Name - Anshuta R Kulkarni

1. Introduction

This dataset is about the trends in International Migrant Stock: 2015 revision, Department of Economic and Social Affairs, United Nations. The excel workbook contains 8 sheets - 6 tables, 1 'Annex', and 1 'Notes'. The overall objective is to clean the data efficiently and create well-structured visualizations explaining the data and its trends.

2. Methods

For the purpose of this assignment, I will be using the cleaned datasets from my midterm assignment and build my graphs using the two final datasets.
To make data sets easier to understand, the objective is to create graphs that help identify patterns and outliers.

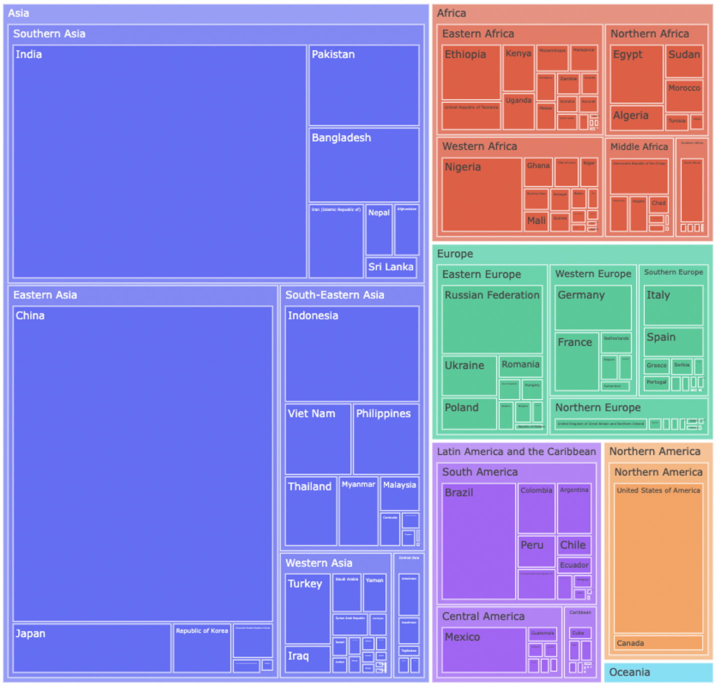
Graph 1:

To start with understanding the breakdown of the information at hand, I wanted to create a visualization that gave an overview of the data. I decided to study the total population values as a base to see how the population was distributed in the data. This would then help us understand the migrant and refugee-based information, relative to the overall total population. To do this, I created a 'treemap' using plotly.express using Data_Final_1 to plot the total population values in a hierarchical order based on the 'Major area', 'Region' and followed by 'Country or area'.

After filtering for 'both_sexes' data for overall data, I started with creating a new data frame - 'Mosaic_DF'; keeping all the relevant variables as stated above and converting the value variable to float type for better readability into the 'treemap' function. Using the 'treemap' function, I filled out the parameters, including stating the hierarchical order for the country information.

One of the features I liked about this visualization was its interactivity. Hovering over each of the panels provided the user with more detailed information about that selected country; while clicking on any panel would fade the remaining panels allowing the user to study that specific path more efficiently.

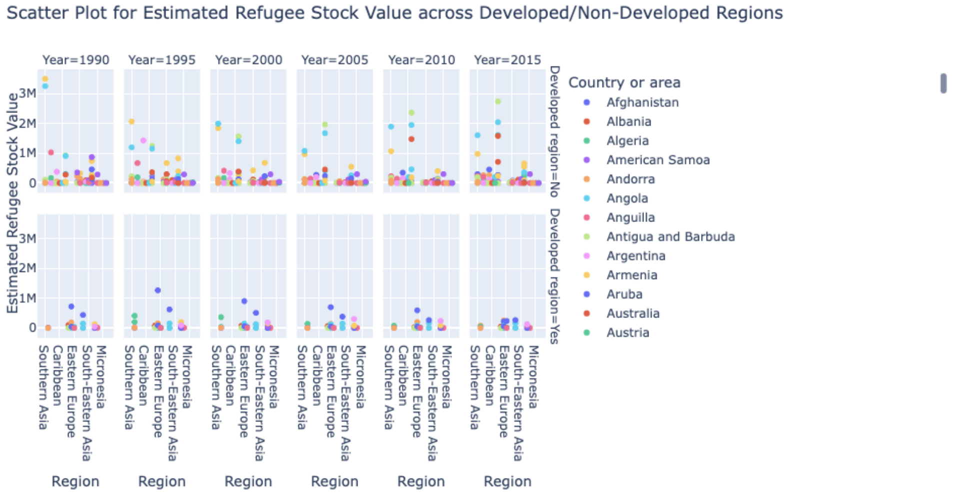
Name - Anshuta R Kulkarni


Population Distribution among Countries (Year 1990-2015)

Since the size of each panel is relative to its total population value, the output shows Asia taking the biggest panel in the overall plot, with China (Eastern Asia) as the biggest contributor among all the countries followed by India with respect to Southern Asia. The smallest section is for Oceania, implying this area has the smallest population among all areas.
Using this graph, we can study the population distributions across major areas and region, while comparing them across different countries in the data. We can see each major contributing country across each region and overall, the major area across the dataset. Overall, this graph helps with understanding the overall proportions of the population in each path of major area, region, and country.

Graph 2:

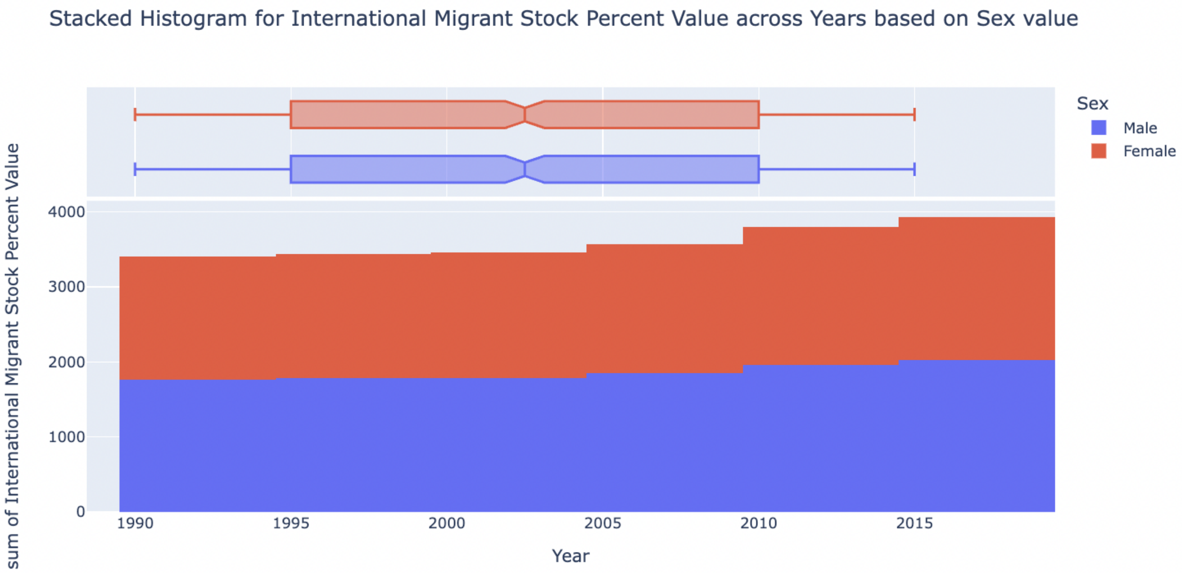
For this graph, I used Data_Final_1 to study the Key Performance Indicator (KPI) – Estimated Refugee Stock Value for both sexes across the regions and countries. Since the variable ‘Developed Region’ was a binary variable (Yes/No values), I decided to use that as a faceting variable so I could relevantly compare the KPI values across countries. Using the ‘Country or area’ variables as color, I created a scatter plot for the KPI to be plotted against all the years in the dataset (1990, 1995, 2000, 2010, 2015) for each of the regions.



The output shows that the spread of the data is much more in non-developed regions when compared to developed regions, which is in line with the number of regions under the non-developed category.
Hovering over the highest point in the whole graph, which is on the top left section of the first panel for the Year 1990 in the Developed region = No – shows more information about that value. This data point is displaying the value for the country – Iran (Islamic Republic of). This data point is closely followed by another value belonging to the country of Pakistan. This is really important information because such high values were seen only once across all those years in the dataset. It would be interesting to understand what caused such a high estimation of these values and study the trend for these across the years.
Overall, this plot helps with analyzing these estimated refugee stock values across years and following these estimations across each year and country for their regions.

Graph 3 :

For this graph, I wanted to study if there were any differences in the international migrant stock percent values over the years among the two sexes – Female and Male in the dataset. To do this, I created a new dataset – Plot_Data, selecting the relevant variables and converting them to their respective types; I filtered the data to remove ‘Both_Sexes’ value so I could use only ‘Female’ and ‘Male’ values required for the graph. I used ‘the histogram’ function to plot the above variables to study their spread across the years and use ‘Sex’ as the differentiating variable for color. I also added the parameter for marginal as ‘box’ to show the overall distribution of each of the sex values.



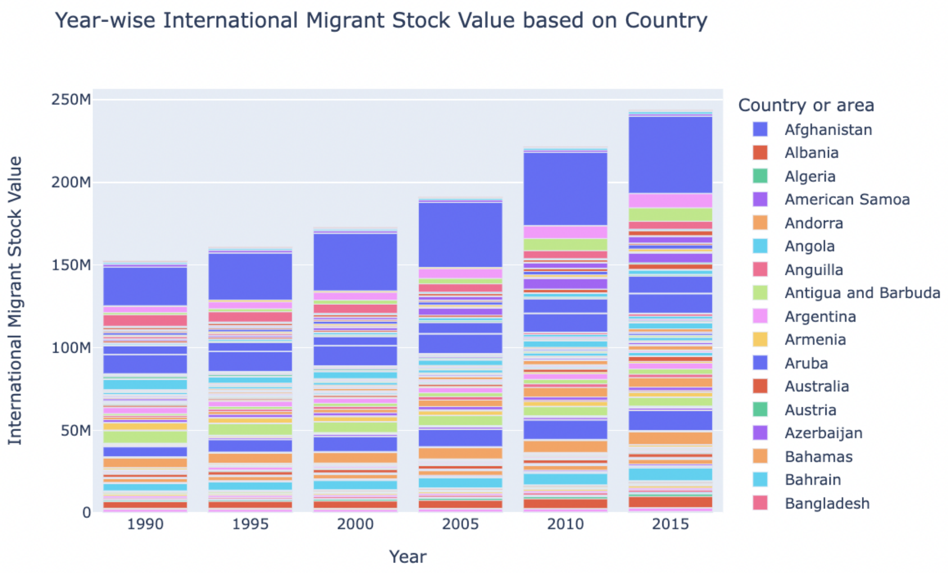
The above output shows similar distribution and spread of the data for both ‘Female’ and ‘Male’ values in the dataset. This comparative graph helps us confirm that there is no significant difference between ‘Female’ and ‘Male’ with respect to International Migrant Stock Percent Value in the dataset.

Graph 4:

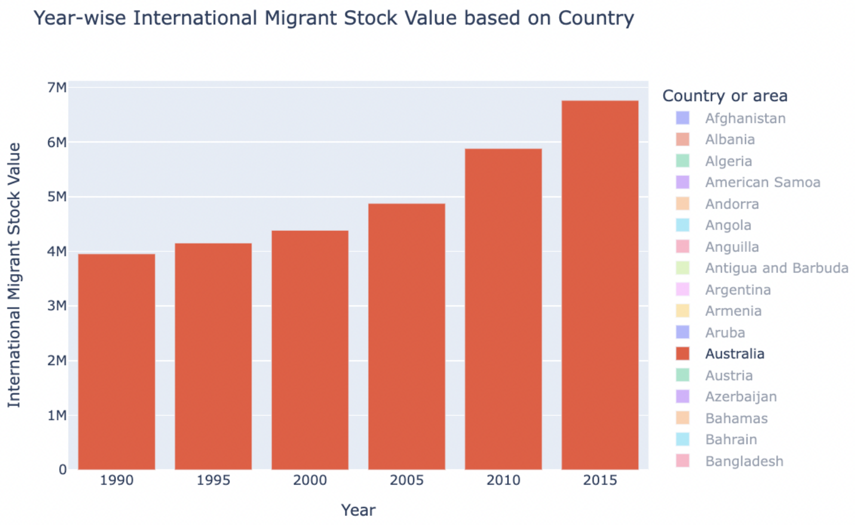
In line with the above graph, I wanted to study the International Migrant Stock Values for each country across the years in the dataset. I used the ‘bar’ function to plot the KPIs values across the years based on the countries.

I reused the dataframe from the above graph and filtered it for ’Both_Sexes’ since I wanted a general overview of the data. Using this data, I created a bar plot and added the color parameter based on ‘Country or area’.

Name - Anshuta R Kulkarni



The above gives the distribution of the KPI – International Migrant Stock Value for all countries across the years. The graph shows really high values for United States of America, showcased as the top blue sections of each of the bars in the above plot. The added advantage of using ‘plotly’ is that it lets the user focus on the data based on legend value selection. Therefore, while the plot does show an overview of the data for all countries, clicking on any of the countries in the legend will update the graph to show values only for that country selection – making it easier to study the trend on the country level as well. For example, the below plot shows the KPI distribution for country: Australia -

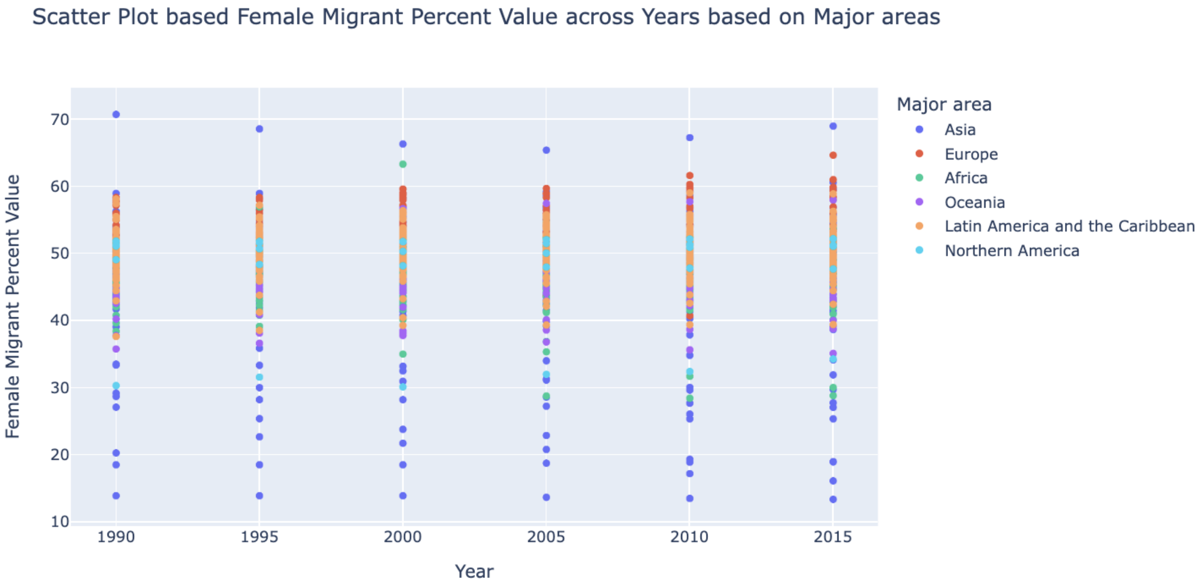


Name - Anshuta R Kulkarni

Graph 5:

For this graph, I wanted to study the female migrant percent values. For a high-level overview, I created a scatter plot using Year and the KPI – Female Migrant Percent Value and used ‘Major area’ as color variable along with ‘Country or area’ information as data point information. This would help understand the female migrant distribution across major areas across years.

I used Data_Final_1 for this plot and created a new dataframe with only valid ‘Female’ values.
Using the ‘scatter’ function, I created a scatter plot for the above variables.



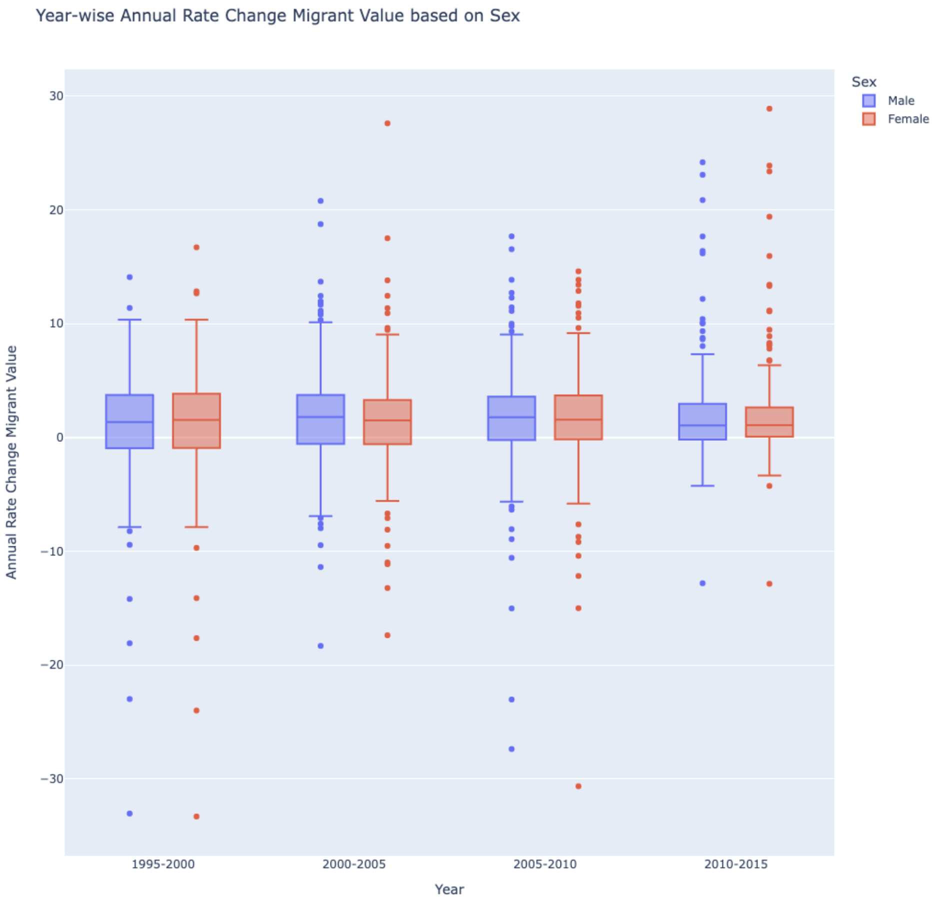
Hovering over the highest data points in the above output plot shows ‘Nepal’ from ‘Major area’: Asia to be the country with the highest female migrant percent values across all the years. The smallest values across all years are seen to be from ‘Bangladesh’ from ‘Major area’: Asia.

Name - Anshuta R Kulkarni

Graph 6 :

For this graph, I wanted to study the KPI - Annual_Rate_Change_Migrant_Value from Data_Final_2. Since we have this information for both ‘Female’ and ‘Male’, I wanted to see if there were any differences in this KPI between these two values across years.

I used Data_Final_2 for this plot and converted the KPI to a float variable. Using the ‘box’ function, I created a box plot for the KPI variable for each year bin, and colored based on the ‘Sex’ value.

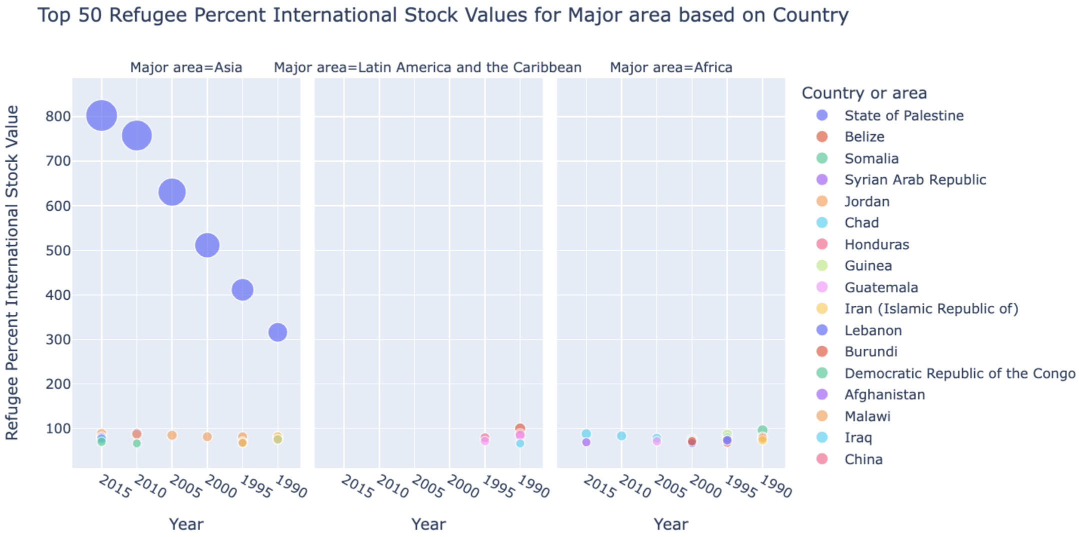


The output shows that the annual rate change migrant value is somewhat similar for both ‘Female’ and ‘Male’ across the years. It is also notable that ‘Female’ values seem to be slightly more spread out than ‘Male’ values for most of the years. For the year bin 2010-2015, ‘Female’ KPI values show to be spread more but have a smaller first and third quartile compared to their respective ‘Male’ values for the same year bin. It would be better to consider some form of outlier analysis in order to understand those extreme values under the ‘Female’ category.

Name - Anshuta R Kulkarni

Graph 7:

For this plot, I wanted to study the Annual Rate Change Refugee value across countries. I used Data_Final_2 for this, and subsetted the data for the top 50 largest values of the KPI. Using the relevant variables, I used the 'scatter' function to plot the KPI values against 'Year' and color by 'Country or area' and finally, faceted by 'Major area'.



The above plot shows the top 50 values to lie in three 'Major area' – Asia, Latin America and the Caribbean, and Africa. Hovering over the largest data points in the plot in the Asia panel, show this KPI value belonging to the State of Palestine. Across all the years, this country shows the largest percentages for Refugee International Stock values, all while progressively increasing value every 5 years.

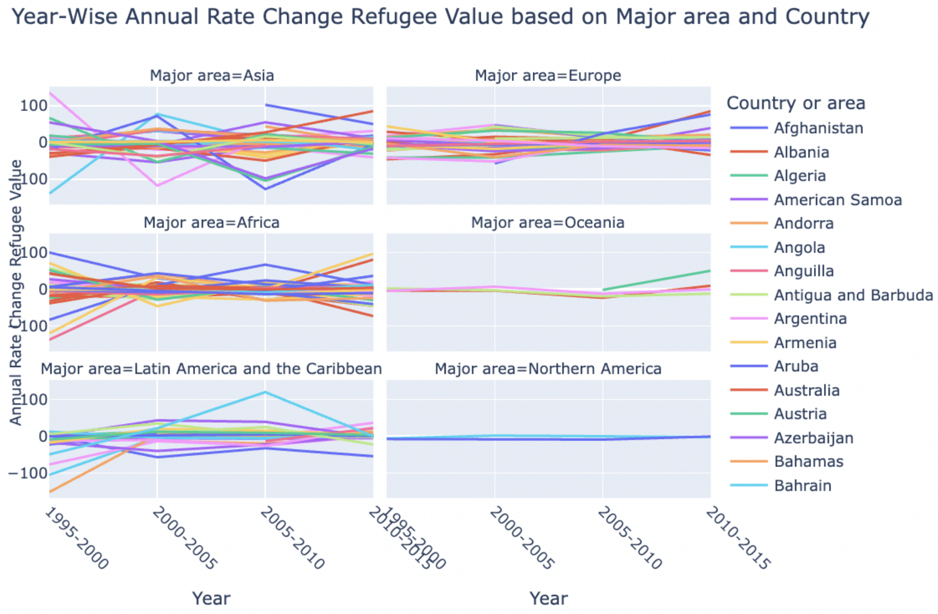
Graph 8 :

For this plot, I focused on 'Annual Rate Change Refugee Value' from Data_Final_2.

I wanted to see this rate change over the years across countries and overall, 'Major area'.

I used the 'line' function in plotly to create a line chart for this KPI across the years, faceted by 'Major area' and colored by 'Country or area'.

Name - Anshuta R Kulkarni



The above graph shows the trends for the Annual rate change for refugees across the year bins. With most of the data in Asia, the panel for the same is heavily populated with trend lines. The most noticeable trend is for Saudi Arabia and Indonesia with extreme highs and lows in consecutive year bins. A similar extreme pattern is observed in Latin America and the Caribbean area, with Venezuela peaking in the year bin 2005-2010 and then dropping low in the next year's bin.

3. Tufte Principle – Small Multiples

The Small Multiples concept refers to the display of multiple images simultaneously, allowing the reader to examine the inter-frame differences immediately, and parallelly.

I found this principle very useful and used it in most of my graphs. Since the data is on such a large scale, it makes EDA quite problematic. Most of the graphs may seem heavily clustered with no method to effectively study their trends. Using higher-level variables like 'Major area', 'Region', 'Sex', 'Developed region' etc., made it easier to split the data and showcase the data efficiently into their relevant sections.

While an overview of the data is important, it is equally necessary to focus on the content of the data. It is imperative to analyze and compare the relevant variables. For example, if I was to understand the relative population and refugee percent across countries in Asia vs countries in Oceania – I would compare based on the relative values within the two regions. This can be done by visually conceptualizing Tufte's principle of small multiples. With the entire concept of making the data easy to understand, using this principle creates a new method of showcasing the data efficiently.

Name - Anshuta R Kulkarni

4. Discussion

My takeaway from this exercise was about learning the importance of data visualization. While cleaning the data and statistical analysis is very important to the data science process, data visualization plays a key role in helping us make inferences quickly and efficiently. Using tidy data principles and principles of minimalism, I can make concise deductions from the 7 tables of raw data using these visualizations.

I would like to point out that I liked using the 'plotly' package and all its features. Especially the function that allowed any graph to be an overview as well as a detailed plot just with a click of a button on the legend. This would be a useful trait since it cuts back on the user having to go through multiple plots to understand the trend efficiently.

I would also like to study these trends better with some more context. For example, in Graph 8, it's seen that the 'Annual Rate Change Refugee Value' for Saudi Arabia drops significantly between two consecutive year bins (2000-2005 and 2005-2010). It would make more sense to look more into this trend and understand if this affected other countries or if there was any reasoning behind such extreme changes. In general, I did notice some extreme values across most plots, so I would like to look more into outlier analysis to understand if they can be categorized as outliers or if there is sound reasoning to keep them as valid data points for further analysis.

5. Conclusion

This overall exercise summarizes the first two important phases of the data science process – data wrangling and exploratory data analysis. While the first assignment taught me the importance of cleaning the data and dealing with white noise efficiently, this assignment has helped me focus on understanding and dynamically delivering inferences from the cleaned data. In line with the concept of storytelling with data, this exercise has helped me see the cleaned data with a new perspective and therefore, build efficient visualizations in order to make the data more concise and accessible when presented to a layman.