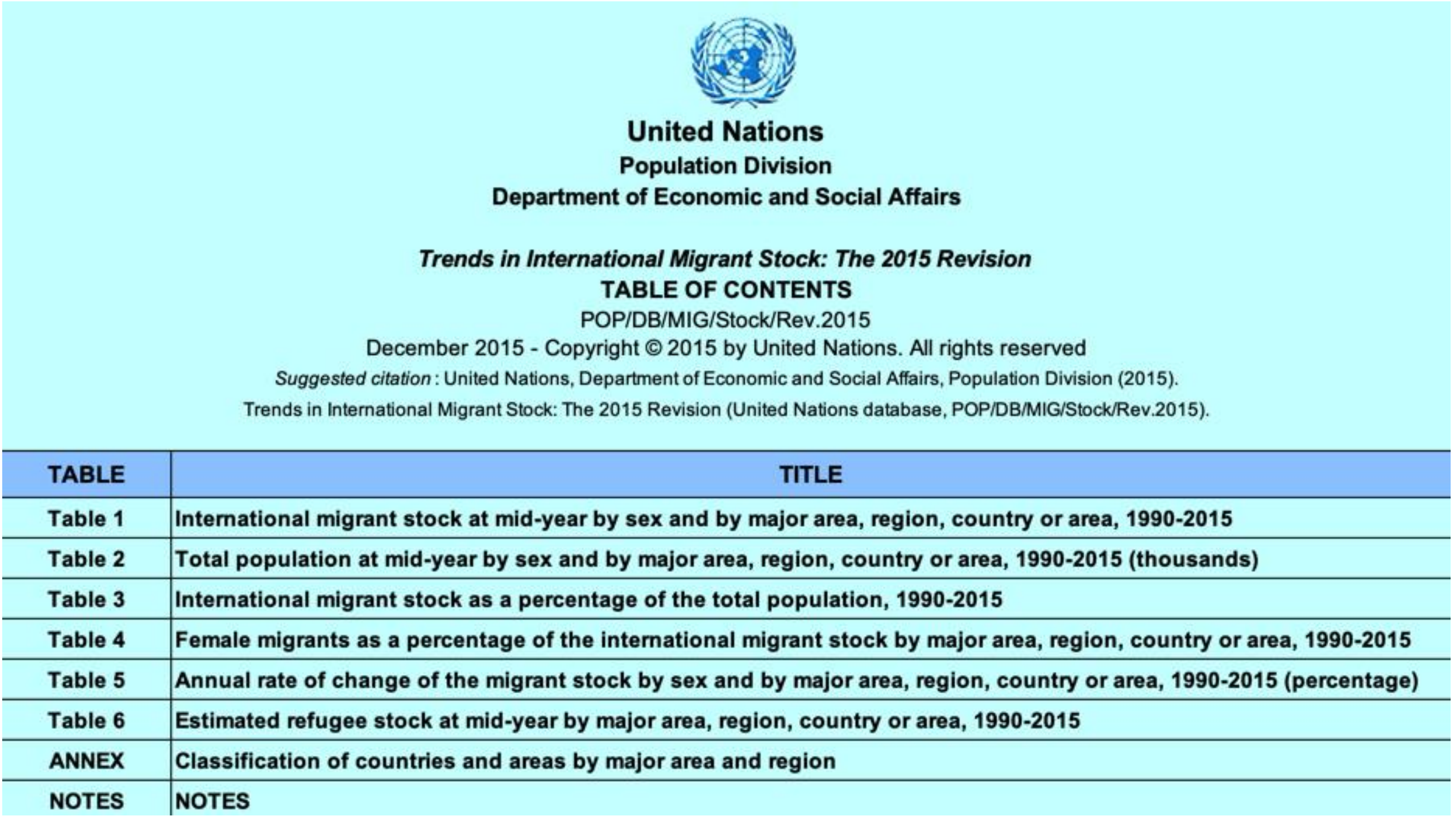
INF1340H: Programming for Data Science
Final Assignment

Exploratory Data Analysis with Python: Trends in International Migrant Stock Part II



United Nations
Population Division
Department of Economic and Social Affairs

Trends in International Migrant Stock: The 2015 Revision
TABLE OF CONTENTS
POP/DB/MIG/Stock/Rev.2015
December 2015 - Copyright © 2015 by United Nations. All rights reserved
Suggested citation : United Nations, Department of Economic and Social Affairs, Population Division (2015).
Trends in International Migrant Stock: The 2015 Revision (United Nations database, POP/DB/MIG/Stock/Rev.2015).

TABLE	TITLE
Table 1	International migrant stock at mid-year by sex and by major area, region, country or area, 1990-2015
Table 2	Total population at mid-year by sex and by major area, region, country or area, 1990-2015 (thousands)
Table 3	International migrant stock as a percentage of the total population, 1990-2015
Table 4	Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015
Table 5	Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)
Table 6	Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015
ANNEX	Classification of countries and areas by major area and region
NOTES	NOTES

Instruction by Professor Shion Guha, Erina Moon, & Priyanka Verma

December 15, 2022

Assignment by Jennifer Puskar (996914629)

Table of Contents

List of Figures

1. Introduction

Aspiring data scientists are often told that most jobs are “quick and dirty exploration” quests—meaning they involve a great deal of data cleaning and exploratory data analysis. The following study is an exploratory data analysis (EDA) that builds on an initial dive into data cleaning using the TIDY data principles. This phase of the analysis uses some of the visualization Python libraries, such as matplotlib and seaborn. It also begins to look at some of John Tukey on Exploratory Data Analysis principles and Edward Tufte’s data visualization principles.

As mentioned in the Part I report, the United Nations regularly reports on a wide range of trends—international migration being one of them. This data is often presented in Microsoft Excel workbooks that are formatted in a way to make it easy for the average person to read them. That doesn’t make these datasets computationally well formatted. This analysis continues to explore what it takes to transform the United Nations Trends in International Migrant Stock: The 2015 Revision TIDY from a computational lens, and how to translate the information into useful insights. This phase of the assessment begins to visualize the global trends, which will also highlight how well the data was cleaned in Part I.

To practice selecting, grouping, and sorting data simply, global trends were focused on for this study. Migrant population volumes from 1990 to 2015, and variables such as sex, and developed region were looked at (from Table 1 and Annex), as well as proportion of refugees (from Table 6).

2. Methods

Cleaning Limitations & Scope

Please note that only Tables 1 and 6 were used for this Part II EDA due to data cleaning limitations from Phase I. The focus of the study is uncovering global trends to develop marco-level questions about the data that will lead to further investigations and areas of study.

In Part I of this study, the UN Migrant Stock data set was cleaned according to the following five TIDY data principles:

1. Column names need to be informative, variable names and not values
2. Each column needs to consist of one and only one variable
3. Variables need to be in cells, not rows and columns
4. Each table column needs to have a singular data type
5. A single observational units must be in 1 table

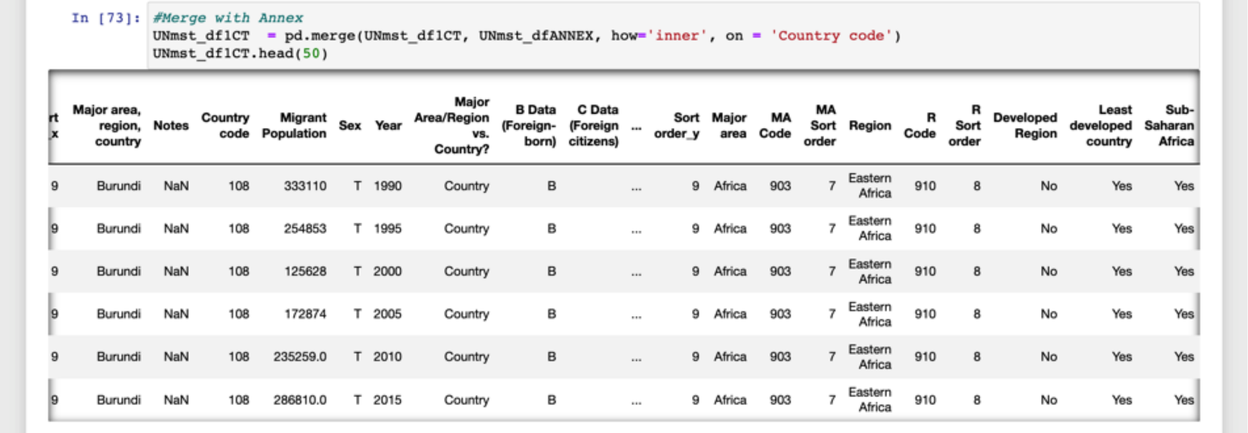
Despite cleaning efforts, principles four and five where breached. Countries, regions, and major areas (singular data types) were not separated into different columns, and repeated observational units, such as country name, sort order, and notes, repeated across the various tables instead of an index being set-up.

In addition to the TIDY data violations, the data types were not converted (e.g., string to float), null values were not addressed, and numerical data was not cleaned of any inconsistencies, such as extra spaces, periods, commas, decimal points, etc.

These limitations impacted the quality of EDA that could be conducted, since numerical canulations would often result in errors and impacted the execution of the study.

Additional cleaning for Part II

To mitigate some of the data cleaning issues remaining from Phase I, the Annex sheet was loaded and merged with the first data frame (Figure 1). This enabled the creation of a countries only column and helped provide useful regional context information that could be used for analysis later (e.g., developed vs. undeveloped).



```
In [73]: #Merge with Annex
         UNmst_df1CT  = pd.merge(UNmst_df1CT, UNmst_dfANNEX, how='inner', on = 'Country code')
         UNmst_df1CT.head(50)
```

rt x	Major area, region, country	Notes	Country code	Migrant Population	Sex	Year	Major Area/Region vs. Country?	B Data (Foreign-born)	C Data (Foreign citizens)	...	Sort order_y	Major area	MA Code	MA Sort order	Region	R Code	R Sort order	Developed Region	Least developed country	Sub-Saharan Africa
9	Burundi	NaN	108	333110	T	1990	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
9	Burundi	NaN	108	254853	T	1995	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
9	Burundi	NaN	108	125628	T	2000	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
9	Burundi	NaN	108	172874	T	2005	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
9	Burundi	NaN	108	235259.0	T	2010	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
9	Burundi	NaN	108	286810.0	T	2015	Country	B		...	9	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes

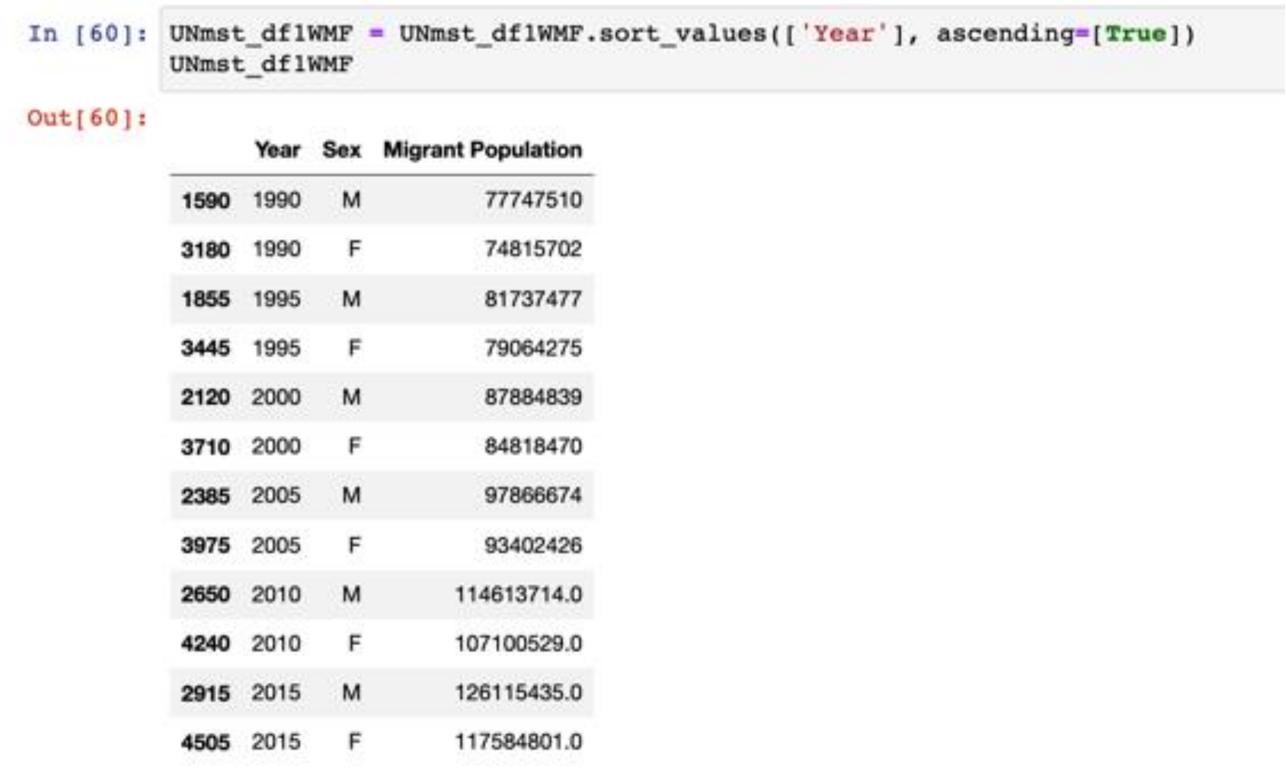
Figure 1: Annex Tab merged with first dataframe

Sorting, Grouping, Sub-setting, and Comparing

John Tukey's Principles were used to set up the data for visualization:

1. Sort
2. Group
3. Subset
4. Compare

To look at global trends, “World” and the total migration population for both genders was selected and sorted by year. A. line chart was then plotted to show the migrant population trend, even though technically the data is not continuous. The same data was then used to develop a grouped horizontal bar chart based developed vs. non-developed regions. To look at sex, both male and female data measurements were selected and sorted by year (Figure 2). The migrant population by sex was then compared using a horizontal bar chart since two items were being compared.



```
In [60]: UNmst_df1WMF = UNmst_df1WMF.sort_values(['Year'], ascending=[True])
         UNmst_df1WMF
```

Out[60]:

	Year	Sex	Migrant Population
1590	1990	M	77747510
3180	1990	F	74815702
1855	1995	M	81737477
3445	1995	F	79064275
2120	2000	M	87884839
3710	2000	F	84818470
2385	2005	M	97866674
3975	2005	F	93402426
2650	2010	M	114613714.0
4240	2010	F	107100529.0
2915	2015	M	126115435.0
4505	2015	F	117584801.0

Figure 2: Selected, subset, and sorted data for global migrant population by sex

The same data selection method was then applied to the proportion of refugees data from Table 6. A scatter plot was used to show the proportion of refugee make-up at specific points in time, and to also practice creating a different kind of chart.

Visualization

Edward Tufte's data visualization principles were considered when preparing the visualizations, including:

- Chart junk (minimalism)
- Small multiples

Charts were kept simple with clear titles and axis labels (Figure 3). No more than two colour were used. Efforts were made to compare no more than 3 variables at a time.

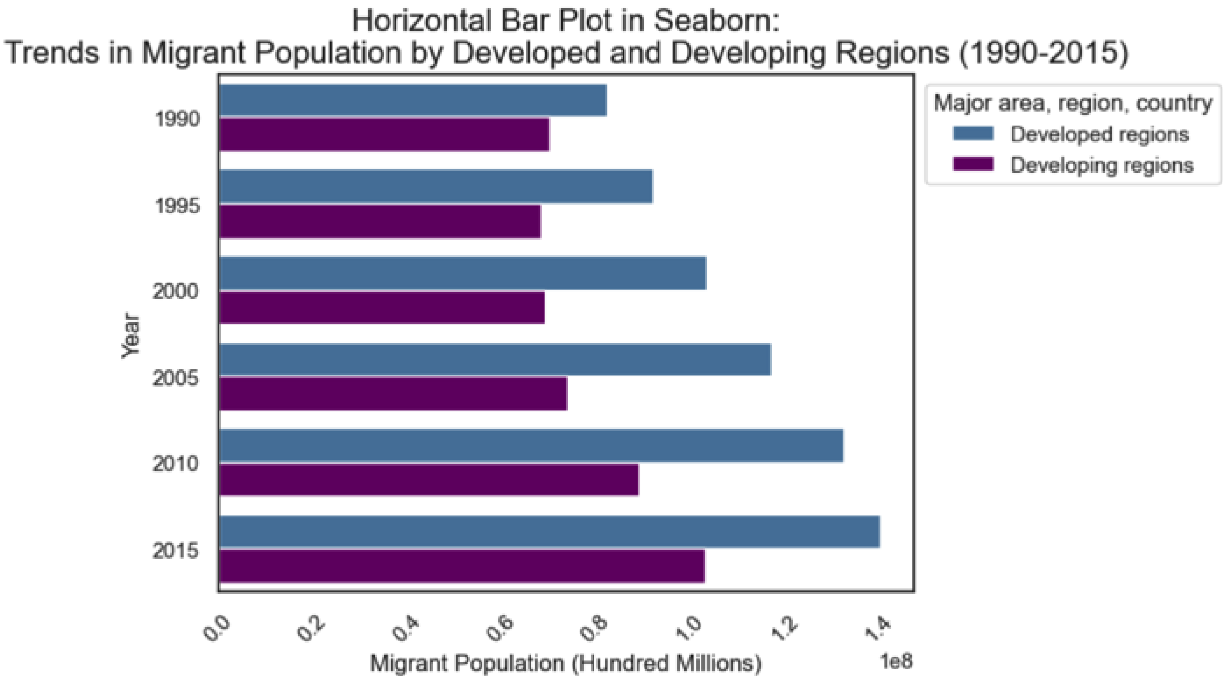
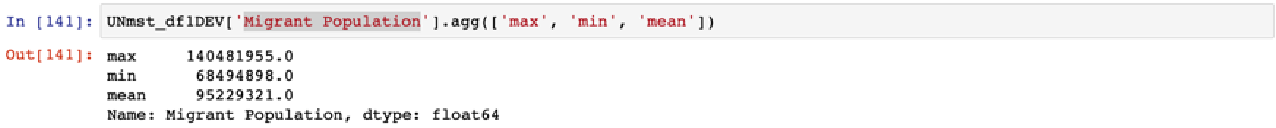


Figure 3: Grouped Horizontal Bar Chart with clear labelling, minimalistic features

The following five descriptive data metrics were also used to set up data analysis, and would be used with larger data, such as when looking at countries:

1. Min
2. Max
3. Mean
4. Median
5. Standard Deviation



```
In [141]: UNmst_df1DEV['Migrant Population'].agg(['max', 'min', 'mean'])

Out[141]: max      140481955.0
          min       68494898.0
          mean      95229321.0
          Name: Migrant Population, dtype: float64
```

Figure 4: Descriptive statistics for migrant population by region type

These metrics were useful to generate before visualizations to get a sense of the data narrative and also when setting up for box plots, etc.

3. Results & Discussion

Data insights and visualizations

Global migrant stock 1990-2015

Overall migrant stock increased from 1990 to 2015 (Figure 5). Two things to note when looking at this visualization:

- First, the y-axis is truncated, meaning is does not start at zero. This make the increase in migrant stock look drastically larger than it is.
- Second, the data has been plotted as a line chart, which is deceiving since the data was collected every 5 years and represents points in time, not continuous data. This chart may have been better as a bar chart or scatter plot, but was plotted as a line chart for demonstration and practice creating different types of charts.

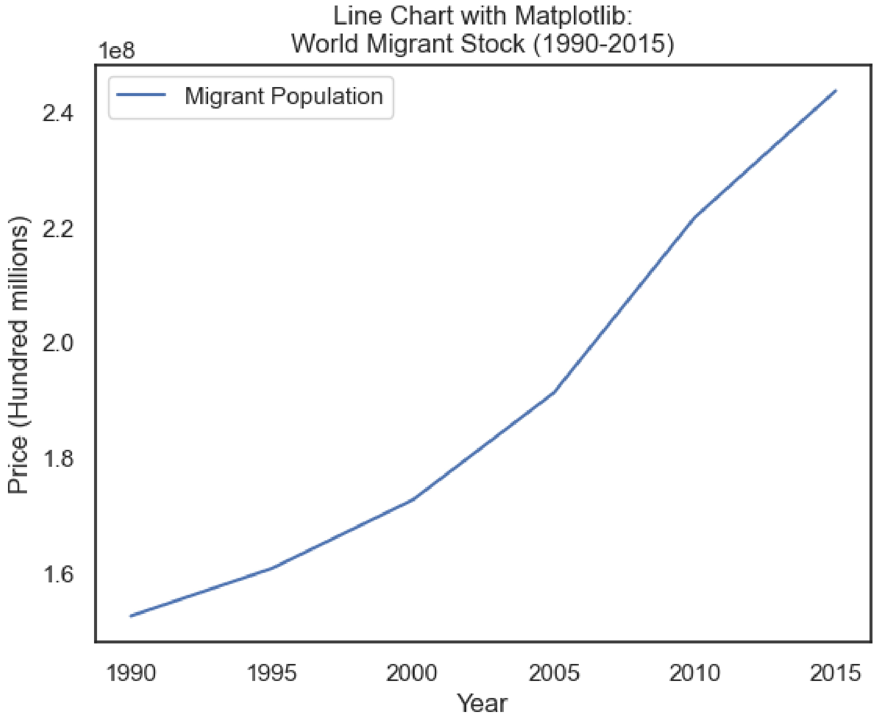


Figure 5: Global migrant stock for 1990-2015 - Chart of featuring "World" migration stock ever five years from 1990 to 2015. Migrant stock increase over the years, but please note the y-axis is truncated, so growth can be deceiving. This is also not continuous data and would likely be suited better as a scatter plot with a trend line.

Refugees as a percentage of international global migrant stock 1990-2015

Overall the proportion of refugees that make up the international global migrant stock decreased from 1990 to 2015 (Figure 6). Two things to note when looking at this visualization:

- Again, the y-axis is truncated, meaning is does not start at zero. This makes the decrease look drastically more than it is.
- Second, the scatter plot may benefit from a trendline and more data (e.g. countries proportion of refugees). This chart may also be clearer as a stacked bar chart of bar chart, comparing the proportion of refugees to the total, which is lost in this format. Again, the scatter plot was created for illustrative purposes.

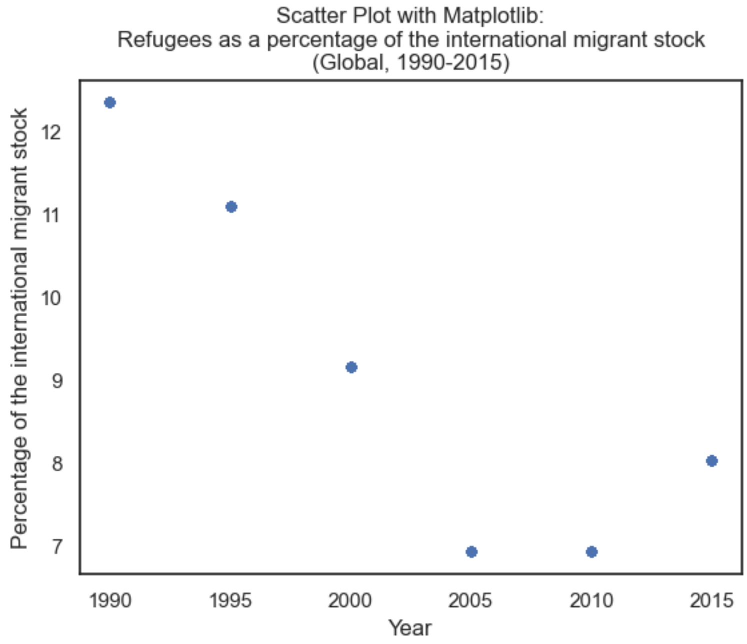


Figure 6: Refugees as a percentage of the global migrant stock for 1990-2015 - Chart of featuring "World" refugee as a percentage of international migration stock ever five years from 1990 to 2015. Proportion of refugees accountin for migrant stock decresed until 2010, but appears to be back on the rise. Please note, again the y-axis is truncated, so decreases can be deceiving. This scatter plot may also be clearer with a trendline or a stack bar chart.

Global trend in migrant stock by sex, 1990-2015

Males consistently accounted for more of the global migrant stock than females. Further investigation here may be interesting to see why and to see how various programs are set up to serve migrants based on sex or gender identity.

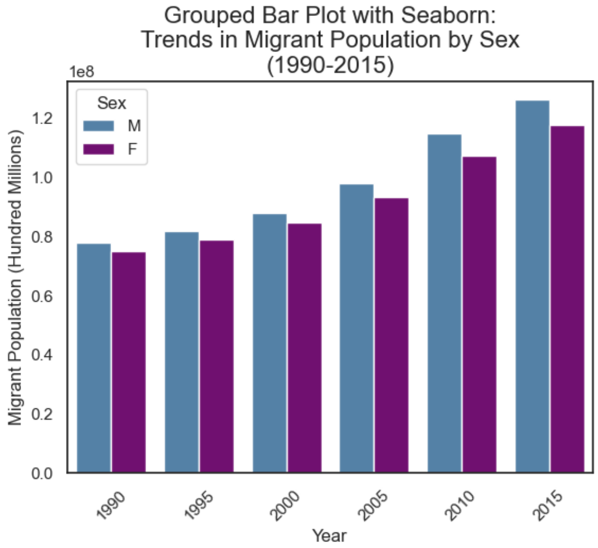


Figure 7: Global trend in migrant stock by sex, 1990-2015 – Chart comparing the number of males vs. females in the global migrant stock.

Global trend in migrant stock by region, 1990-2015

The global migrant stock is consistently higher in Developed regions compared to developing regions (Figure 8). A box plot was also developed to show the max, minimums, and median (Figure 9). There are two things to note when looking at this visualization:

- The x-axis uses 1e8. This is likely not understood by most people and should be converted for increased comprehension.
- The legend was places outside of the chart area to increase white small in the chart area, particularly since two colours and variables where already being compared. This was an extra effort from the default chart. It will be important to increase understanding of these customizations as a data scientist conducting Eda.

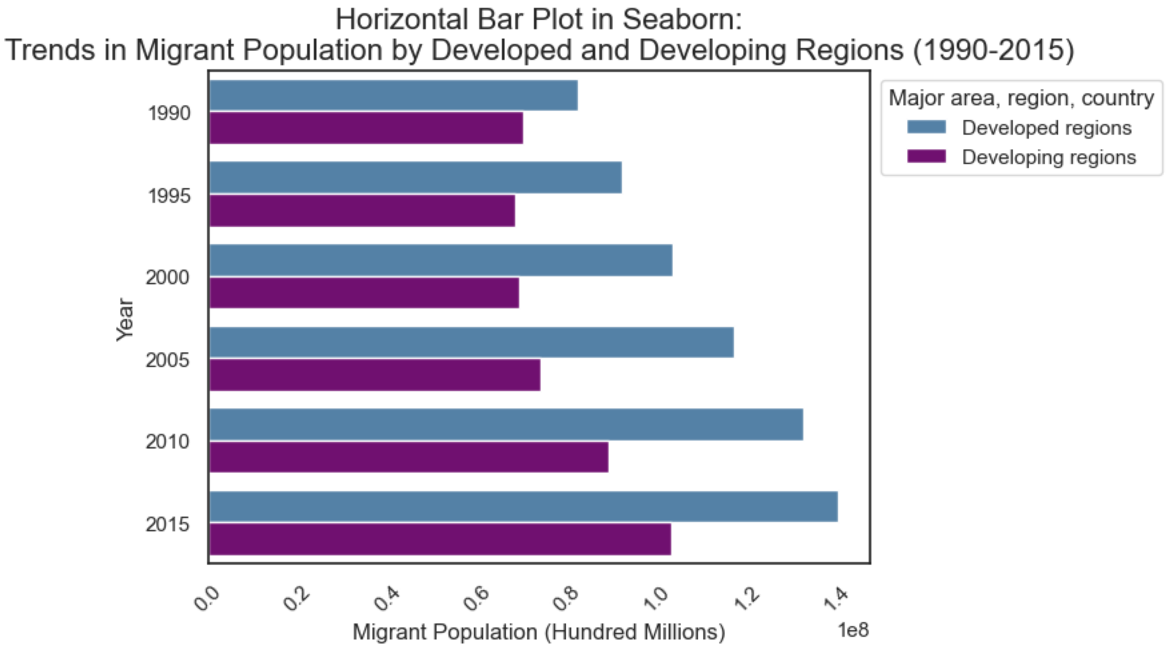


Figure 8: Trends in migrant population by region for 1990-2015 – The chart shows that developed regions consistently have a higher migrant stock.

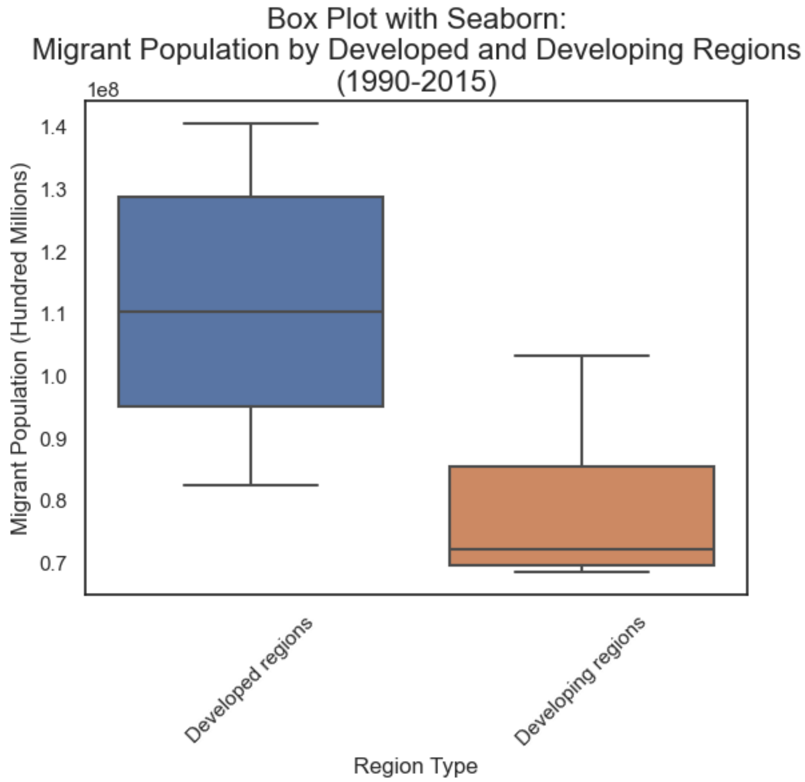


Figure 9: Boxplot comparing the global migrant stock by region – Box Plot comparing the migrant population by region. The chart shows a higher median for developed regions compared to developing regions. The maximum for Developing regions is also further aware from the median than the minimum, meaning there are more instances closer to the minimum.

4. Conclusion & Next Steps

One of the largest lessons reinforced through this exercise was that clean data is the foundation to smooth exploratory data analysis. Admittedly the study did not accomplish almost have of the analysis it wanted to cover, as highlighted in the Methods section, due to limitations with data cleaning. Further data cleaning, such as the removal of spaces and characters from numerical data as well as the proper conversion of data types, would need to be conducted to dive deeper into EDA.

The following outlines questions and areas of further exploration with respect to data cleaning and EDA:

- When do data types need to be converted?
- What are some best practices for working through errors when trying to convert data types?
- What should be done with Null/missing values?
- What are the best practices for visualization different types of data? Dive deeper into Tukey and Tufte.
- How can the functions and libraries be used to make EDA more efficient?

Data visualization tools/libraries, such as seaborn and matplotlib, both proved to be useful tools to present data. It will be important to expand beyond the defaults of these scripts to make cleaner and more comprehensible charts. Further exploration of the capabilities and functions would enhance the

EDA skill set of any becoming data scientist. Online sources, like Stack overflow, often contained helpful tricks or more efficient functions to complete the data cleaning and visualization.

The next phase of the project would be returning it data cleaning. As expected, the EDA exercise exposed gaps in the data cleaning and proved to be an important lesson in teaching foundational skills.

References

How to Create a Grouped Bar Plot in Seaborn (Step-by-Step)
https://www.statology.org/seaborn-grouped-bar-plot/

Pandas Drop Rows Based on Column Value
https://sparkbyexamples.com/pandas/pandas-delete-rows-based-on-column-value/#:~:text=Use%20drop()%20method%20to,or%20on%20another%20column%20value

Selecting rows based on multiple column values in pandas dataframe
https://stackoverflow.com/questions/29219011/selecting-rows-based-on-multiple-column-values-in-pandas-dataframe