Introduction

After cleaning the whole UN dataset, it is time to do the exploratory data analysis. This concept was brought by Tukey in 1977, and he suggested that we need more emphasis on using data to suggest a hypothesis in a test, which is a part of our next stage - building models and algorithms. Exploratory data analysis has a lot of objectives, includes:

- Enable unexpected discoveries in the data
- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the elections of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

In this project, I will use several data visualization tools to conduct the exploratory data analysis, and find some pattern or trend inside this dataset. We will have a discussion about what this exploratory data analysis tells us and try to suggest some hypotheses for the later model constructing process.

Method

Exploratory data analysis has a lot of tools including graphing techniques and quantitative techniques. In this project, our exploratory data analysis will focus on graphing techniques. We used boxplot, scatterplot, barchart, line chart, and violin plot to explore the value inside different variables. Below are the descriptions of each visualization tool.

Boxplot can be used to graphically demonstrate spread and skewness of the data through their quantiles. In the box plot, we will have information of minimum, maximum, first quartile, third quartile, median and dots for outliers.

Scatter plot uses dots to represent values for two different numeric variables. We use the x-axis to represent the first variable, and the y-axis is for the second variable. We can use the scatter plot to observe if there is a further linear relationship between these two variables. In the later modeling construction, scatter plots can also help us to predict future values.

Bar chart is a graph that represents categorical data with several rectangular bars. The height or lengths of these rectangular bars can be used to represent the value of each category. One of the advantages of a bar chart is that it can show the distribution among the data, and it is easy to read and construct.

Line chart can be implemented from the scatter plots since it is a type of chart which displays information through a line connected by several dots. However, there are still some differences between, such as the measurements points of line plots has to be ordered so that we can see the trend of the dataset over intervals of time.

A violin plot can be seen as a hybrid of a box plot and a kernel density plot, which shows peaks in the data. While box plots can only give us the information of summary statistics, the violin plot can further show us the density of each variable, which is the feature of kernel density plot.

In order to construct the above plots, I also imported several Python packages for plotting such as seaborn and matplotlib. In order to make it easier to compare graphs in different areas and time periods, the major method we applied is to use the subplots function. Moreover, we have created some specific data frames used for plotting.

Result

Tufte has developed a set of principles for effective data visualization, which includes following:

1. Show comparisons
2. Show causality
3. Show Multivariate data
4. Use appropriate encoding
5. Maximize data-ink ratio
6. Use appropriate scales
7. Label carefully

Following these principles, I conducted two sets of box plots, one bar chart, one scatter plot, two line charts and one violin chart. According to the principle of label carefully, all these plots come with a proper label and a title. Also based on the principle of showing multivariate data, these plots analyze data in different years, regions. Moreover, they almost covered all columns in our original dataset. For the principle of maximizing data-ink ratio, I only show the necessary grid lines in order to read the plots easier and the legends for multi-color line charts.

These plots did bring us some unexpected discoveries, help us make hypotheses and assumptions, and provide the basis for further data collection and model constructing. In this section, we will focus on the discoveries of the plots. All the future discussions and how these plots inspire us will be talked about in the later discussion session.

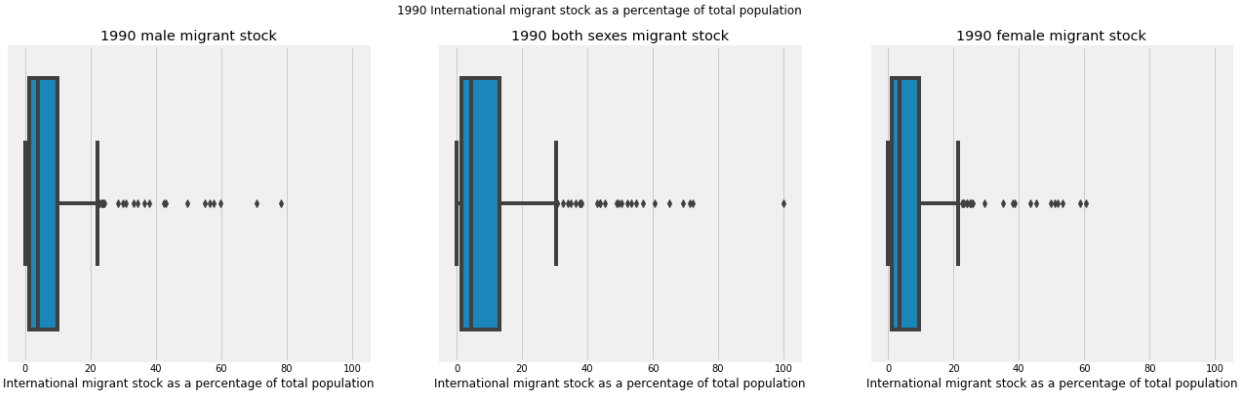


Figure 1: 1990 International migrant stock as a percentage of total population

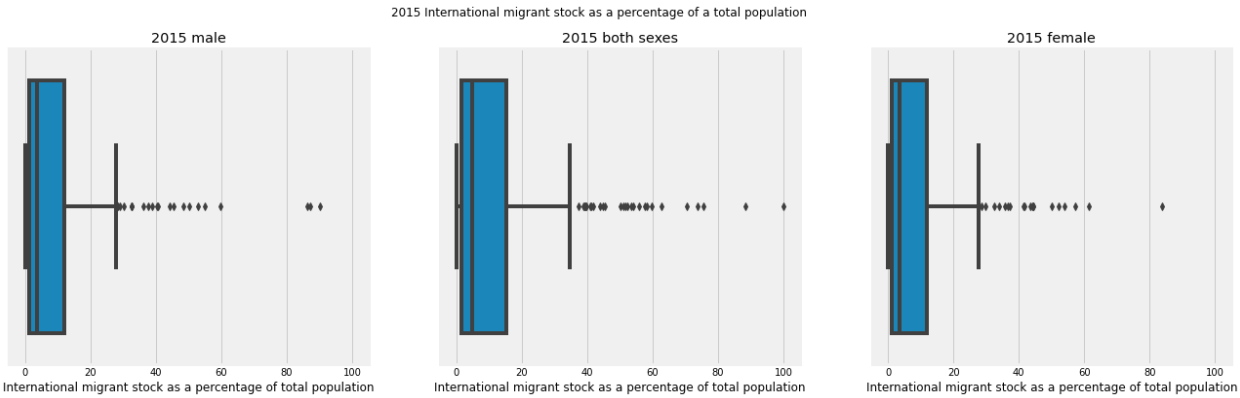
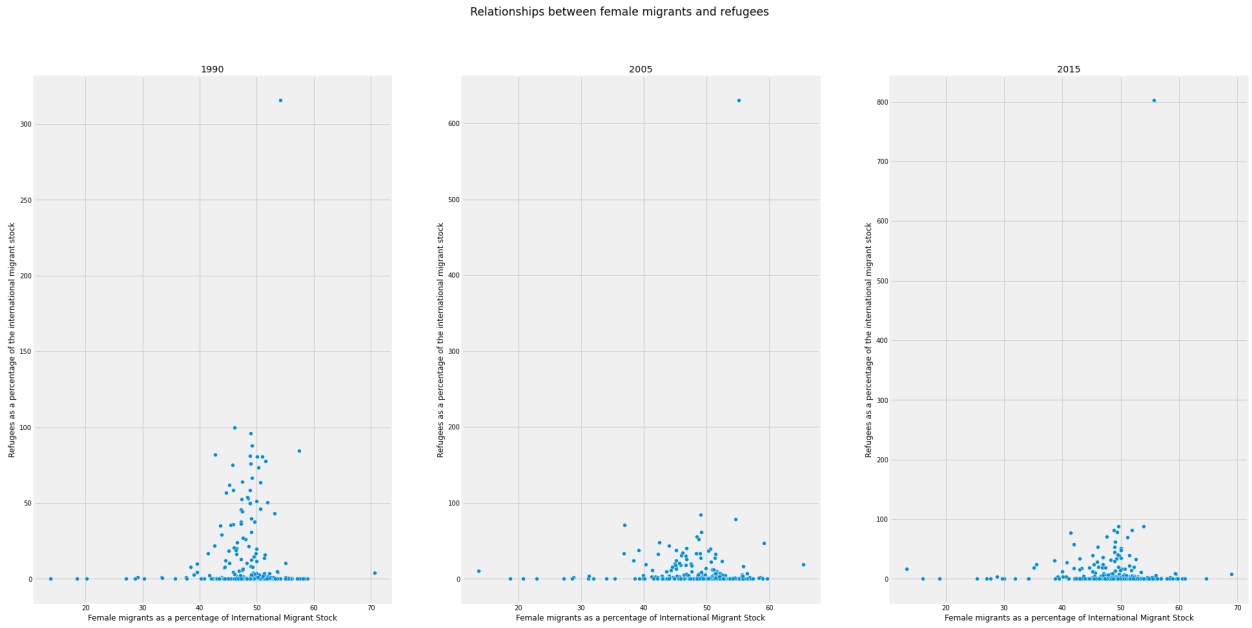


Figure 2: 2015 International migrant stock as a percentage of total population

First, we created the two sets of boxplots for international migrant stock as a percentage of total population. Based on the principle of showing comparison, I combine 3 box plots together to compare 3 different groups in the same year. Also I created another plot for 2015 data to compare with data of 1990.

Based on these two plots, we can easily see that there is a substantial increase in the maximum in 2015 compared to 1990 in all 3 groups. Although the main distribution of this data is below 40, I still keep the scale of 100 in order to show those outliers. Other changes in larger outliers all tell us that there is an increase in the international migrant stock as a percentage of total population. On the other hand, 3 groups' minimum always stayed around 0. This discovery can tell us that there are some countries that have had extremely low migrant stock since 1990. Also, compared to graphs of male and female, it is obvious that graphs of both sexes have more outliers and wider interquartile range indicating that the central portion of both sexes' data spread out further.

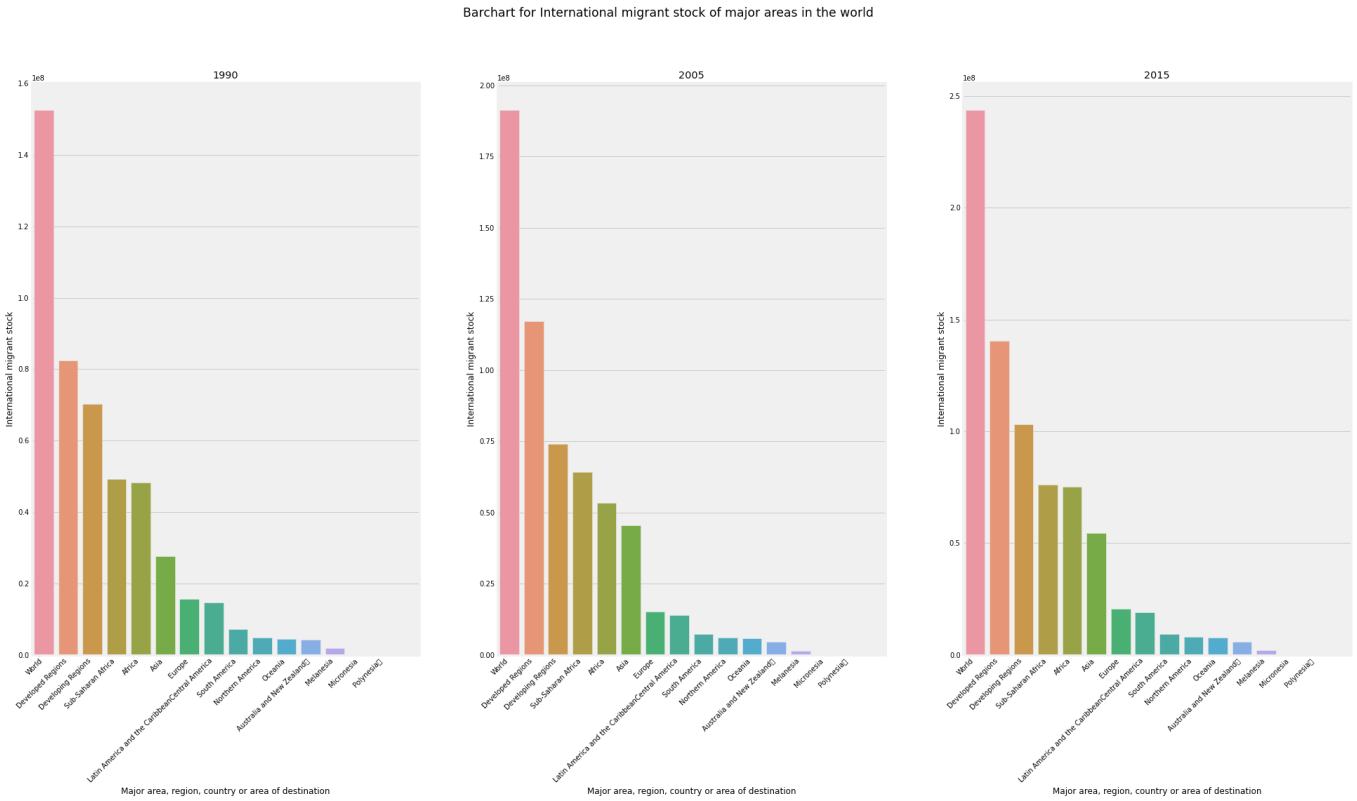


Next, I created 3 scatter plots to explore if there is any relationship between female migrants percentage and refugees percentage, and to compare if this relationship would change with time based on the visualization principle of showing comparison. The x-axis represents the female migrants as a percentage of International Migrant Stock while the y-axis represents the refugees as a percentage of International Migrant Stock. All the dots in the graph represent a single area or country. From the above three graphs, we can easily observe that most areas had a female migrant percentage around 40-60% in 1990, 2005 and 2015. Similar to above box plots, I kept the original scales in order to show those outliers. The main goal of these plots is to explore the relationship between these two variables, so it is less necessary to see the detailed distribution as long as we can identify the trend.

In the graph of 1990, we can see that refugee's percentage will increase as female migrants' percentage increases. However, this trend is not so obvious in later years, which could mean that female migrants are a large proportion of refugees in the last century, but this proportion is decreasing with time. Although correlation cannot equal causality, there must be

some reasons that cause the trend to decrease. We can conduct more research and advanced models to explore this.

Besides the discovery in the above trend, another interesting thing is that I observed that most countries have a refugee percentage of 0 or around 0, which means that refugees as migrants may be a rare phenomenon in most countries. Based on the above discovery, we can explore further if we can build a simple linear regression model between these two data, and we can use statistics like p-value to see if this model can give us the correct prediction.

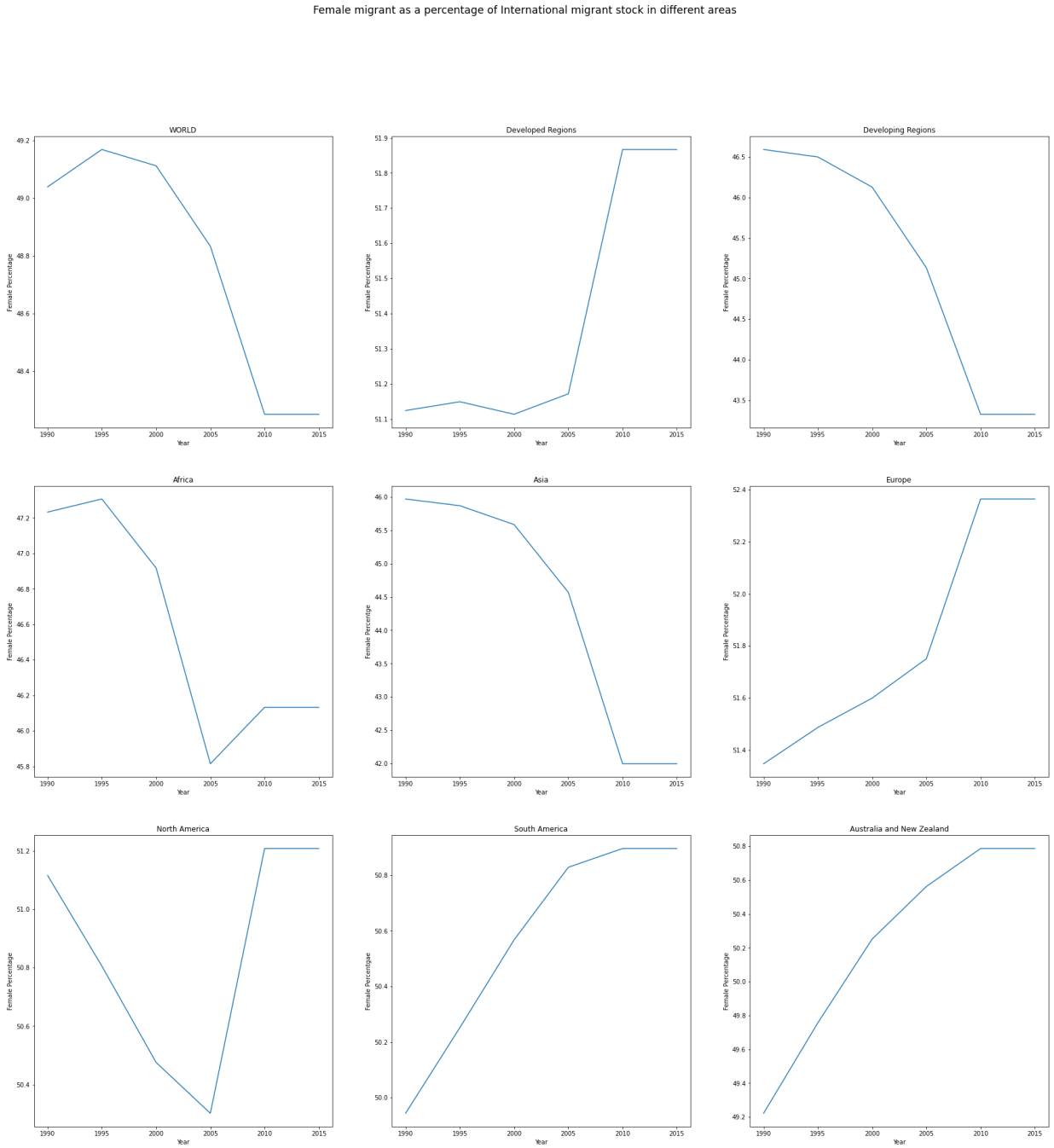


Bar charts can be used to visualize the distribution of data in different categories. When designing these bar charts, I use the set_xticklabels to make it easier to see each label along the x-axis. Also I specified the color of each bar to compare each region based on the principle of showing comparisons and using appropriate encodings. I sorted the data frame in a descending order so that we can see directly which region has a higher international migrant stock.

In the above 3 bar charts, I first observed that the scale of the y-axis increases with year, so it is clear that the international migrant stock in major areas of the world is increasing with time. Also developed regions always have a higher international migrant stock than developing regions. Among the major continents in the world, Africa has the highest international migrant stock, and then is Asia and Europe. Actually this observation is very interesting because it conflicts with the observation that developed regions have a higher international migrant stock than developing regions since there is a large proportion of developing regions in Africa and

Asia. Of course, this may be due to the base number of developed countries, but it is still worth exploring further.

Now if we look at the difference between developed regions and developing regions, we can see that the difference was not so large in 1990, but the difference came to the peak in 2005 and back to normal in 2015. One possible explanation could be some factors that prevented it from increasing during 2005 and 2015. Another possible explanation could be something that stimulated the international migrant stock to increase during 1990 and 2005. Based on these hypotheses, we can further collect related data and conduct experiments.

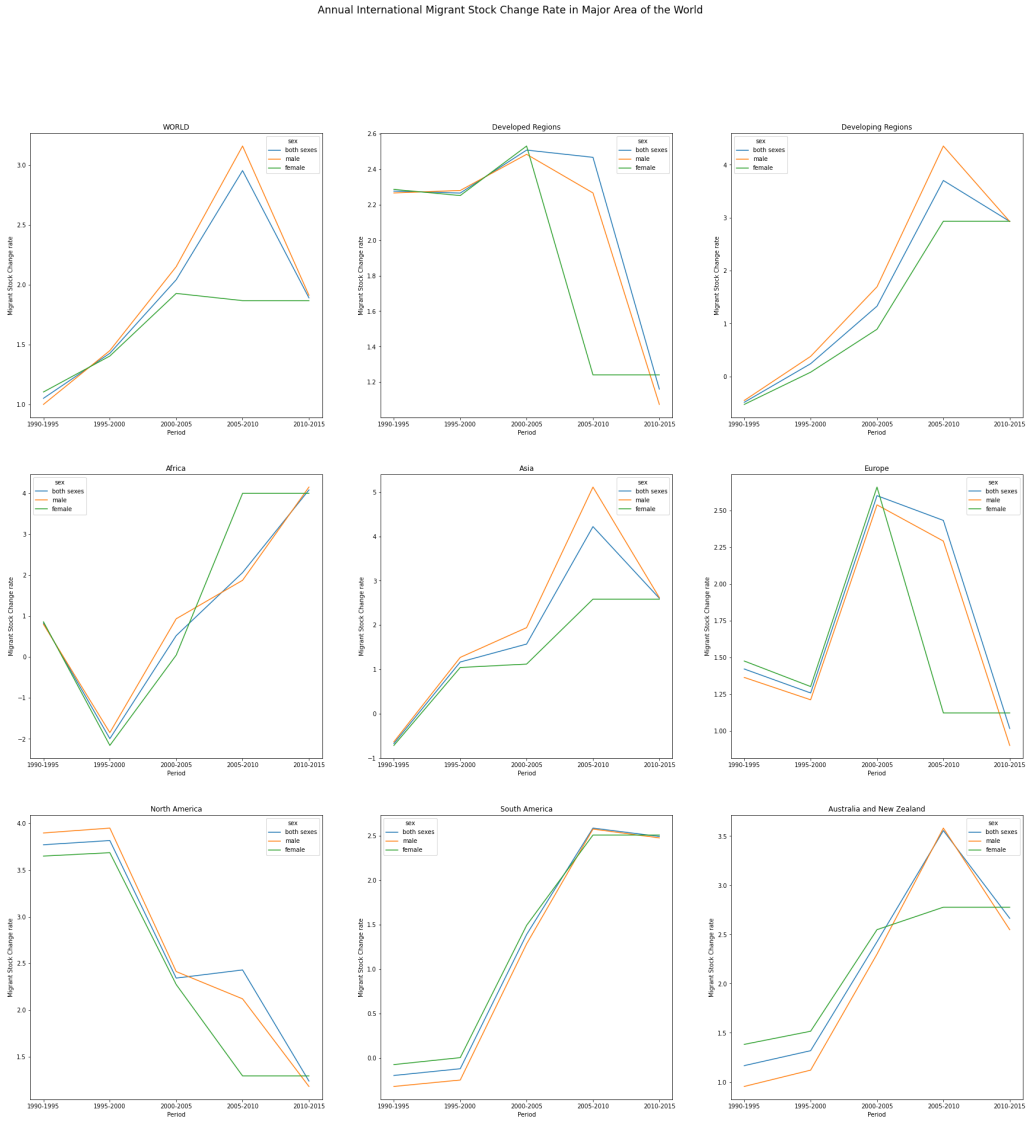


Having a line chart can further help us understand the trend of the data in the timeline. In this set of line charts, I created a new temporary dataframe storing the female migrant percentage

in the main continents of the world. And then I created 9 line charts of these continents and each line chart comes with their own x-label, y-label, sub-title, and their own y-axis scale.

From the above chart, we can easily see that the trend of the overall world and developing region is decreasing while developed regions have an increasing trend. One thing that needs to be noticed is that developed regions have had a sharp increase since 2005. Meanwhile, Europe and North America also have had the same sharp increase since 2005. The reason for this phenomenon may be that most countries in these two areas are considered developed. On the other hand, the trend of developing regions is very similar to the trend of the overall world. This may tell us that the situation of developing regions can better reflect the whole world.

Another interesting point in this graph is that we can see that the female migrant percentage has become stable since 2010. In the period of 2010 to 2015, there was no obvious increase and decrease in all regions. This must be caused by some factors. We can further explore this in the future.

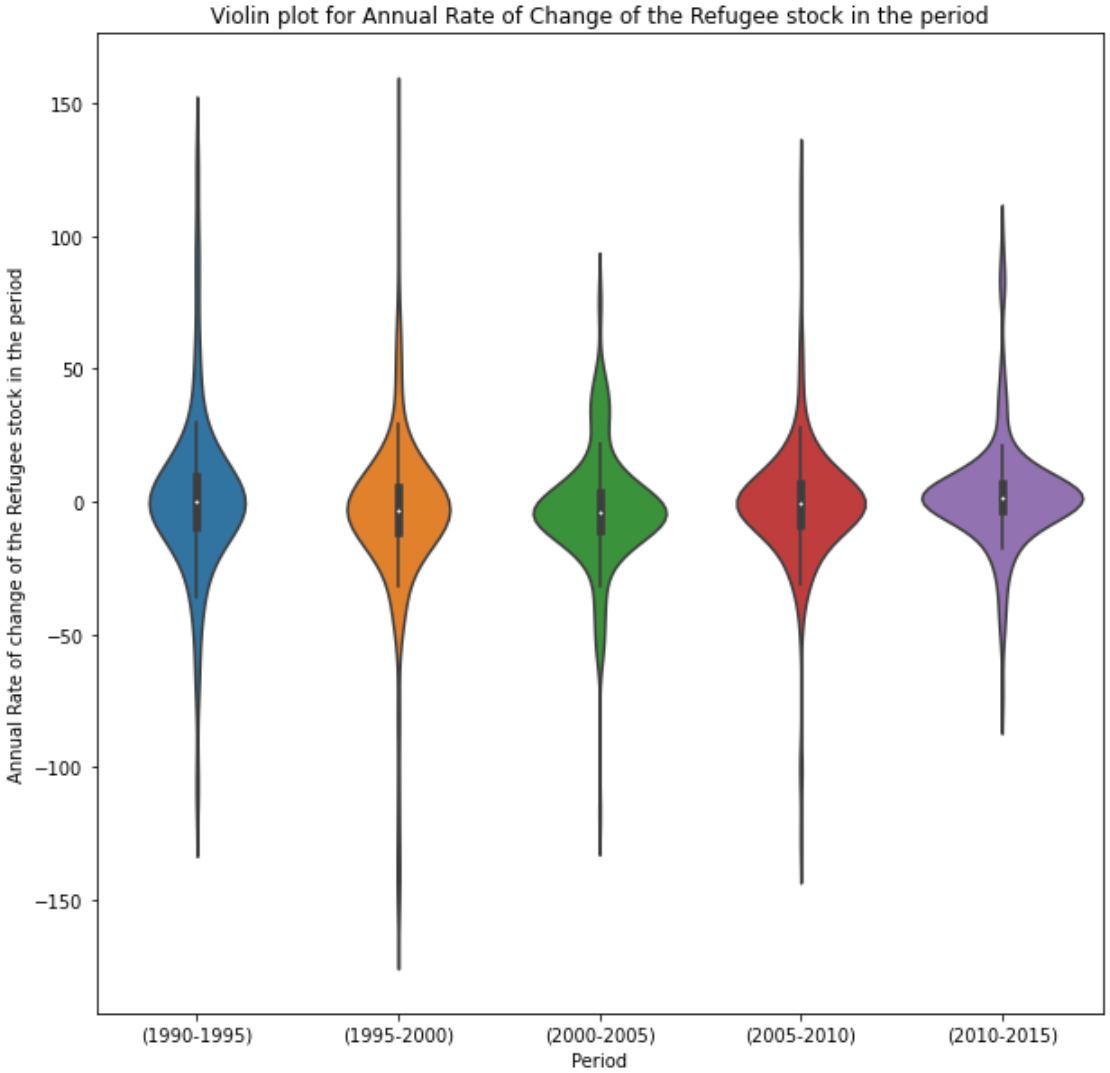


Now we conducted the line chart of international migrant stock in these continents. We further separate the data into 3 groups by sex, which is female, male, and both sexes. In the graph, we used the hue parameter of the Python lineplot function to show 3 groups in 3 different

colors for comparison. Based on the principle related to data-ink ratio, I only show the legend so that we can see the color of each group. It is also clear that each region will have a different level of female migrant percentage, so each subplot has their own y-axis scale based on the principle of using appropriate scales.

What surprised me is that the trend of international migrant stock is completely different from the trend of female migrants percentage. More properly said, it is in contrast to the change of female migrant percentage. Combine these two sets of line charts, we can observe that the world international migrant change is decreasing with time while the female percentage keeps increasing.

However, we still have some exceptions like South America and Australia and New Zealand. Both female percentage and international migrant stock keep increasing in these two areas. There may be some special factors like environment in these two areas, so we need to have more data to find it out.



Similar to the box plot, the violin plot can help us identify summary statistics. Moreover, it can tell us the density of the data. Above is the violin plot for the annual rate of change of the

refugee stock in 5 different periods. Same as previous graphs, colors here help us better compare data in different periods.

Each side's shape of the violin plot represents the density. From the above plot, we can see that density around the mean value increases with year, and the increase is relatively stable according to the shape, which means the annual rate of change is approaching to the mean level. Mean level is around 0, so we may conclude that the annual rate of change of the refugee stock is very tiny in each period. Violin plot is also a useful tool for showing non-normal distribution. From the above 5 violin plots, we can see that most of them are very close to the normal distribution except for the period between 2000 and 2005.

Discussion

From the analysis above, we have got a lot of useful information to build our future models and provide a basis for additional data collection besides bringing us some unexpected discoveries. For example, both violin plots and line chats gave us information that some factors have caused the change of line chart and the non-normal distribution of the violin plot. These potential factors may include some social and environmental factors that are not included in our database. Based on this hypothesis, we can further collect data in some specific directions because we now know that something special may happen in 2005 since these non-normal data are all in 2005. Knowing the specific year can help us collect data more efficiently. Of course, we all know that sometimes data is not up to date since a data collector would collect data a little bit later which means something in 2004 can also influence the data of 2005.

Based on the data visualization, I have already found some relationships between different variables such as female refugees and migrants. The scatter plot of these two variables help me form a hypothesis that there may be a linear relationship between them, so the next step that I will take is to conduct the linear regression model. Splitting the train and test dataset can help me examine if I can use this model to predict some values. Also according to the scatter plot, their relationship seems to reduce with time, and this phenomenon is also observed in 2005 which emphasize the previou discoveries

Besides looking at the whole set of data visualization, we also need to keep an eye on some particular regions such as South America, Australia, and New Zealand. In the two line charts we got, the lines of these three regions keep increasing unlike other regions that have a sharp change in some particular time. We can further explore what caused the increasing international migrant stock of these regions. This can help the country make its immigration policy.

While the EDA inspired us a lot, it also brought us some problems that need to be noticed. For example, we can see that developed regions will have a higher total international migrant stock than developing regions. However, if we look at some continents specifically, we will see that continents like Asia and Africa have a higher international migrant stock, but there are more developing countries in these continents. We need to find out the reason for these conflicted phenomena in the future research.

After using and comparing these different visualization tools, I have realized that some of them have an obvious advantage on specific problems and others may have some disadvantages. For example, the violin plot and the box plot is designed for those audiences who have a statistical background since it requires some statistical knowledge to read it. We may add more texture parts using these two plots. On the other hand, plots like bar charts and line charts are easy to read and understand for everyone, and it can reveal some patterns and distributions.

However, if we have a mass dataset which contains a lot of categories, we have to be careful when we design the bar chart since it will be a messy and unreadable chart.

Conclusion

By doing this EDA progress, we have a lot of unexpected discoveries and several hypotheses. Also EDA gave us the direction for the future data collection. Some data visualization tools did help us reveal some patterns, but some tools may have some obvious disadvantages which we should consider in a more proper way in the future.

In conclusion, we can use the current data to build some useful models in the later model constructing progress while we have to keep an eye on the data outside of the dataset because it is very possible that some data in our dataset is easily influenced by the outside. After the EDA process, our first task is to find if there are special circumstances around 2005 that makes the data non-normal. This EDA process also gave us information like which region will have a higher international migrant stock.