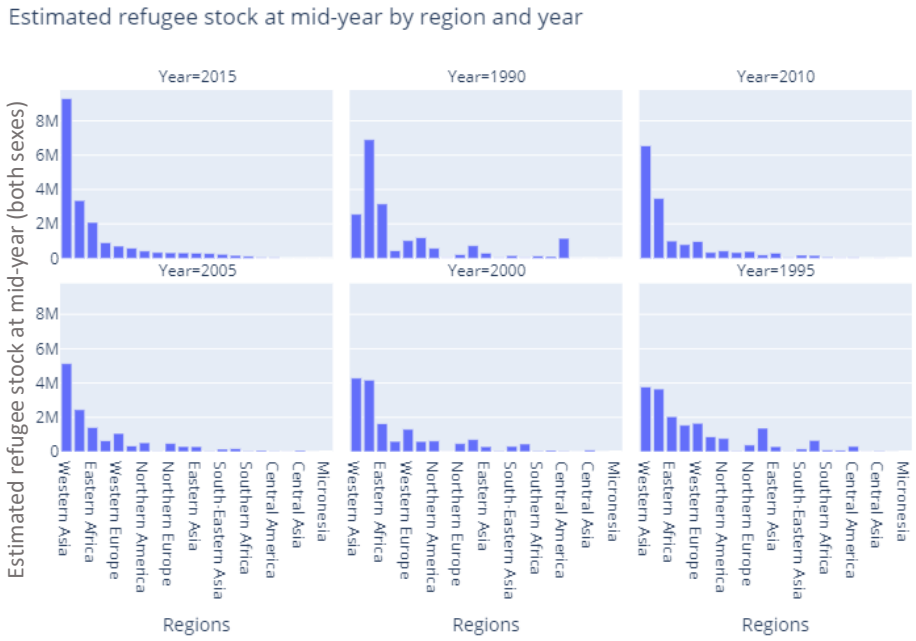
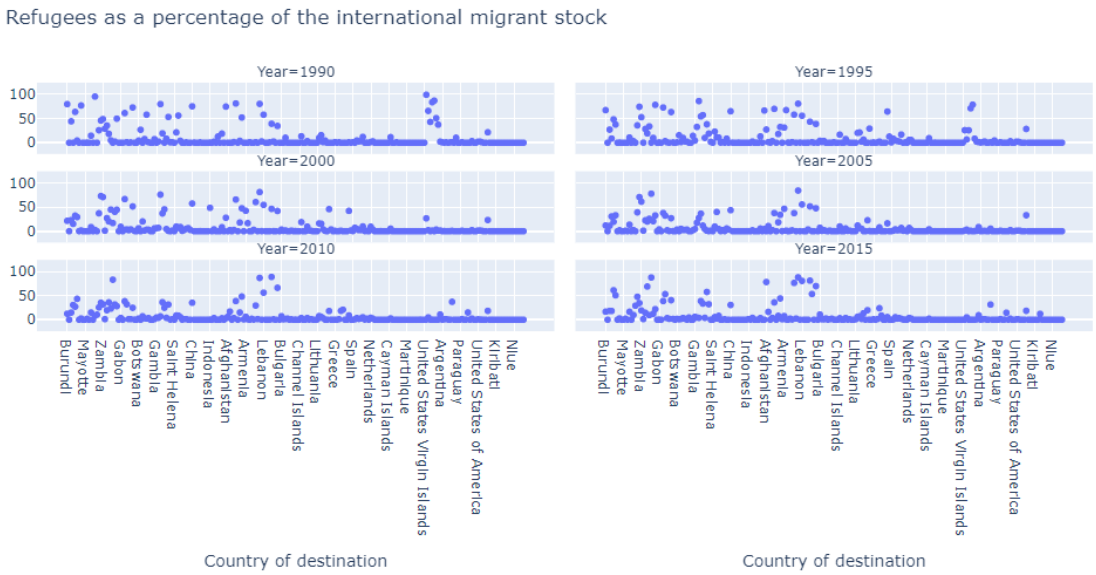


Figure 26 Refugee as a percentage of the international migrant stock



Appendix 2: Data Cleaning Continued

The data cleaning I did for the midterm was sufficient for data visualization. Therefore, I cleaned the six sheets again, trying to follow the Five Principles of Tidy Data. The final products include five new '.csv' tables for each of the first five and three separate tables that are split from Table 6. The following summarized the additional steps I took to clean the tables.

a. After reading in each table, directly drop the first 14 rows of contents
b. Renaming the first row as the headers for the columns as 'Year + Sex', such as '1990B', '1990F', and '1990M'. I used 'B' to represent both sex, 'F' to represent female, and 'M' to represent male (for the ease of melting columns).

```
# Check the results of renaming the headers, all the columns look correct now.
# Then we move on to splitting columns (column names) with multiple variables
df1_processed.head(5)
```

	Order	Major area, region, country or area of destination	Country Code	Type of Data	1990B	1995B	2000B	2005B	2010B	2015B	...	2000M
15	1	WORLD	900	NaN	152563212	160801752	172703309	191269100	221714243.0	243700236.0	...	87884839
16	2	Developed regions	901	NaN	82378628	92306854	103375363	117181109	132560325.0	140481955.0	...	50536796
17	3	Developing regions	902	NaN	70184584	68494898	69327946	74087991	89153918.0	103218281.0	...	37348043
18	4	Least developed countries	941	NaN	11075966	11711703	10077824	9809634	10018128.0	11951316.0	...	5361902
19	5	Less developed regions excluding least develop...	934	NaN	59105261	56778501	59244124	64272611	79130668.0	91262036.0	...	31986141

c. By using the *pandas.melt* function, I moved the 'Year + Sex' rows into one single column, originally named as 'variable'

```
[17] # Principle 2 of Tidy Data: Multiple variabes should not be stored in one column
     # Sex and year are stored in one column, so we need to split, and melt the table and seperate the information
     # I will keep "Order", "Major area", "country code", and "type of data" constant.
     # The value is "International migrant stock at mid-year", as mentioned in the EXCEL sheet
     df1_melted = (df1_processed.melt(id_vars=['Order','Major area, region, country or area of destination','Country Code', 'Type o
                                      value_name = "International Migrant Stock at Mid-year"))
```

```
[18] # Check if I melted the table correctly
     # I successfully splitted year+sex from "international migrant stock at mid-year"
     df1_melted.head(10)
```

	Order	Major area, region, country or area of destination	Country Code	Type of Data	variable	International Migrant Stock at Mid-year
0	1	WORLD	900	NaN	1990B	152563212
1	2	Developed regions	901	NaN	1990B	82378628
2	3	Developing regions	902	NaN	1990B	70184584
3	4	Least developed countries	941	NaN	1990B	11075966
4	5	Less developed regions excluding least develop...	934	NaN	1990B	59105261
5	6	Sub-Saharan Africa	947	NaN	1990B	14690319
6	7	Africa	903	NaN	1990B	15690623
7	8	Eastern Africa	910	NaN	1990B	5964031
8	9	Burundi	108	B R	1990B	333110
9	10	Comoros	174	B	1990B	14079

d. I wrote a *Lambda* function to split the ‘Sex’ and ‘Year’ column

```
[278] # Now I will use a lambda function to slice "Year" into "Year" and "Sex"
      df1_melted = df1_melted.assign(Sex = lambda x: x.Year.str[4].astype(str), Year = lambda x: x.Year.str[:4].astype(str))
```

```
[279] # Chech if the lambda function worked or not
      df1_melted.head(10)
```

	Order	Major area, region, country or area of destination	Country Code	Type of Data	Year	International Migrant Stock at Mid-year	Sex
0	1	WORLD	900	NaN	1990	152563212	B
1	2	Developed regions	901	NaN	1990	82378628	B
2	3	Developing regions	902	NaN	1990	70184584	B
3	4	Least developed countries	941	NaN	1990	11075966	B
4	5	Less developed regions excluding least develop...	934	NaN	1990	59105261	B
5	6	Sub-Saharan Africa	947	NaN	1990	14690319	B
6	7	Africa	903	NaN	1990	15690623	B
7	8	Eastern Africa	910	NaN	1990	5964031	B
8	9	Burundi	108	B R	1990	333110	B
9	10	Comoros	174	B	1990	14079	B

e. I kept the for loop from the midterm to separate the major regions and countries from the ‘Major area, region, country or area of destination’.

```
[281] # Split the Major area column into 'Regions' and 'Major area, region, country or area of destination'
      last_region = None
      regions = []
      kept_rows = []
      for row in df1_melted.iterrows():
        if row[1][2] >= 900:
          last_region = row[1]["Major area, region, country or area of destination"]
          continue
        regions.append(last_region)
        kept_rows.append(row[0])
      df1_melted = df1_melted.iloc[kept_rows]
      df1_melted.insert(0, "Regions", regions, allow_duplicates=True)
```

```
[282] # Check if we correctly split the two columns.
      df1_melted.head()
```

	Regions	Order	Major area, region, country or area of destination	Country Code	Type of Data	Year	International Migrant Stock at Mid-year	Sex
8	Eastern Africa	9	Burundi	108	B R	1990	333110	B
9	Eastern Africa	10	Comoros	174	B	1990	14079	B
10	Eastern Africa	11	Djibouti	262	B R	1990	122221	B
11	Eastern Africa	12	Eritrea	232	I	1990	11848	B
12	Eastern Africa	13	Ethiopia	231	B R	1990	1155390	B

f. Replacing the capital letters in the ‘Sex’ column with "Both","Female","Male"

```
[389] # Replace the capital letter "B", "F", "M" with "Both","Female","Male"
      df1_melted = (df1_melted.replace(to_replace =["B","M","F"],value =["Both","Male","Female"]))
```

```
[390] # Check if change the content properly
      df1_melted.head()
```

	Regions	Country of destination	Country Code	Year	International Migrant Stock at Mid-year	Sex
8	Eastern Africa	Burundi	108	1990	333110	Both
9	Eastern Africa	Comoros	174	1990	14079	Both
10	Eastern Africa	Djibouti	262	1990	122221	Both
11	Eastern Africa	Eritrea	232	1990	11848	Both
12	Eastern Africa	Ethiopia	231	1990	1155390	Both

g. For the purpose of data visualization, I replaced the missing values (‘..’) with zeros.

```
[420] # Replacing missing values
      df2_melted["Total population of both sexes at mid-year (thousands)"] = df2_melted["Total population of both sexes at mid-year (thousands)"].apply(lambda x : 0 if x == ".." else x)
```

```
[646] df2_melted.tail(10)
```

	Regions	Country of destination	Country Code	Year	Total population of both sexes at mid-year (thousands)	Sex
4759	Micronesia	Palau	585	2015	0.000	Female
4761	Polynesia	American Samoa	16	2015	0.000	Female
4762	Polynesia	Cook Islands	184	2015	0.000	Female
4763	Polynesia	French Polynesia	258	2015	138.468	Female
4764	Polynesia	Niue	570	2015	0.000	Female
4765	Polynesia	Samoa	882	2015	93.584	Female
4766	Polynesia	Tokelau	772	2015	0.000	Female
4767	Polynesia	Tonga	776	2015	52.931	Female
4768	Polynesia	Tuvalu	798	2015	0.000	Female
4769	Polynesia	Wallis and Futuna Islands	876	2015	0.000	Female