Rui Hu 1002147376

INF1340H: Programming for Data Science

Professor Shion Guha

Final Project

Nov 14, 2022

1. Introduction

Based on the definition in Wikipedia, "exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods." EDA is a powerful tool or method used by a Data Analyst to better understand the story hidden behind a complicated dataset. The first two steps of EDA, which are data sourcing and data cleaning, are done in the mid-term project. Some slight changes are made to the clean dataset for better visualization purposes. The next step, also the key to the whole analysis, is data visualization.

According to Tufte's Principles, visuals should be clear, and detail-labeled to maintain graph integrity. A good graphical representation should maximize data-ink and erase non-data-ink. Thus, remove unnecessary words and unrelated data and keep the most valuable information on the plot. More details of Tufte's principles shall be discussed in the following section.

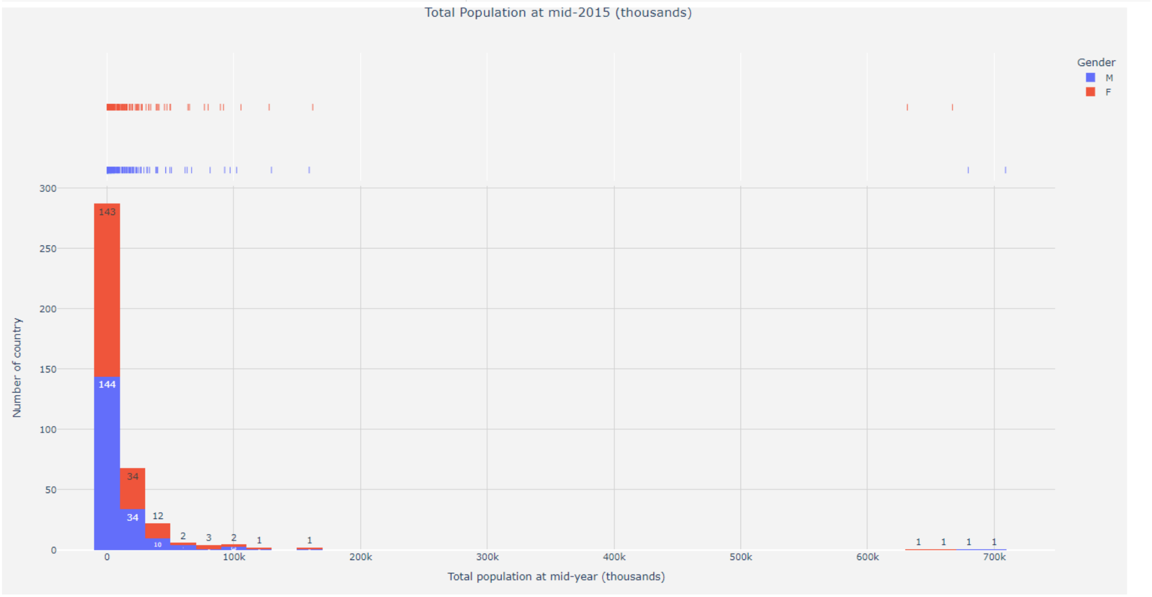
2. Methods

The data for this project is based on migrants stock for each country from 1990 to 2015. I used 6 different plots to analyze data from different aspects.

2.1 Histogram

The first one is histogram, which is one of the most commonly used plots in analysis.

The plot at the bottom is a vertically stacked histogram, classified by gender. The chart shows the total population of each country in 2015. Populations of most countries range between 0 to 100,000 (male and female separately). Two outliers on the right are China and India. The plot on the top is the rug plot. It will put all points on the horizontal axis based on their values. From this plot, we can see how each point is distanced from the other, instead of grouped in bins in the histogram. And it is easier to compare male and female datasets. The female population is less than the male population in most countries in 2015, except for the USA. Normally, we can also use histograms to build density plots. Instead of the count, use probability on the y-axis to prevent distortion by outlier effects.

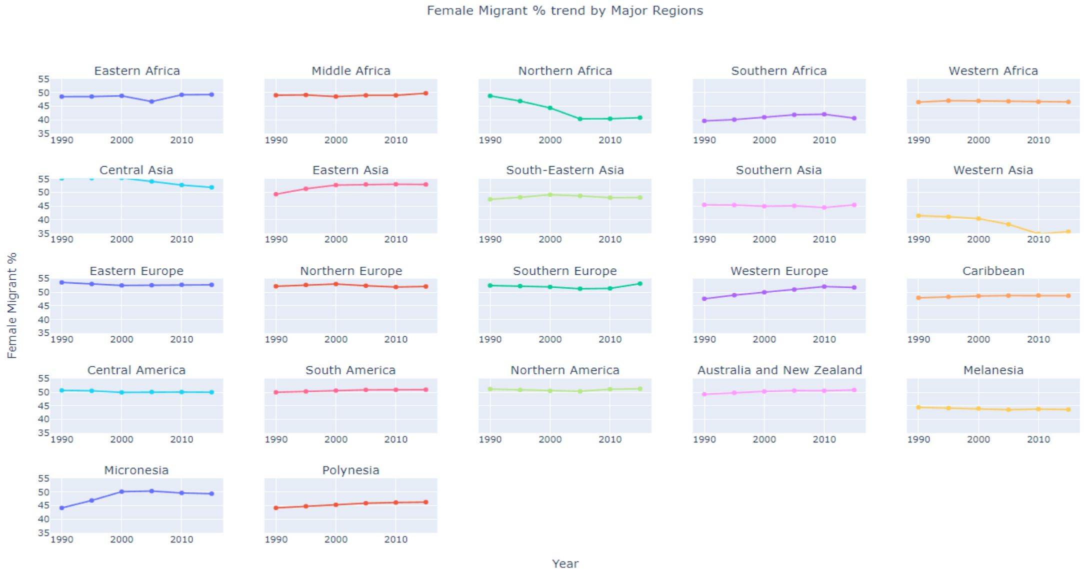


This plot aligns with Tufte's Principle in the following way. Meaningful titles are given and data legend is clearly displayed on the side. X- and y-axis have reasonable intervals and the total population number is formatted in thousands for easier reading. I set the background color to white and the grid line to light gray so they won't overlay the actual data point.

When moving the mouse on the data point, more detailed information will be displayed. I will not put the screenshot in this write-up. Feel free to try in the actual notebook.

2.2 Line Chart

The second plot is the line chart. This type of plot is normally used for time-series data and to show trends of data change over time. I use this plot to analyze the female migrant % change over time in major regions. Country-level data is too detailed and requires too many plots. So I chose to use major-region-level data and use a subplot to put each region on a separate plot, to avoid seeing twenty lines on one single plot.

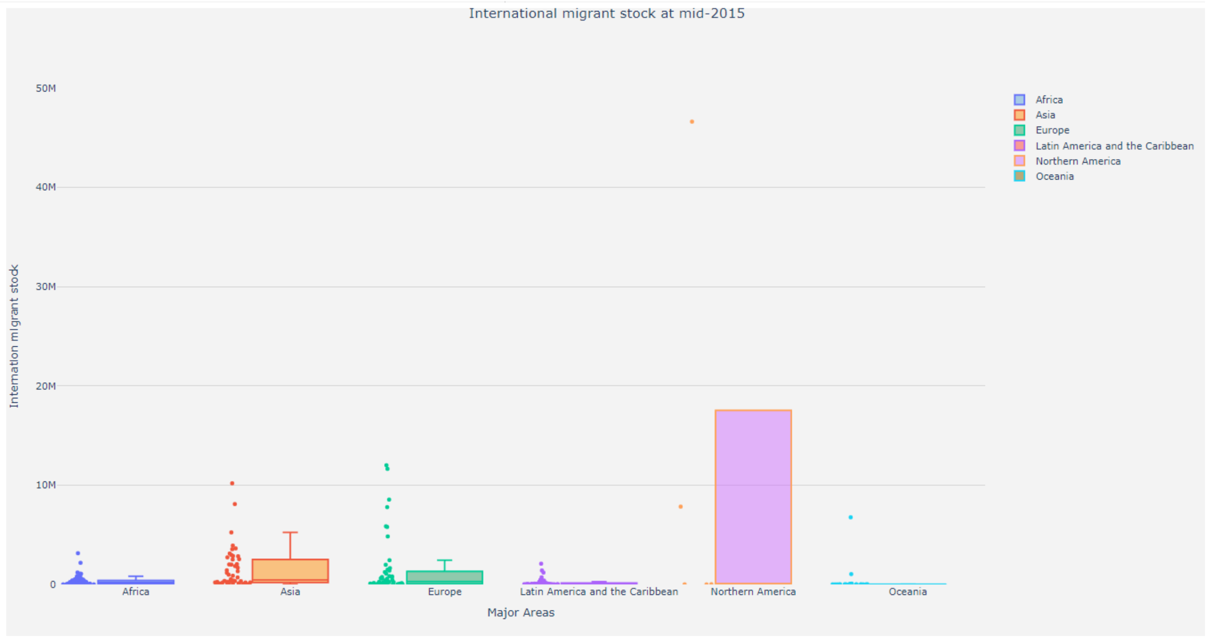


From this plot, we can see that the female migrant % in Central Asia is over 50% and slightly drops over time. % in Western Asia was 41% in 1990 and dropped to 35% in 2015. In all European areas, more than 50% of migrants are female. Line charts allow trend comparisons across different regions.

All y-axis are shared with the same range and general titles are given to the x- and y-axis. I removed legends because it will be hard to distinguish each line color and look for it in the legend table. Instead, I set subtitles for each plot to the region it belongs to. I use markers to clearly display each data point and use a light color for the background and grid lines.

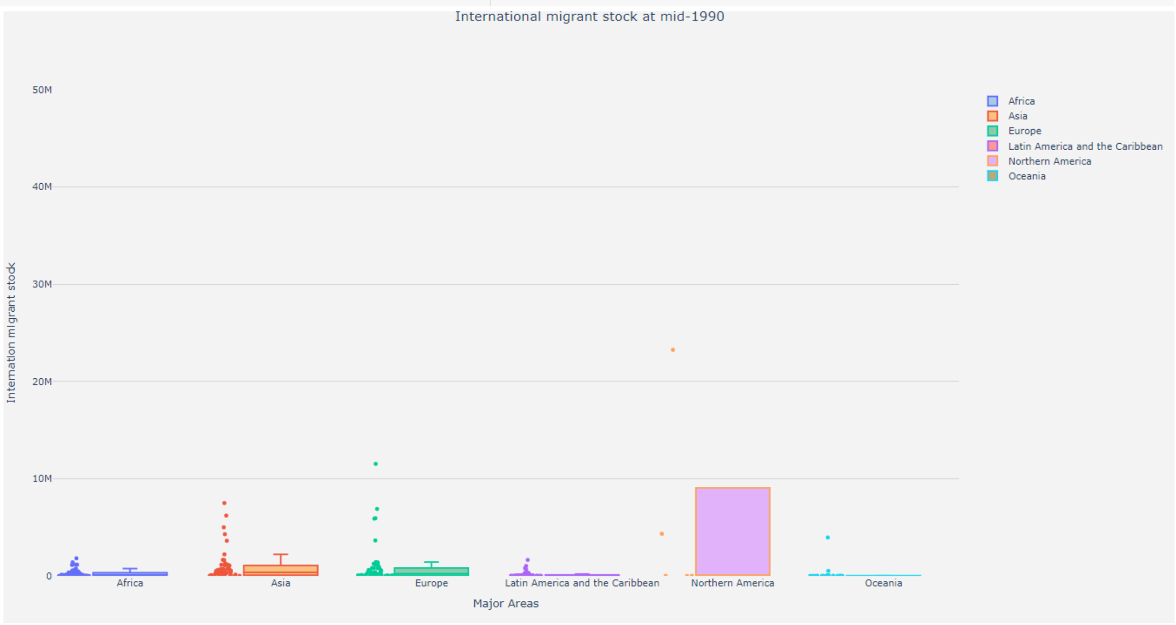
2.3 Box-and-Whisker Plot

A box plot is used to show the distribution of data points and can be used to measure different groups. The distribution is measured through quantiles. It is easier to find outliers with this type of visual.



I chose the box plot for the analysis of the international migrant stock for each country, grouped by major areas. Since box plots will not fit for time series data, I chose to only use data from 2015. If the comparison between different times is necessary, simply change the year filter and create several plots.

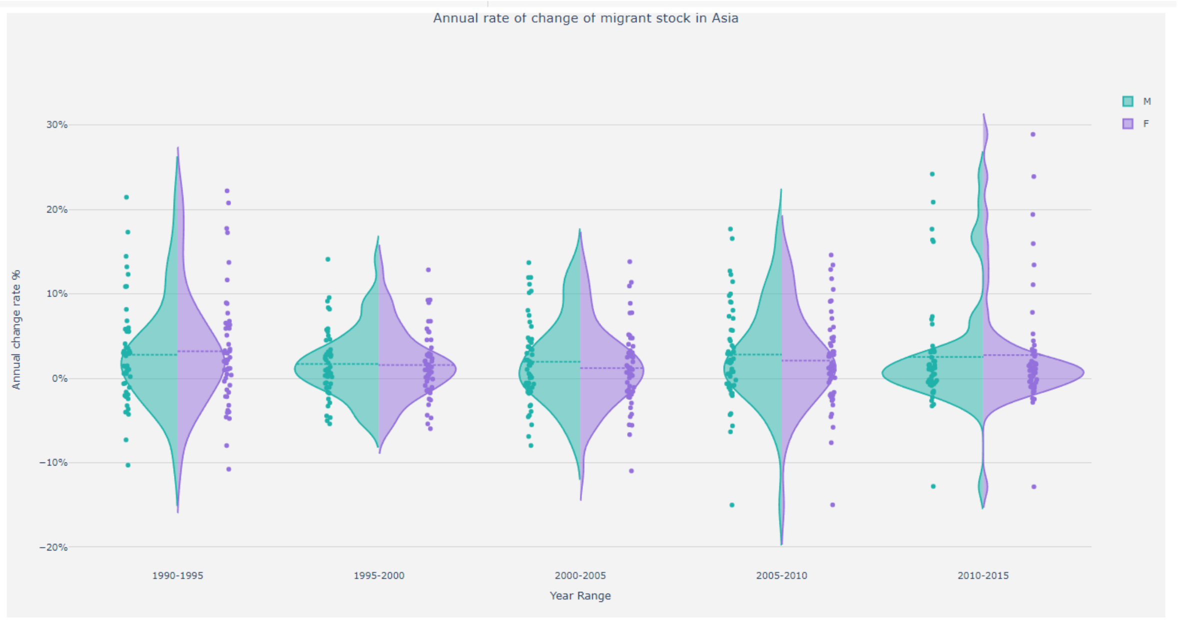
From this visual, we can see that Oceania, Latin America, and Africa have the least number of immigrants. The outlier in Oceania is Australia. North America is the most popular for migrants to come and the outlier is Canada with over 40M migrant stock. Larger the box, the higher the migrant stocks for most of the countries in that major area. Asia and Europe are showing similar patterns in 2015. If we change the year to 1990, the number in North America is reduced by half, but the shape of the boxes remains similar.



Axes in the visual are clearly labeled and a meaningful title is given. The Y-axis is formatted in “millions” to avoid showing a list of zeros. Each box is colored differently and each point is visible. Detailed information on each data can be viewed by moving the mouse on the data point, so I chose to erase the number label for each point.

2.4 Violin Plot

The Violin plot works similarly to the box plot. It shows the distribution of data points instead of quantiles as boxes. Thus, I chose this visual for the analysis of the annual rate of change of migrant stock in Asia from 1990 to 2015.



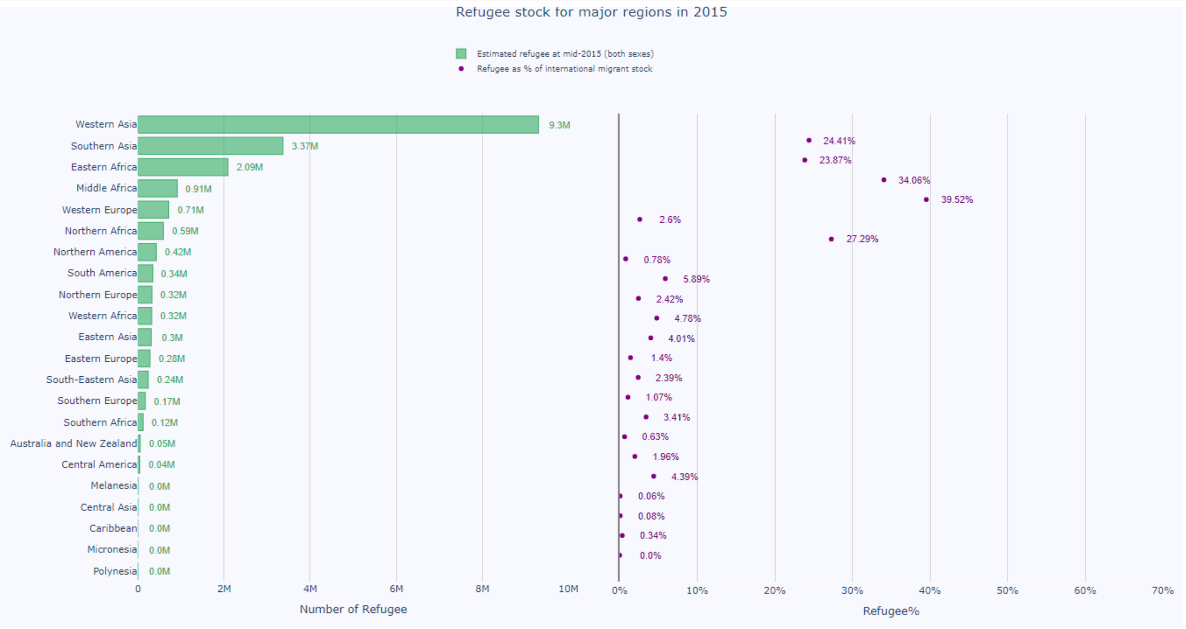
I used a box plot to compare numbers between different major areas at the one-time point and used a violin plot to compare trends over time within one major area. This demonstrates two usage of similar visuals. If interested, plots for other major areas can be created and compared horizontally.

From this plot, we can see the mean for male and female migrant changes are positive over time and staged around 3%. Thus we can say, on average, migrant stocks of all countries in Asia increased by 3% every five years. The “2010-2015” figure has a fatter belly and a longer tail upwards. This means several countries have large positive changes and most of the countries ' change rates are close to the median value. Data points near the center are less spread. On the other hand, data in “2005-2010” are more evenly spread out. Female migrant change rates are higher than male migrant change rates in ‘1990-1995’ and ‘2010-2015’.

Similar treatments on labels and titles are performed. Y-axis is formatted in percentages.

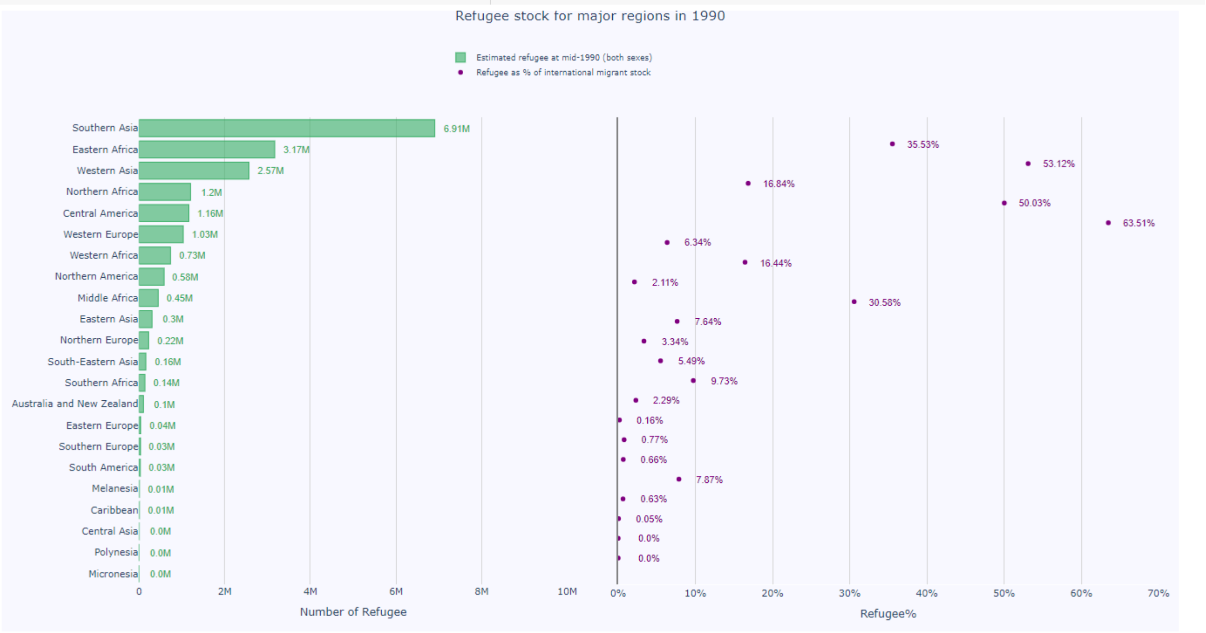
2.5 Horizontal Bar Chart and Scatter Plot

I found that the vertical bar chart is easier to compare different categories or view the distribution of data points. In contrast, horizontal bar charts are easier to show ranks of different categories. I combined a bar chart and scatter plot to create a comparison view between the actual refugee number and the percentage of refugees. I chose 2015 as a time filter and aggregate to the major region level to reduce the number of bars displayed at once. This visual can also be performed by scatter plotting, putting an actual number on the y-axis, a percentage on the x-axis, and labeling all points by name.



Western Asia and Southern Asia are the main source of refugees with a total of 12.7M in 2015. However, this only accounts for 24% of their total migrant stock. Middle Africa has 0.9M refugee stock and this accounts for 40% of its total. North America has 0.42M refugees which accounts for 0.8% of the total. Regions have high refugee numbers and a high percentage are mostly less developed countries and wars happen a lot. For wealthier places like North America, fewer refugees and more general migrants are coming, which will push up the total migrant stocks and reduce % of refugees. This result aligns with the real-world situation.

If we look at 1990, the refugee number was reduced by half, but still exists. Fewer people were migrating between countries, and this leads to a higher refugee percentage in all regions. For some Asia and African places, the ratio spiked over 60%.



Labels and titles are treated similarly to before. I added a number label to each bar and point for a better visual experience. I chose to keep 2 decimals for all numbers to reduce redundant data-ink without loss of data precision. The X-axis value range is fixed and it is easier to compare across different time points.

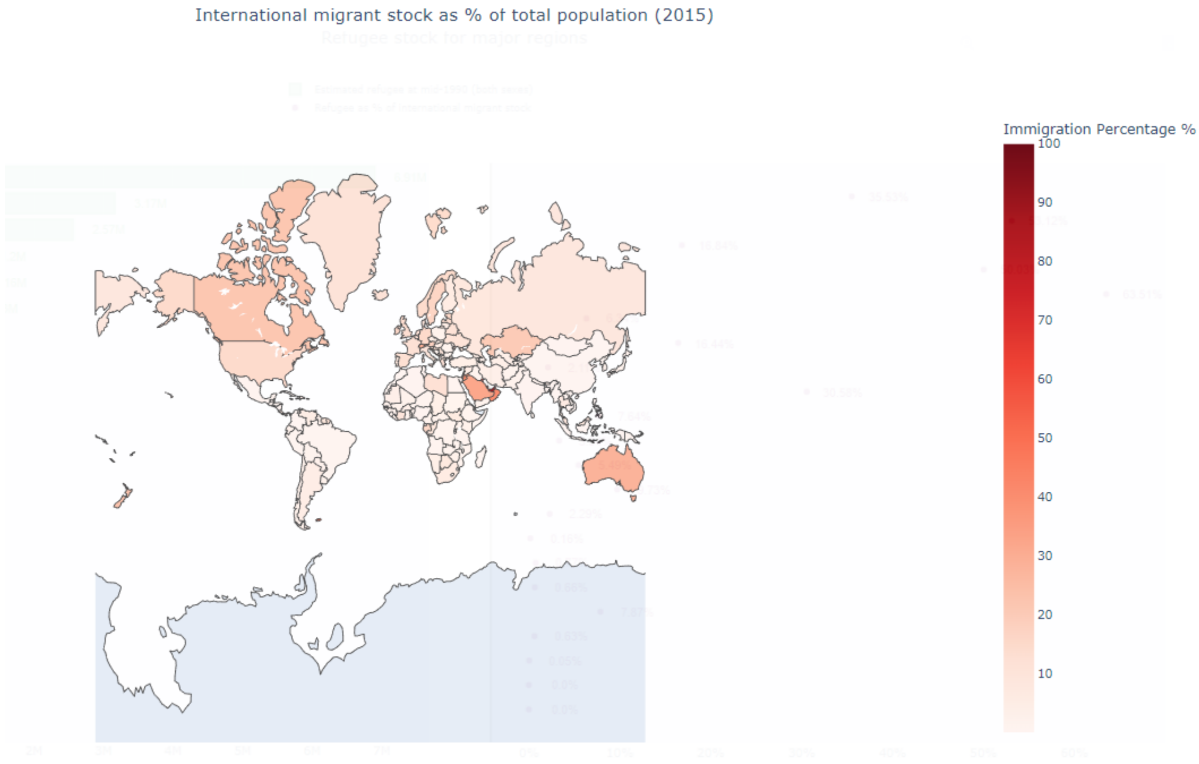
2.6 Choropleth Plot

Lastly, I chose to use a choropleth plot for the visualization of migrant stock % of the total population. I set the time filter to 2015. A Choropleth map is a map composed of colored polygons. It will visually show which areas have higher numbers and which have lower ones.

Darker the color, the higher the percentage of migrants. We can see most North American countries are colored in orange, especially Canada, which reached 21.9%. In contrast, China only has 0.7% and is colored pretty much white.

A Choropleth map is a straightforward way to compare different geo-locations. It cannot be applied if the data is based on other categorical ways.

Detailed numbers can be viewed by moving the mouse onto locations. The title and legend are clearly named.



In summary, all plots in this project are clearly and efficiently visualized. They all comply with Tufte's Principles. Detailed analysis of each plot and why they are chosen have been discussed. The following section will be based on the overall result and next step.

3. Result

In general, Northern America has the largest migrant population, followed by Asia and Europe. China and India have the largest total population. The trend stays constant over years. Female and male migrant size doesn't show significant differences in all years. Less developed areas like Africa or Southern Asia have high refugee % due to geographic and political conditions. Female migrants % are less than males in most areas of Africa and part of Asia. These results are not super complicated but will be helpful in future stages like constructing models or in-depth analysis.

4. Discussion

From this project, I learned about Tufte's Principles and how to apply them to different plots.

I believe all my plots comply with Tufte's Principles in the following ways:

Data Integrity: representation of numbers is directly proportional to numerical quantities. All numbers are rounded to a reasonable decimal place. Clear labeling on all axes, legends, and plots. Axes are set to the same intervals and range for better comparison across plots.

Data-Ink: unnecessary legends or words are excluded from plots. Labeling data points only when necessary.

Chartjunk: no excessive use of graphical effects. All graphs are clearly displayed. Use of subplots to avoid too much data presented on one graph. Use distinct colors to separate different groups. Use a light color for the background and grid lines for an easy read.

All plots allow for self-interpretation and for visual comparisons.

Moreover, some details about different graphs I learned from this project:

- **Histogram:** vertical bar chart, easy to view the distribution of data points, grouped into bins.
- **Line chart:** more likely to be used for time-series data, easy to see the trend over time
- **Bar chart:** easy to compare across different categorical groups.
- **Box chart:** show data distribution by quantile, easy to identify outliers and compare median, Q1, and Q3 across groups.
- **Violin chart:** similar to box chart, shows density plot instead of quantile.
- **Choropleth Map:** mostly used for data with geometric location, visually displays the quantity of data by color or size of the bubble.
- **Rug plot:** put data on either the x- or y-axis, show the distance between each data point, represent numerical value difference, and used for univariate analysis.
- **Scatter plot:** can be univariate or bivariate. Shows a relationship between two numerical data sets, identifying correlation existed.

5. Conclusion

Data visualization is one of the most important skills for a Data Analyst. After cleaning the dataset, one should be able to use data and generate a story for the audience. It is hard to find insights by staring at numbers. Selecting the appropriate visual type and transforming data into graphs will make the whole process easier. Clearly read visuals will also be efficient to share insights with others.

Overall, I believe I have covered all points mentioned by the professor. Please read the whole write-up clearly and try to comprehend it. I have put a lot of thought into this and please do not only look for keywords for marking.