Audrey Medaino-Tardif

INF 1340: Introduction to Python Programming

December 15th, 2022

Final Project:

Data Visualization of the UN Migrant Stock Dataset

1.Abstract

The files have been submitted via GitHub (username: @odgerey). The contents of my final projects include a working Jupyter file "final-project.ipynb" & a pdf, No readme files have been added. This is the write up accompanying the final project. It should be read while looking at the accompanying ipynb file.

2. Introduction

The purpose of the final project was to do some exploratory data analysis combined with data visualizations using the basic principles of Edward Tufte's *Envisioning Information Graphics*. Prior to beginning visualizations, I encountered a first issue, which was that despite my best efforts, my tidying was not tidy enough for reusing the code easily and for creating visuals. I therefore approached the project in a very different way than anticipated with visualizations coming last in my process:

1. Tidying data
2. (Re)grouping data
3. Analyzing what questions the data could answer
4. Creating visuals

3. Process & Troubleshooting EDA & Visualization

The first problem I encountered were the titles of my columns that were difficult to work with. For example, titles with spaces, capital letters, special characters are not practical for coding. More than this, in the saving process, my document added an extra column that needed to be

removed. I also realized that multiple columns I had kept avoiding data suppression, were not all that useful to the dataset and served more as clutter than anything else.

While I do not think removing those columns is the right way to go, a project needs two distinct files, one of which is a clean data file with as little information suppression as possible and a second of which is a visualization file. This second file is further cleaned to offer a bigger picture scope to the data. Human-centered data science principles are to avoid over essentialization of data, however, I realized that in practice too much unnecessary data only contaminates very useful information in a data set. The columns Country Code, Type, and Notes were not necessary; thus, I did not use them for the visualization file.

A second issue I encountered was that the document numbers which I had assumed were floats in the columns “migrants” and “refugee stock” etc., were actually strings. Simple boxplots did not work. Thus, I converted the strings to floats and in the process removed a row of special characters that had not been dealt with in the tidying phase. Every little bit left matters. It taught me that the tidying phase is essential.

I really struggled with the organization and process of EDA & Visualisations. I spent more time cleaning my data and really understanding what I had constructed as tables in the midterm. I focused on building off and making do with the state of the data from the midterm file, rather than go about restarting the cleaning process. It was a better learning experience to deal with the data as is and it furthered my understanding of the tidying process.

For the visualisations, I began by working with a smaller sample set using a Summary table I had created that had less rows. This enabled me to practice understanding visualizations before grouping my data from the dataset I wanted to use and represent. One of the first things I noticed with my early plots is that when dealing with a large number, the plots turn out strange with exponential numbers, making it difficult to get a good visual. I spent quite a bit of time to see if I could standardize the numbers found in my table, as this would have put every number on a scale of 0 to 1. However, this did not work out and overly flattened the data.

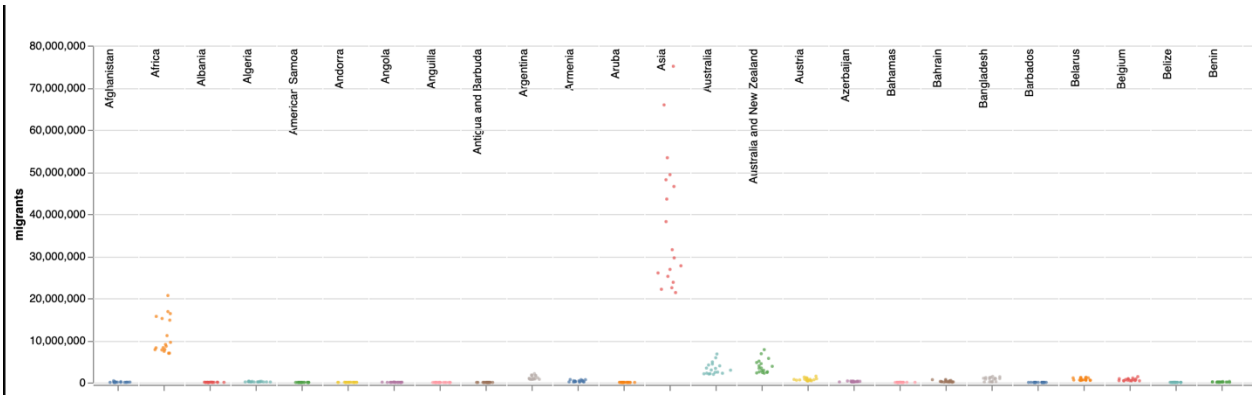
I also struggled with proper x & y parameter visualisations. Long titles like “Least developed regions not including less developed regions” simply do not fit on the x-axis, or at least I did not

find a solution to make it fit within the x-axis. Thus, I circumvented this issue by placing most of my longer titled parameters on the y-axis for better legibility.

Another issue pertained to the sheer quantity of countries and the difficulties of exploring this data without grouping it. For the most part, the countries did not fit on a single plot. I spent quite a bit of time finding a library that could plot all the countries and did with Altair, as I did not feel comfortable grouping the data prior to visualizing the bigger picture. HCDS asks us to be weary of using data that easily and readily available; what if my questions were missing vital information? While this scatterplot did not allow for details from smaller countries, it allowed me to see there were continents, continental regions, and countries mixed, which even during the tidying phase I found difficult to manage. The figures below demonstrate part of the full graphic, but as seen, the correct number of migrants are measured rather than exponential numbers, the full titles are present unperturbed, and the scatterplot has colored points to help differentiate the countries. This plot includes migrants all years and genders encompassed, which is why there are multiple points and respects two essential visualization principles:

Tufte Principle #1: Comparisons must be enforced within the scope of the eyespan. (p.76)

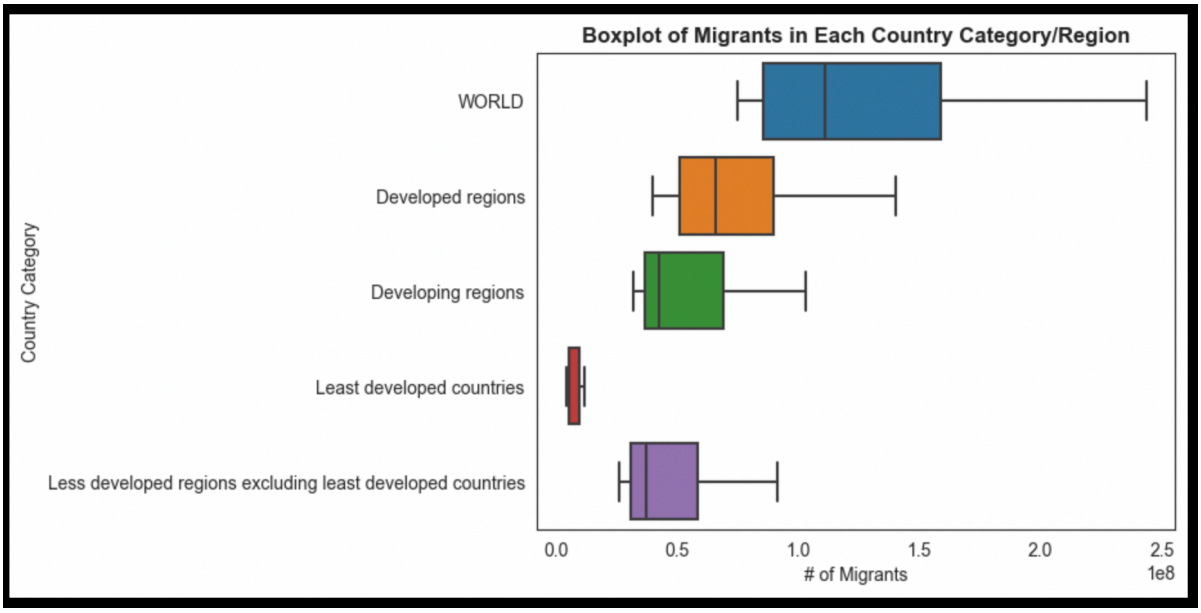
Tufte Principle #2: Use color in information design to label, to measure, to represent, or imitate reality, and to enliven or decorate (p.80).



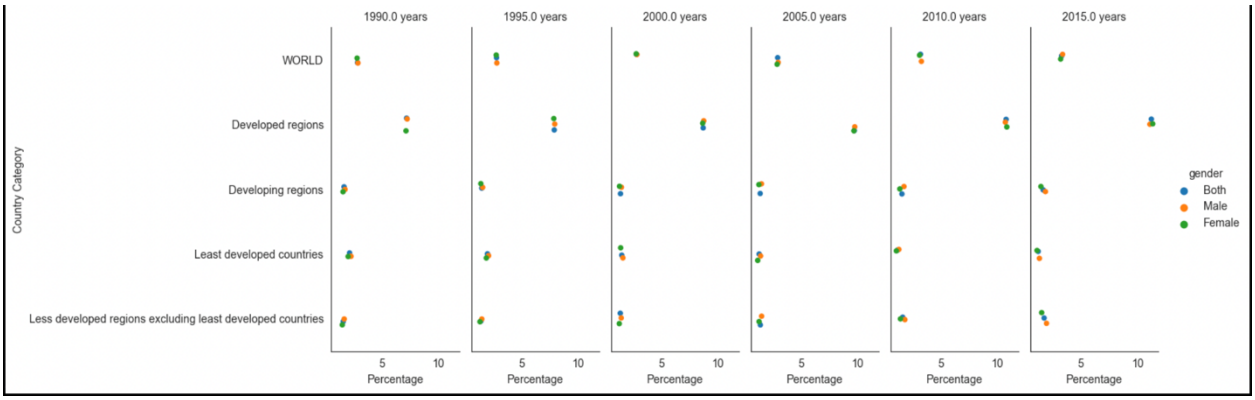
As a point of improvement, this graph would need to be bigger to see the elements with more clarity. Putting the countries on the Y-Axis would make the lettering more legible, rather than sideways as such. The ratio between for example Asia and Aruba is visibly unequitable and not a good comparative. However, the goal of this visual was not to compare, but rather offer a portrait

of ensemble to see how I could better group the data, which served its purpose. In the section of the graph shared below, Asia has the most migrants, followed by Africa. Africa has a larger migrant stock than Australia.

If I compare the above plot with a boxplot that covers approximately the same information, but rather than specific countries, they are summed into categories:



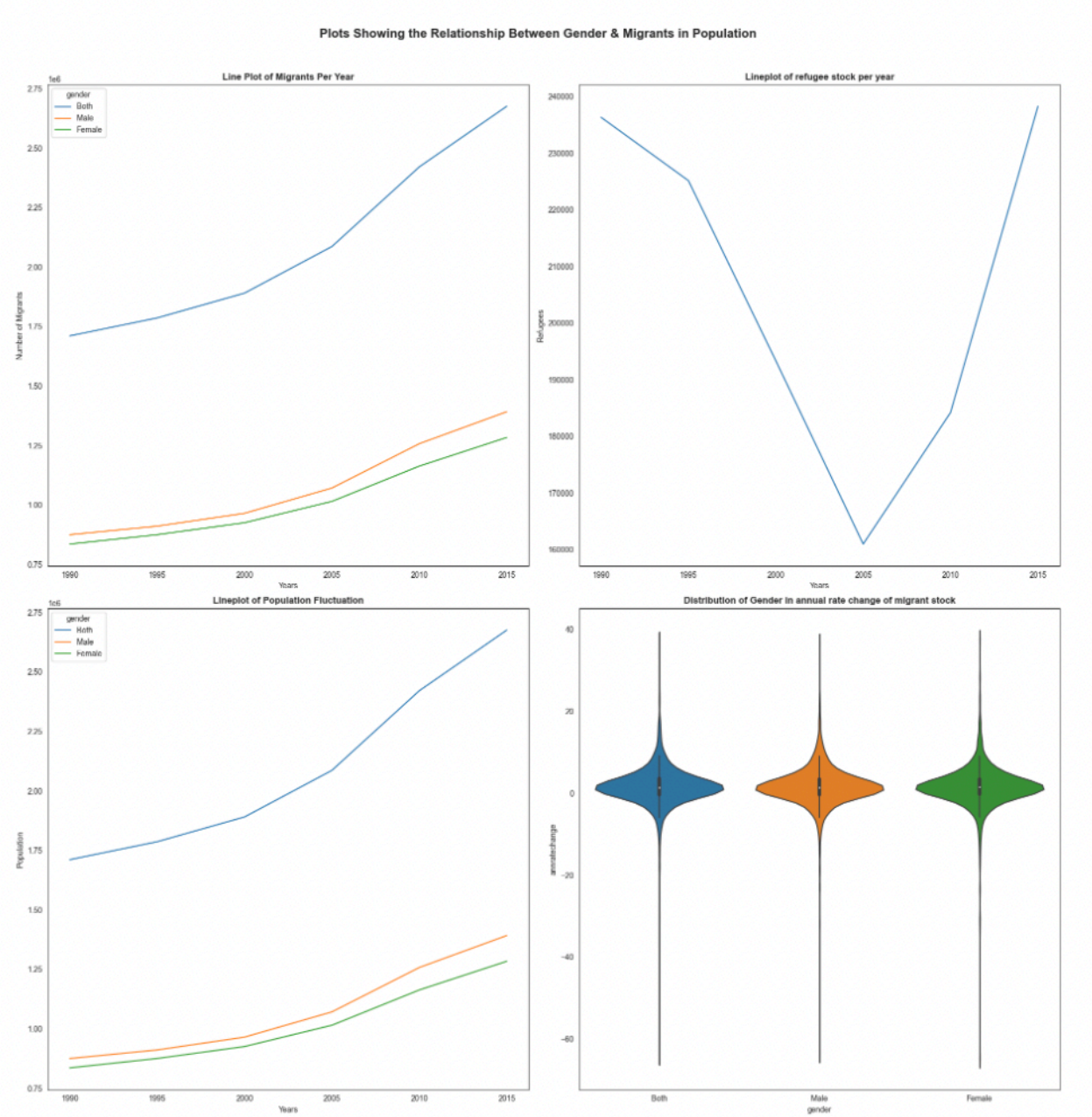
Here, while the visualization is much simpler to follow because there only 5 categories, other than offering which category hosts the most migrants, the numbers are not clear (0, 0.5, 1, 2 etc.). This goes back to one of the first issues I encountered, which is that dealing with exponential numbers is not ideal and that standardizing them properly would offer more precise results. The information of this boxplots tells us that there are more migrants in developed regions than in developing regions. Least developed countries have the least amount of migrant stock.



This catplot offers insight on the percentage of migrants within the population of country categories or regions. It is extremely interesting as it demonstrates that there has been little next to no fluctuations in the percentage of migrants versus population numbers from 1990 to 2015. Thus, despite their being an increase in migrants, as will be demonstrated lower, the population also increases, making that the ratio or percentage is relatively the same. This is information I would not have expected and thus would bridge a knowledge gap, as I would have assumed that the number of migrants would have augmented due to the deteriorating conditions of certain countries due to ongoing civil war and climate change. It also follows another visualization principle:

Tufte Principle #3: Small multiples reveal, all at once, a scope of alternatives, a range of options. (p.67)

While it is not the best example of small multiples because there is relatively no change between 1990 and 2015, the plot below demonstrates this principle more at length.



Here, we can better analyze the fluctuations between the population and migrant numbers between 1990 and 2015. The lineplots act as a checkpoint for the previous catplot, as effectively, it seems both the migrant and population numbers are augmenting at a very similar rate per year. Furthermore, we can analyze the number of refugees per year, which suggests 2005 was the year with the lowest amount of refugees within a 25 year period. The refugee stock is much less in sync with the population fluctuation, which suggests there is another facture that dictates the number of refugees that is unrelated to the natural ebb and flow of population fluctuations.

Failures & Experimentation:

After having analyzed from every possible angle more simplistic measurements like gender, years, migrants, population numbers without touching “countries” as a parameter, I tackled larger datasets and began exploring the use of more complex parameters that required grouping and filtering to get a smaller segment of countries to work with. I began by looking into “binning” the numbers based on a condition to categorize the population and migration numbers into smaller and more workable ratios:

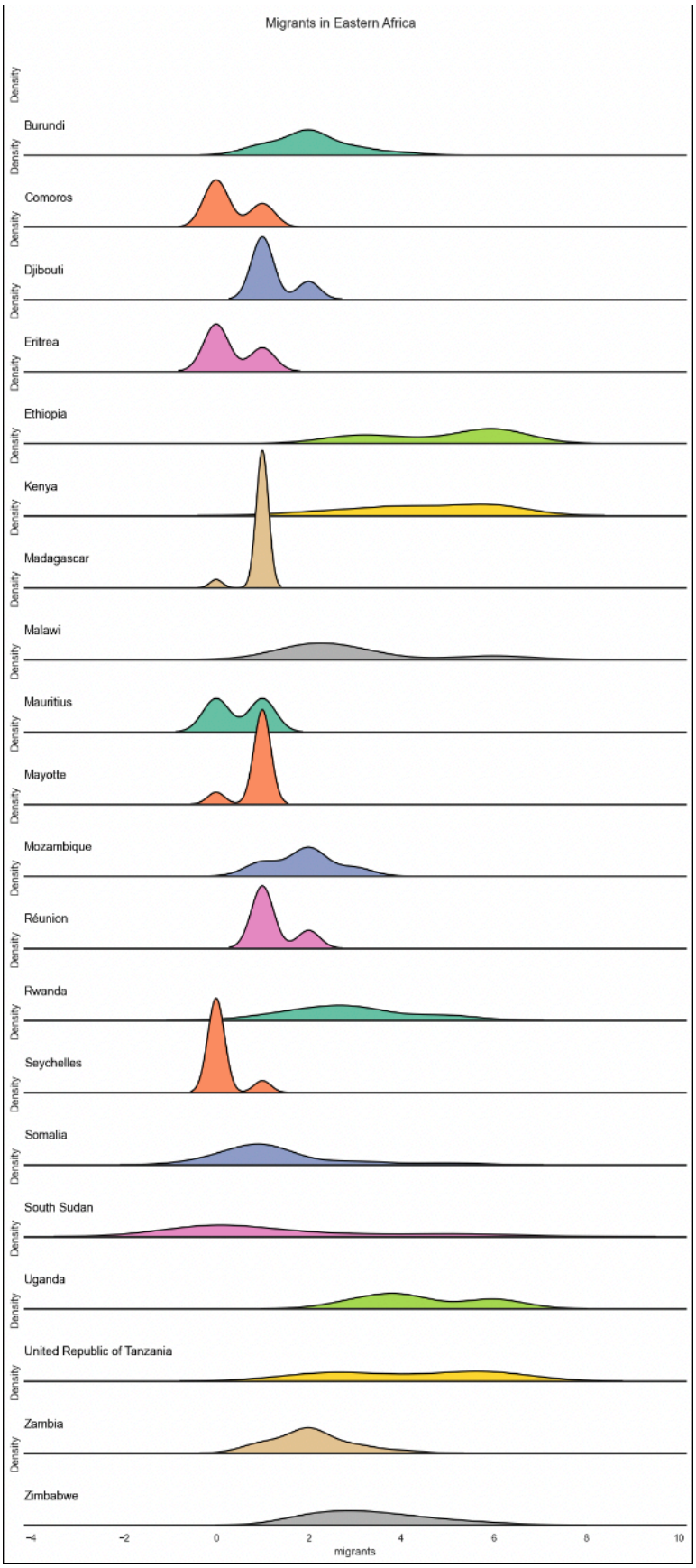
```python
# Binning of the data based on a condition
df_t1.loc[df_t1.migrants < 10000, 'migrants'] = 0
df_t1.loc[(df_t1.migrants > 10000) & (df_t1.migrants < 100000), 'migrants'] = 1
df_t1.loc[(df_t1.migrants > 100000) & (df_t1.migrants < 200000), 'migrants'] = 2
df_t1.loc[(df_t1.migrants > 200000) & (df_t1.migrants < 300000), 'migrants'] = 3
df_t1.loc[(df_t1.migrants > 300000) & (df_t1.migrants < 400000), 'migrants'] = 4
df_t1.loc[(df_t1.migrants > 400000) & (df_t1.migrants < 500000), 'migrants'] = 5
df_t1.loc[df_t1.migrants > 500000, 'migrants'] = 6
```

Python

```python
# Binning of the data based on a condition
df_t1.loc[df_t1.population < 10000, 'migrants'] = 0
df_t1.loc[(df_t1.population > 10000) & (df_t1.population < 500000), 'population'] = 1
df_t1.loc[(df_t1.population > 500000) & (df_t1.population < 1000000), 'population'] = 2
df_t1.loc[(df_t1.population > 1000000) & (df_t1.population < 5000000), 'population'] = 3
df_t1.loc[(df_t1.population > 5000000) & (df_t1.population < 10000000), 'population'] = 4
df_t1.loc[(df_t1.population > 10000000) & (df_t1.population < 20000000), 'population'] = 5
df_t1.loc[df_t1.population > 20000000, 'population'] = 6
```

More so, I tried experimenting with online tutorials to achieve better visuals. I well understood Tufte’s other visualisation principle: Principle #4 No clutter and no junk information, the goal of a data visualisation is to gather information, not demonstrate your skills as a designer.

Online tutorials can send you down the rabbit hole and make you forget that underrated principle. Thus, some of my experiments were successful, while others were not. I did begin a process of filtering and grouping which made the information much more valuable. The plot below has successes and failures.



Seaborn's RidgePlot

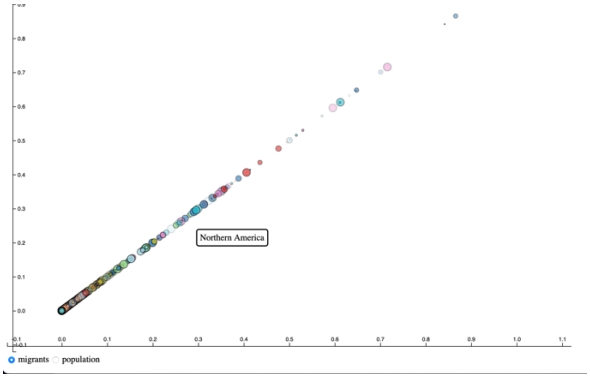
First, it is a great way to compare as per Tufte's principle#1, Eastern African countries. It demonstrates a good use of the groupby and filtering methods to segment and extract useful information. Secondly, the use of colors and the ridge effect makes for a great comparative. However, the standardized numbers turned into a density plot, which can be useful to demonstrate fluctuations in migrant stock from country to country, but offers no concrete numbers, no years, and no gender. Thus, other than being nice to look at as an overview plot, it does not serve the same useful information that simpler scatterplots, catplots, or barcharts can offer. The ridgeplot was therefore not the most efficient way of translating the information at hand in this case.

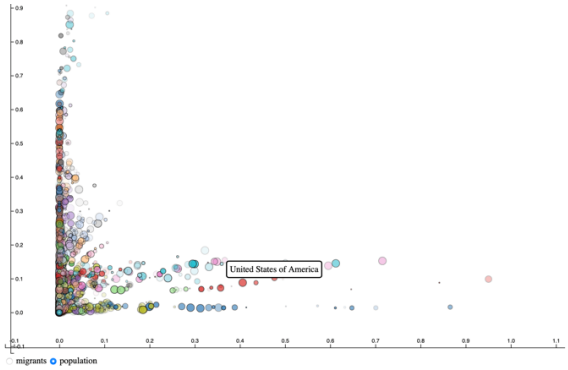
I attempted to use the libraries JoyPlot, d3Blocks, and Bokeh; all of which I had functional success, but not visual success. My reflection on these libraries is that they are 1) very advanced for my relatively still burgeoning skillsets, but 2) not all plot types are useful for all datasets. Each dataset is unique and the numbers you are using will have various high priority needs versus lower priority needs. More than this, often these 'fluffy' tutorials are more fluff than information, which is exactly what we to avoid. The mantra should be to make sure we are mastering basic visualizations with useful information, rather than littering with fluff without any useful

information. Below with D3Block, I was attempting to demonstrate in a dynamic way the difference between migration and population numbers. What was interesting about this library was the possibility of doing a plot that moves with colors demonstrating the countries (each clickable) and the circle sizes representing either the population number or the migrant stock. That is the main issue with this visualisation: I simply cannot explain what I have represented. However, it was the closest I got to my initial idea which was to create a clustermap with an actual map in the background of each country in the data base, the circle sizes representing the number of migrants, refugees, or population.

Unfortunately, I am not yet skilled enough to achieve this ideal visualization that would perfectly match Tufte's main principles of 1) Comparing data and using small multiples 2) Layering with high information priority versus low information priority differentiated in color schemes (see figure below) 3) Using color for measurement .

D3Block failed experiment:

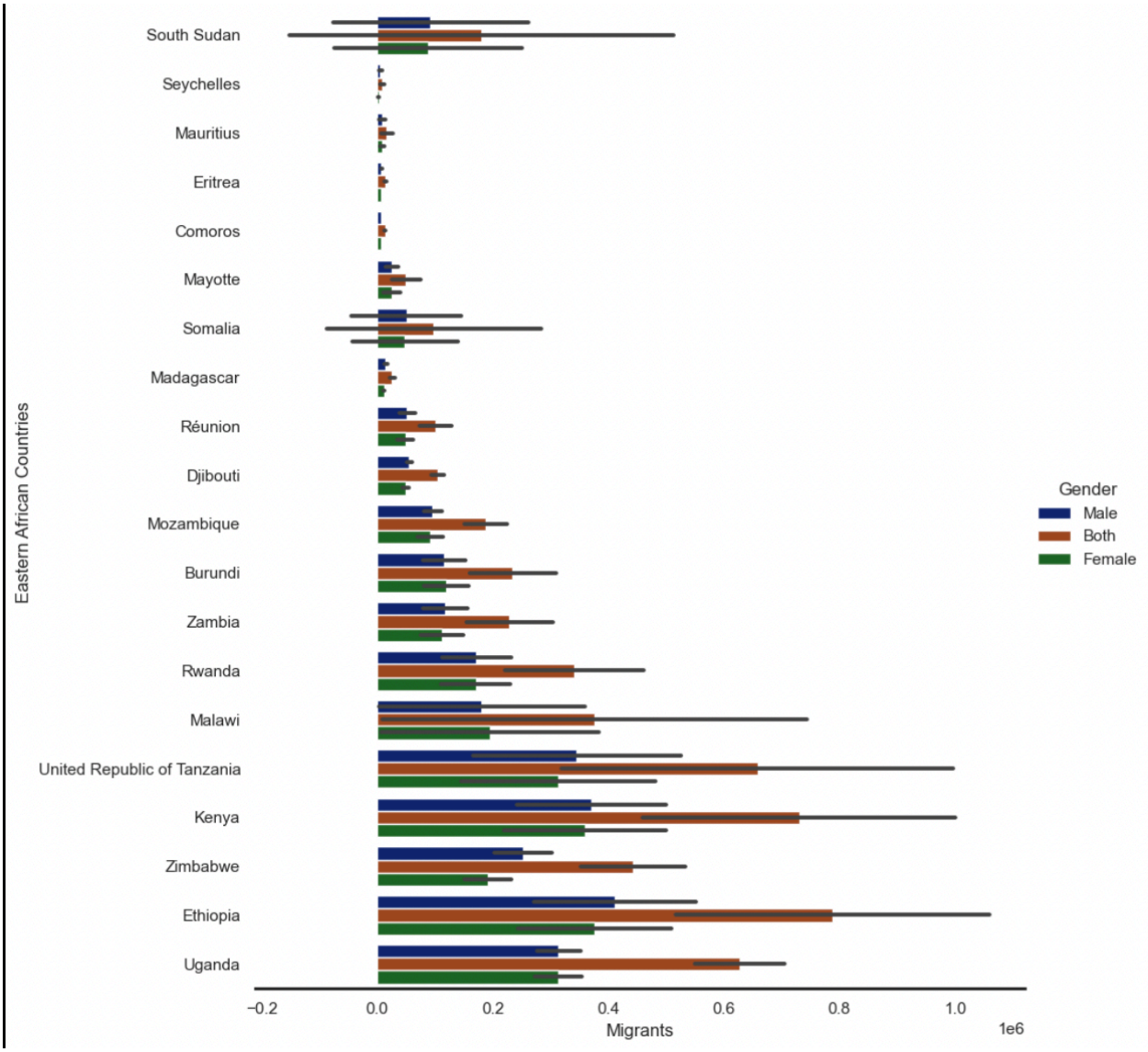




Tufte's discussion on color composition and layering (p.57)

4. Results

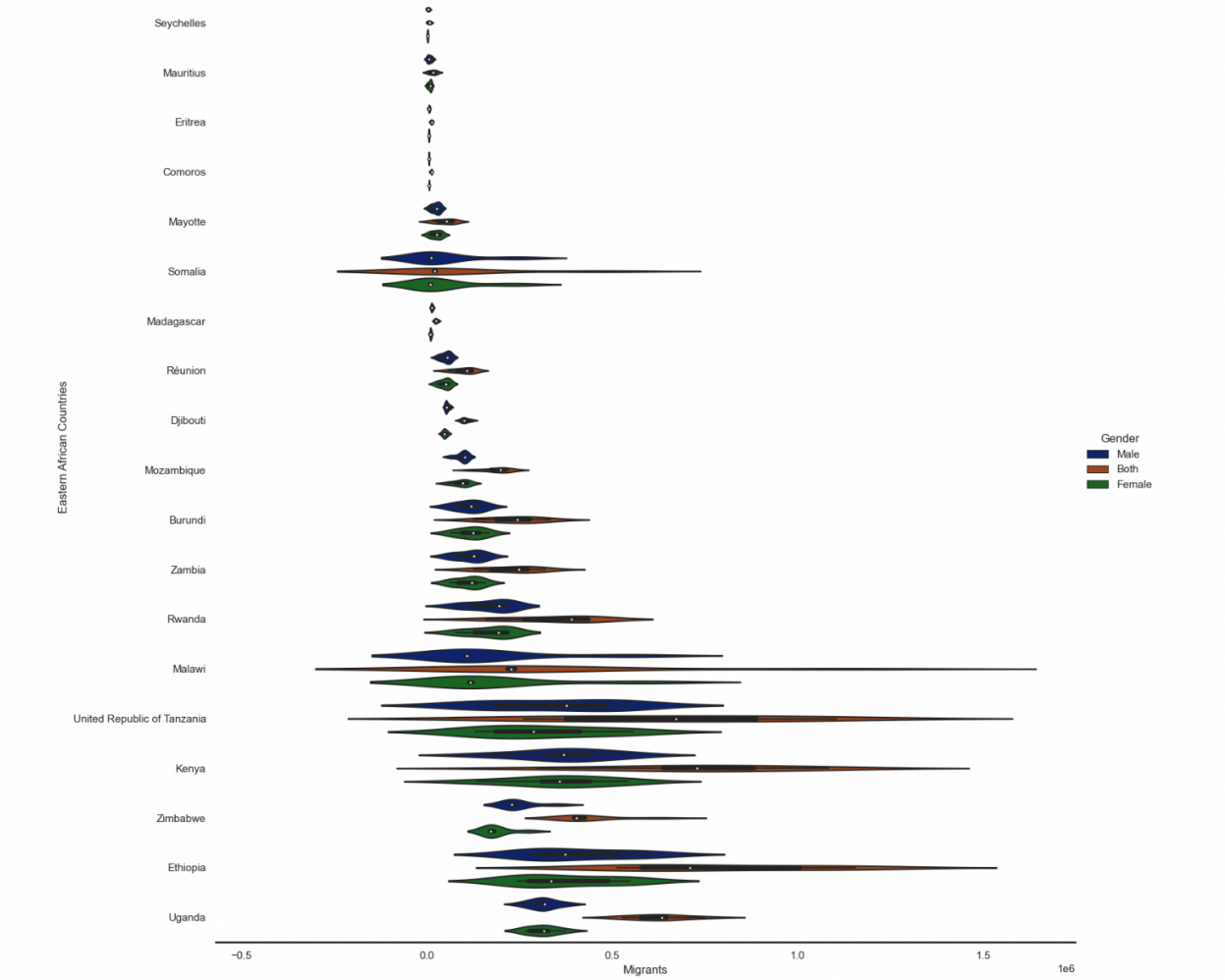
In this last section on visualization, I present what I think are my best visualizations going back to basics with bar graphs. This extrapolative process was essential in my learning curve to understand that sometimes too much is like not enough, which I think Tufte preaches best. His book repeats that harmony and balance is the trickiest task in visualizations. Bar charts are simplistic, but can sometimes offer the most flexibility because they are common and customizable. More than this, the comparative features ended up being the most useful to represent the UN Migrant Stock database respecting Tufte's principles.



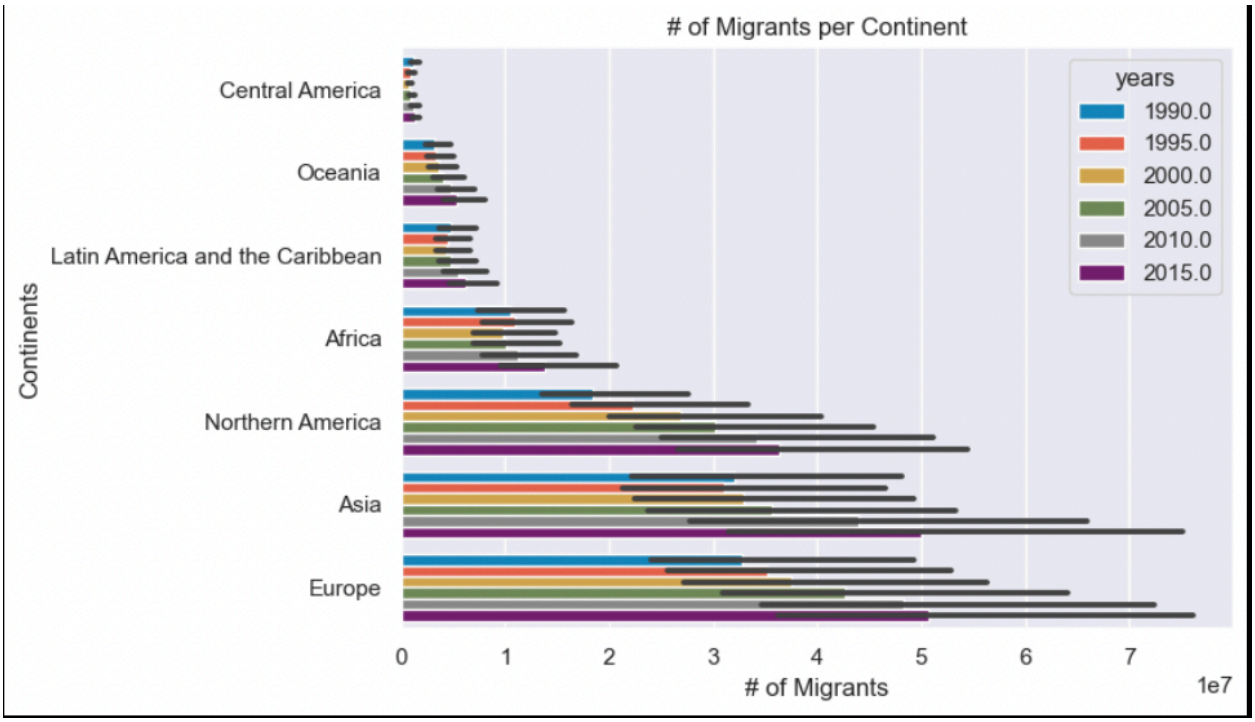
This vertical bar plot is well ordered, demonstrating for example that Ethiopia has the highest number of male migrants within Eastern African Countries, whereas Kenya and Ethiopia are neck to neck in number of female migrants. On the other hand, Seychelles has very low number of migrants. This is a perfect example of how the macro and micro lens work hand in hand, from

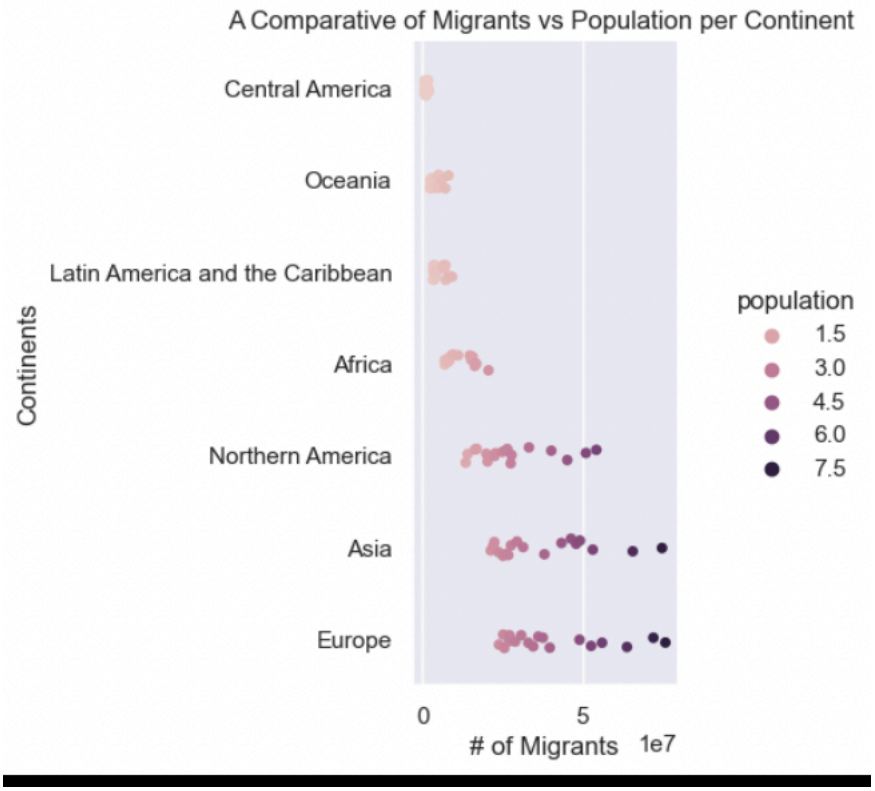
Big Data to Think Data. A researcher might want to take a closer look at why Ethiopia has more migrants than Seychelles: Better job prospects? Economic reasons? Better quality of life? All these factors could be better analyzed with qualitative data, complementary to quantitative data.

A violin plot could also better situate the mean, which alters the data slightly. Malawi has the most outliers.



Finally, if we look at my bar chart *# of Migrants Per Continent* we get a bigger picture view of where migrant populations are most situated. Europe and Asia are the most prized destinations. Again, asking good questions might provide answers as to why that is: Are the policies friendlier to migrants? Why Europe more than North American, proximity? *A Comparative of Migrants vs Population per Continent* might best answer this question, as the population numbers are much higher in Asia and Europe. Therefore, it would make sense that the migrant concentration is higher. In any case, these last visuals are filtered and segmented, which helps to make them much more conducive for research and analysis.


# of Migrants per Continent
Central America
Oceania
Latin America and the Caribbean
Africa
Northern America
Asia
Europe
Continents
years
1990.0
1995.0
2000.0
2005.0
2010.0
2015.0
0
1
2
3
4
5
6
7
# of Migrants
1e7


A Comparative of Migrants vs Population per Continent
Central America
Oceania
Latin America and the Caribbean
Africa
Northern America
Asia
Europe
Continents
population
1.5
3.0
4.5
6.0
7.5
0
5
# of Migrants
1e7

5.Reflections:

In conclusion, this final project was the culmination of what we have learned and put in practice all semester, from tidying, manipulating, curating, analyzing, and visualizing. The opportunity to build off the same project for the entire semester was also quite useful to learn the dos and do nots. My final project took a lot longer than anticipated and was much less about constructing visualizations, but much rather about reorganizing, re-tidying, and ultimately understanding the data to better segment or curate it. My workflow was considerably slowed down because I did not complete the tidying steps in an ideal way. With more time, I would have restarted the tidying process from scratch, understanding what it now means to create code that will be legible, reusable, and making sure to leave the data in workable conditions, i.e., checking if the numbers are floats, making sure strings are strings and not objects or other types of data. Most importantly, I would have cleaned my worksheet not by step or task, but rather divided per all the work done on one table, as the incessant scrolling and littered work sheet hinders the workflow. I had to keep column names short to remember them and code faster, but this also means you consistently lose track of where you are in your worksheet and despite clear titles, it is difficult to find in large coding sheets. Lastly, I set an unrealistic expectation from the task at hand and singlehandedly underestimated the importance of the tidy data process. This reflexivity has been essential to understanding the process of human-centered data sciences.