The Exploratory Data Analysis of Trends Changes of **Female Migrants** in Different Developing Stage Areas during the Year from1990 to 2015.

Name: Danni Yang (Student# 1009535558)

GitHub ID: JosephineYang228

Introduction

This report demonstrates the exploratory data analysis (EDA) of the trends changes of female migrants in different developing stages area from 1990 to 2015. The report used UN_MigrantStockTotal_2015.xlsx as the dataset and made data cleaning and adjustments before doing further statistic data analysis.

There are 6 sub-datasets focus on different aspects separately which include migrant stock, area population, international female migrant stock as a percentage of the total population, female migrants as a percentage of the international migrant stock, annual rate, and estimated refugee stock. Since Table 6 only contains both-sex migrants' data, thus, we remove this sub-dataset as it will not influent our study. We also remove the "male" and "both sex" rows in the Table 1-5 and only keep female data for study.

Methods

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, by using statistical graphics and other data visualization methods. An EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing, which is more intuitionistic and visualized and easy understanding for people with no related domain knowledge.

Edward Rolf Tufte is noted for his writings on information design and as a pioneer in the field of data visualization. The data visualization principles he mentioned which could guide and be applied to our work are show as following:

Principle 1: The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented. Data visualization should show multivariate data to provide a comprehensive picture of the data.

Principle 2: Maximize data-ink ratio to show data variation instead of design variation. Data visualization should remove the non-data elements to mess the graphics.

Principle 3: Data visualization should apply appropriate encodings such as color, shape, angle, to represent different data values.

Principle 4: Clear, detailed and through labeling should used to defeat graphical distortion and ambiguity. Write our explanations of the data on the graph itself. Label important events in the data.

Principle 5: Use appropriate scales to display different data points could make it easier to find significant data nodes and changing trends. The number of information carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

Principle 6: Graphics must not quote data out of context. Data visualization should show causality that how different variables are related and how independent variable changes effected other variables, etc.

This report doing data visualization as per all these Tufte's principles mentioned above.

Results

All the 5 tables are applied to bar chart, line chart and box plot to make date visualization. The outputs are shown as following:

Table 1 – bar chart:

It is clear to see that the overall female international migrant stock is increasing from 74,815,702 to 117,584,801 during the year of 1990 to 2015. Female international migrants in developed countries occupied more than the half amount of all female migrants' stock, while the least developed countries have very small amount of stock around only 5 billion in the past 25 years.

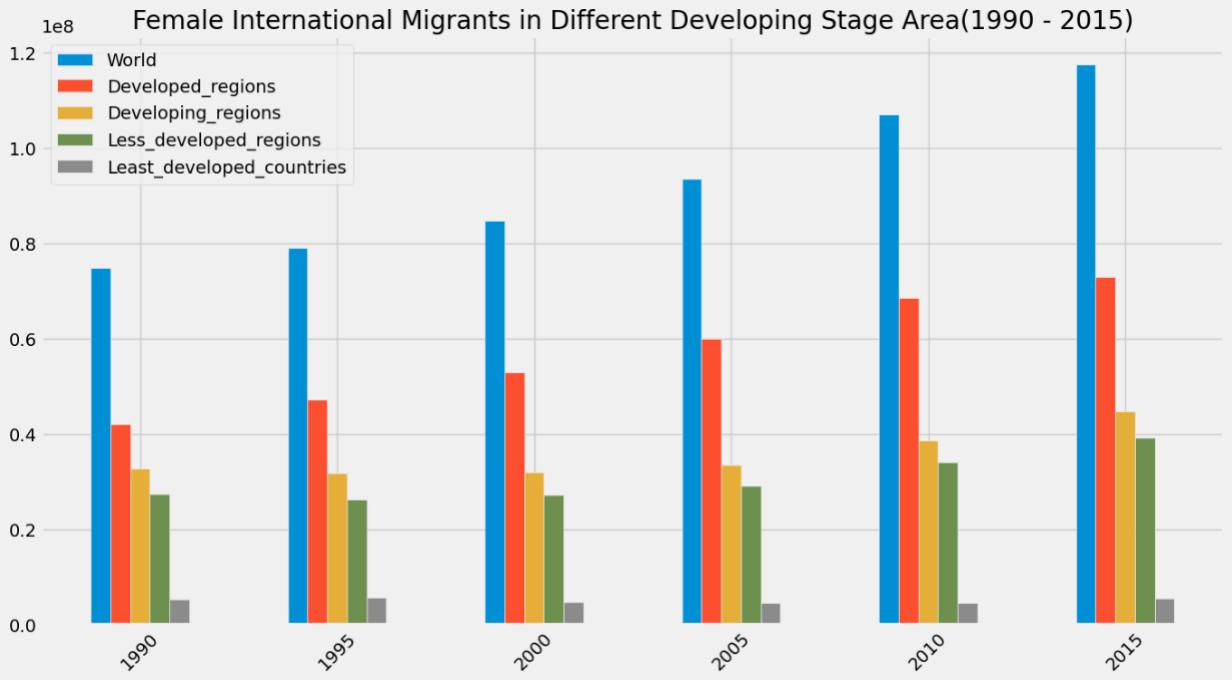


Table 1 – line chart:

The line chart of Table 1 shows that the overall female international migrant stock is increasing which is highly depending on the increase of female migrants' stock from developed regions. Female international migrants in developed countries experienced the most significant increasing from 42,115,231 to 72,863,336, while the other regions all keep the same stock level for 25 years.

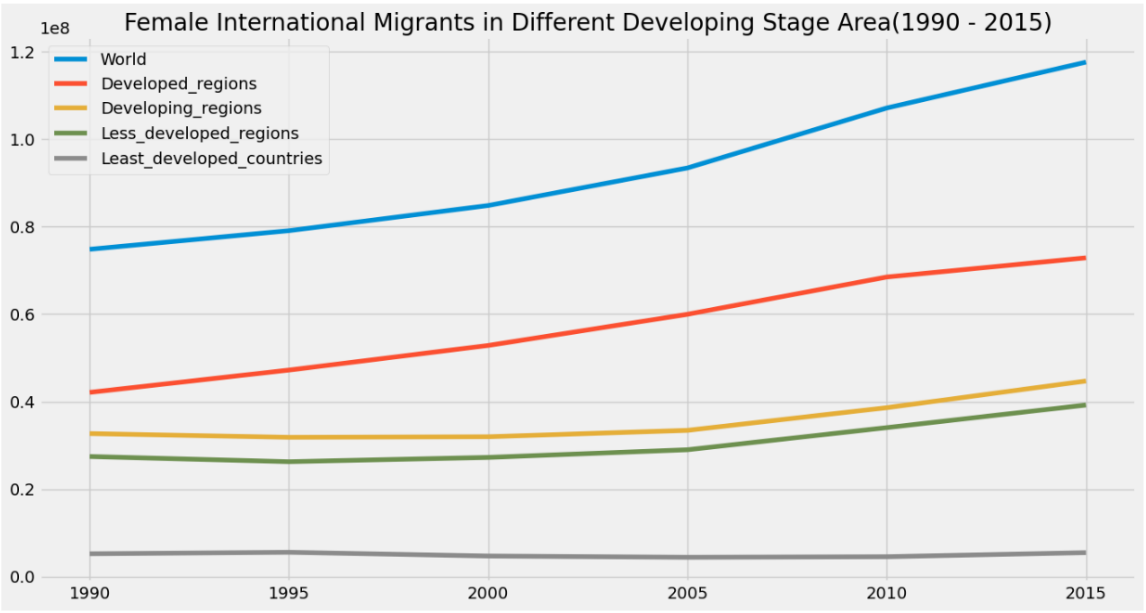


Table 1 – box plot:

The box plot of Table 1 shows that there are no outliers in all independent variables. The median of female international migrants' stock differs greatly among these 5 groups. However, all the median lines of the variables are closer to the minimum amount in the dataset, which illustrates that the higher stock amount is happened only in a few recent years.

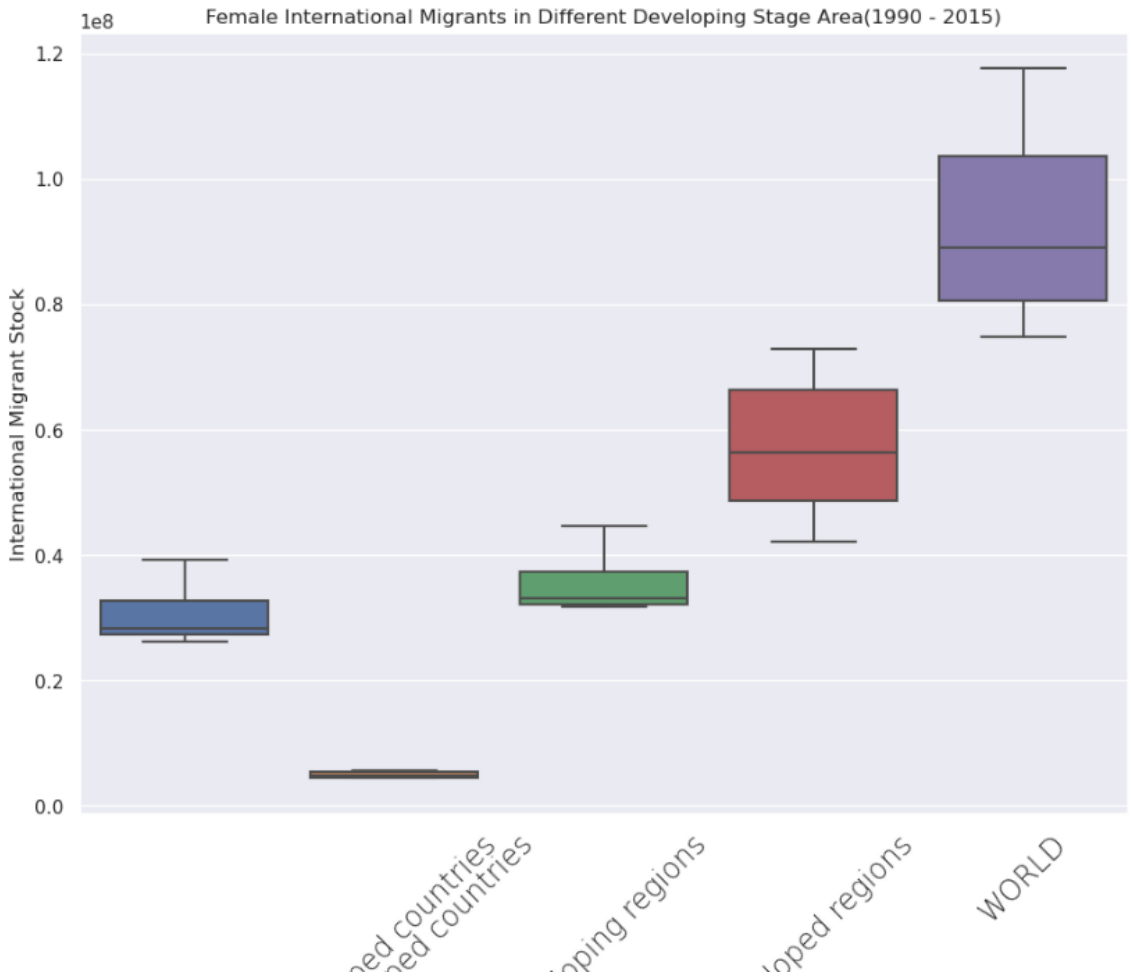


Table 2 – bar chart:

It is clear to see that the overall female population stock is increasing from 2,639,243,998 to 3,642,266,346 during the year of 1990 to 2015. Developed regions has approximately 20% female population stock throughout the whole world. Developing regions and less developed countries has the most of female population stock.

The female population stock all increased during the past 25 years.

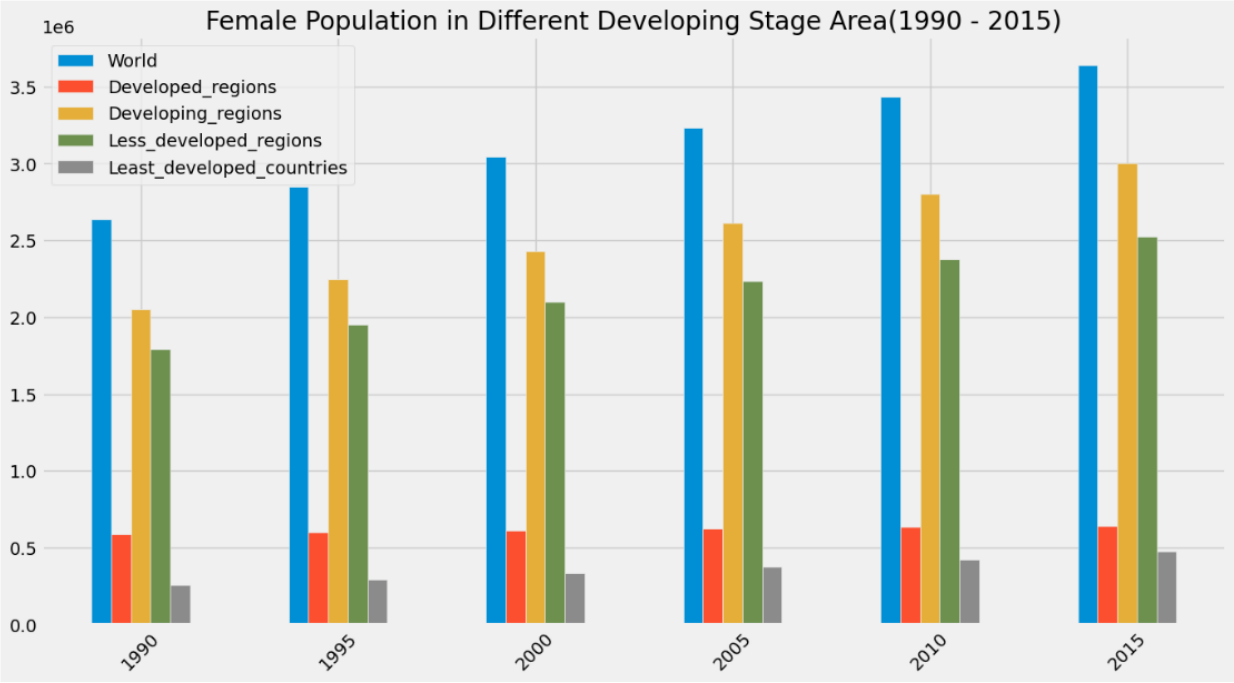


Table 2 – line chart:

The line chart of Table 2 shows that female population stock in all regions were experiencing a constantly raising during the past years while only the developed regions almost keep the same with the amount around six hundred million.

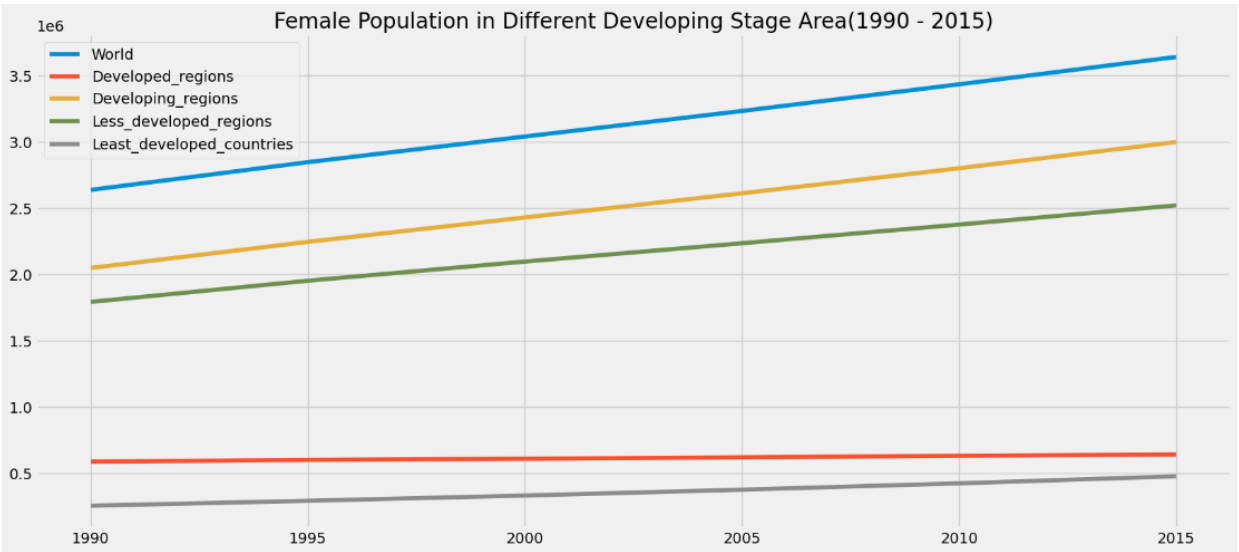


Table 2 – box plot:

The box plot of Table 2 shows that there are no outliers in all independent variables. The median of female international migrants' stock in least developed regions and developed regions, while the developing countries and less developed regions are also sharing a similar median line. However, all the median lines of the variables are close to the mean amount of the dataset, which illustrates that the female population stock is smoothly increased during the 25 years.

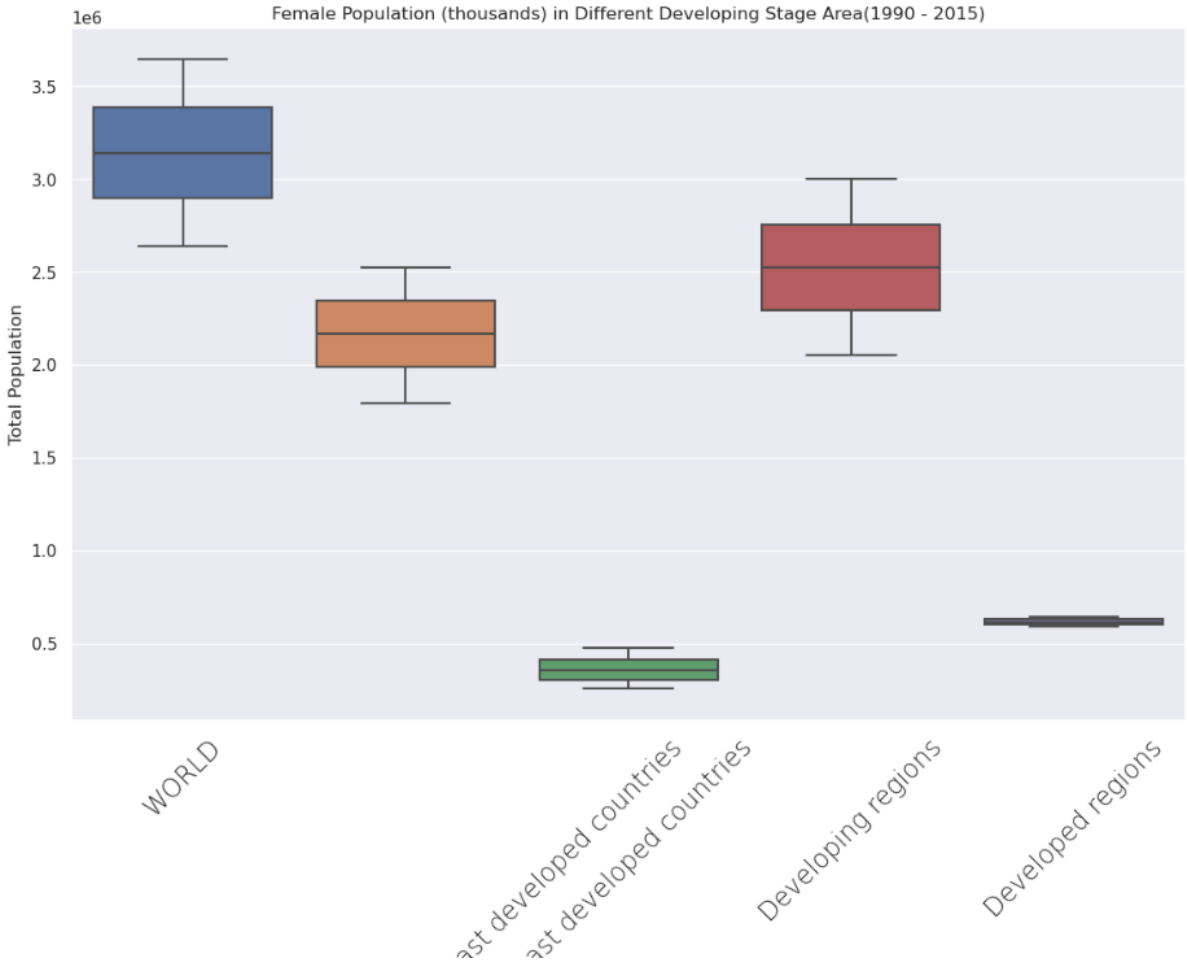


Table 3 – bar chart:

It is clear to see that the overall female international migrant stock as a percentage of the total population kept a stable level. However, the percentage of developed regions significantly differs from other regions. I believe such result is reasonable based on the Table 1 and Table 2 corresponding data.

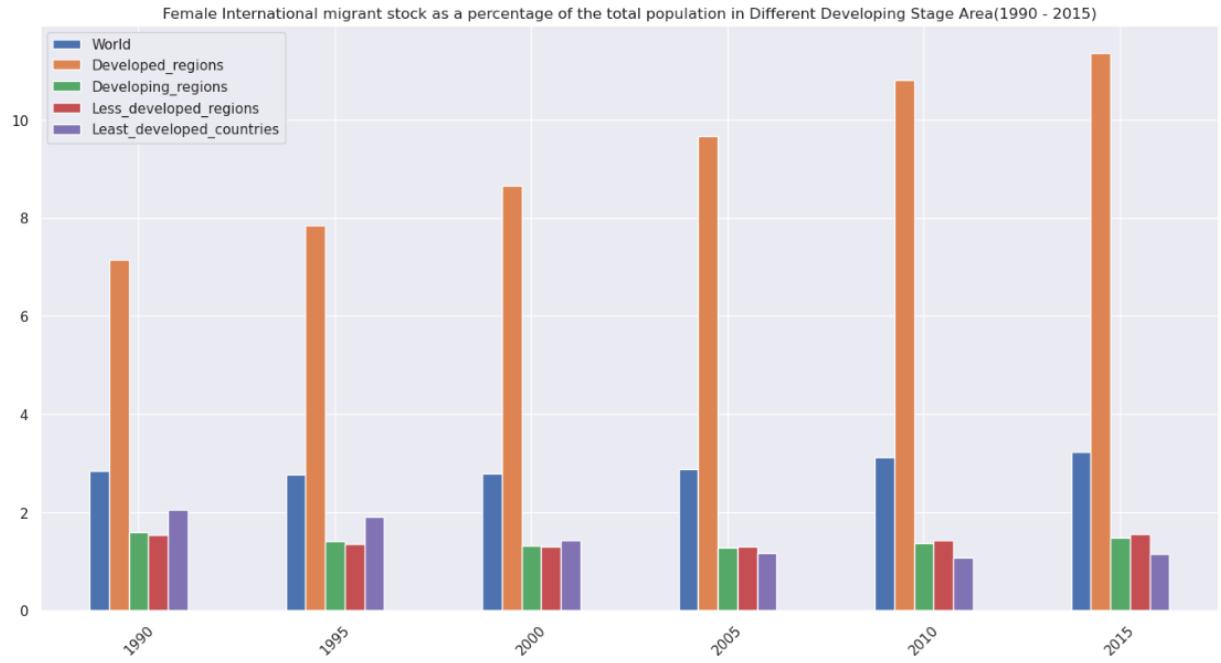


Table 3 – line chart:

The line chart of Table 3 shows that although the female international migrant stock as a percentage of the total population changed greatly in developed countries, however, since the developing regions, less developed regions, and least developed countries almost keep the same percentage level, the changes of percentage for the whole world only experience a gentle rise.

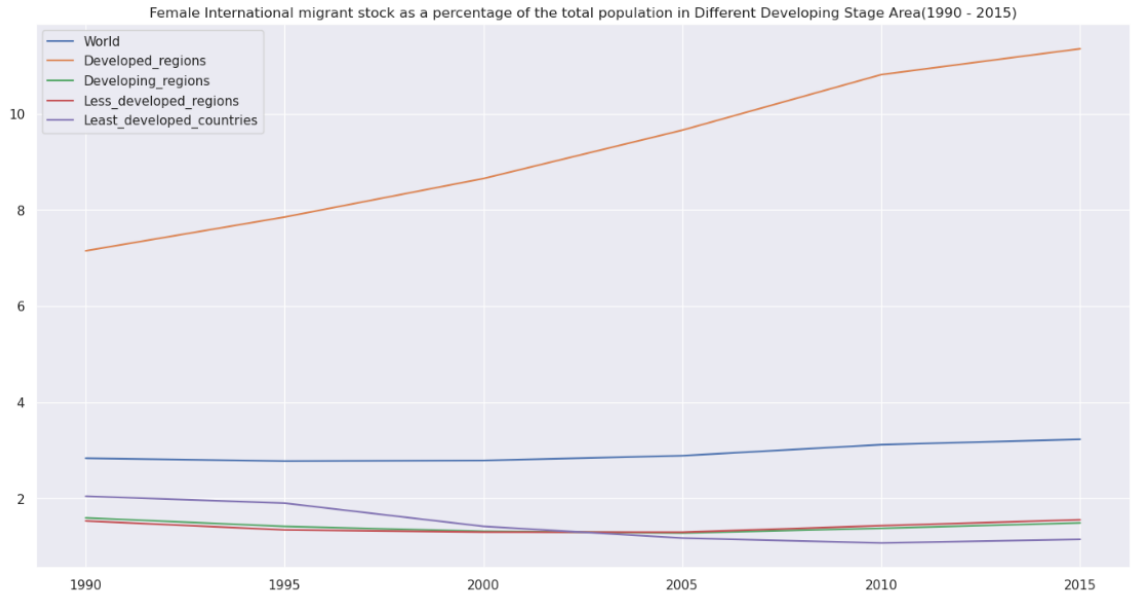


Table 3 – box plot:

The box plot of Table 3 shows that there are no outliers in all independent variables. The median of female international migrant stock as a percentage of the total population almost shares the same median line among the developing regions, less developed regions, and least developed countries. The median of percentage in developed regions is almost 9 times to the rest of other regions.

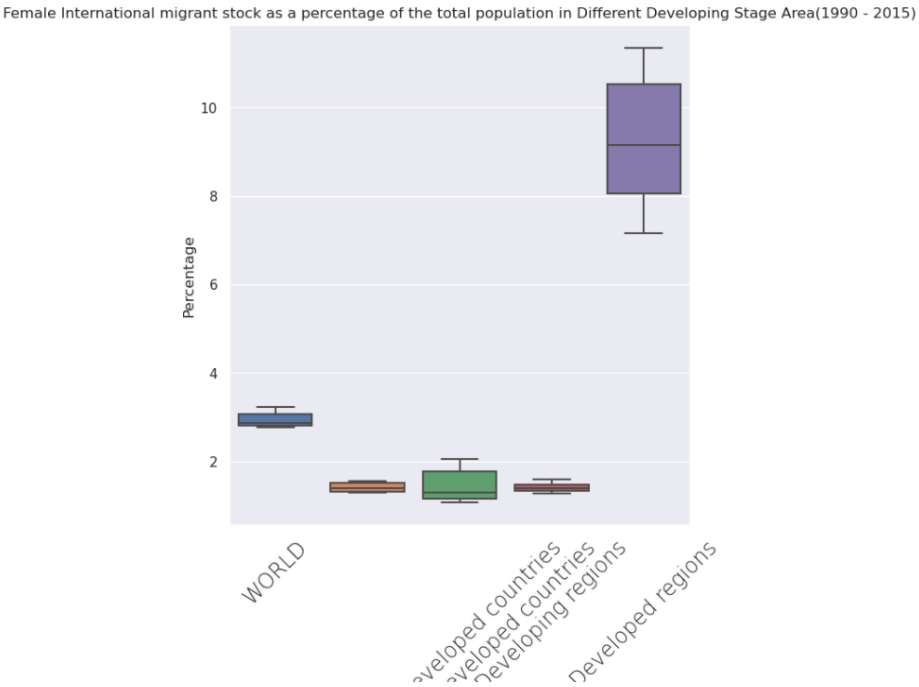


Table 4 – bar chart:

It is clear to see that the overall female migrants as a percentage of the international migrant stock kept a stable level. Such percentage stays near 50% in all regions.

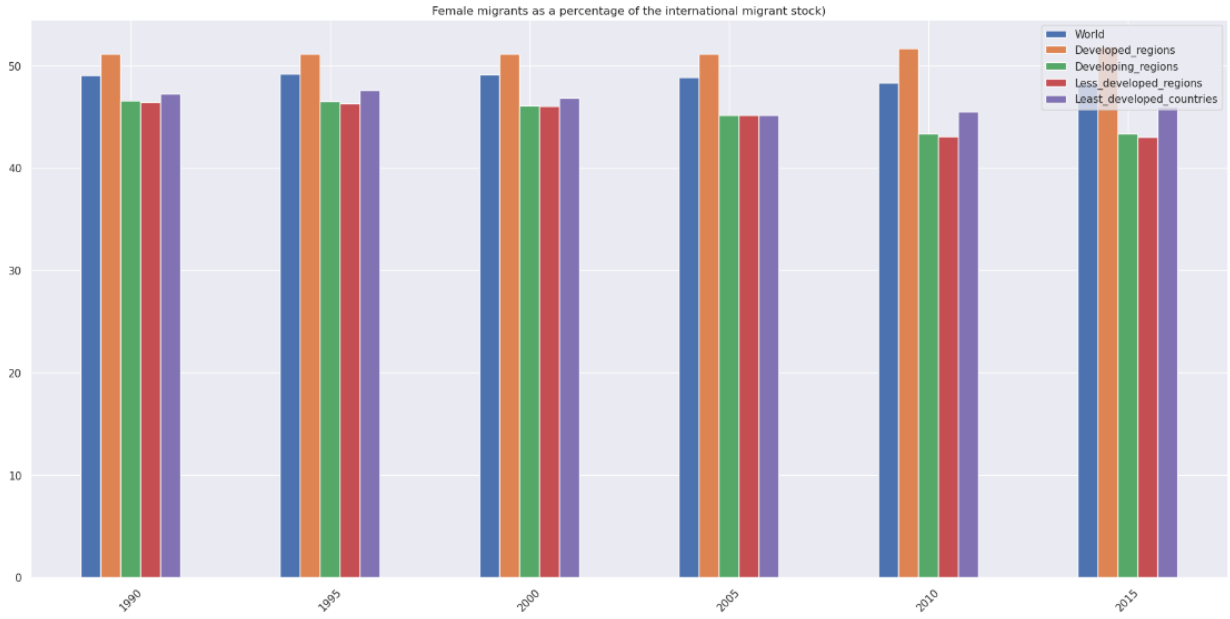


Table 4 – line chart:

The line chart of Table 4 shows that female migrants as a percentage of the international migrant stock are similar among all regions. However, only the percentage of developed regions experienced a slightly increase while the rest of other countries all gently decreased during the past 25 years.

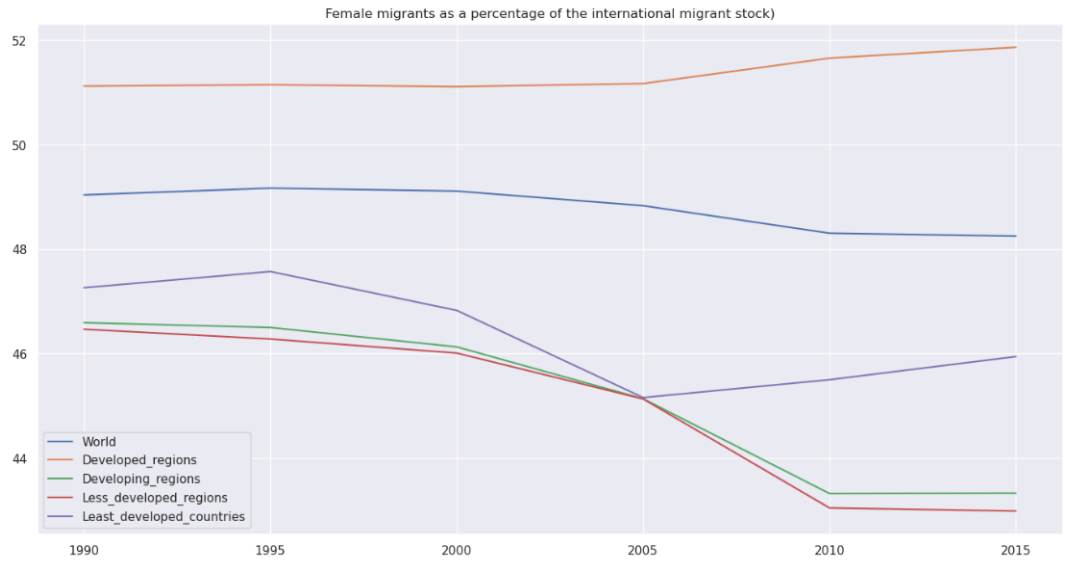


Table 4 – box plot:

The box plot of Table 4 shows that there are no outliers in all independent variables. The median of female migrants as a percentage of the international migrant stock in developing regions, less developed regions, and least developed countries are all near 46%, while the developed regions have a higher median line (51%). Thus the overall percentage of the world neutralize with a number of (49%).

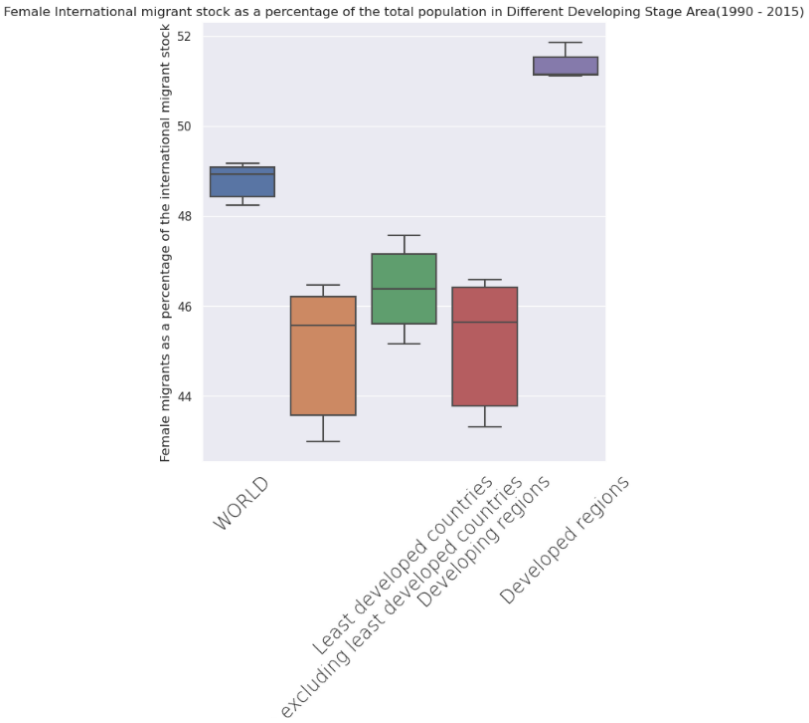


Table 5 – bar chart:

It is clear to see that the annual rate of change of female migrant stock in developing regions, less developed countries and least developed countries has negative growth during the past 25 years. In the year period from 1995 to 2000, the least developed regions experienced a great negative growth for annual rate of changes for female migrants (-3.31%).

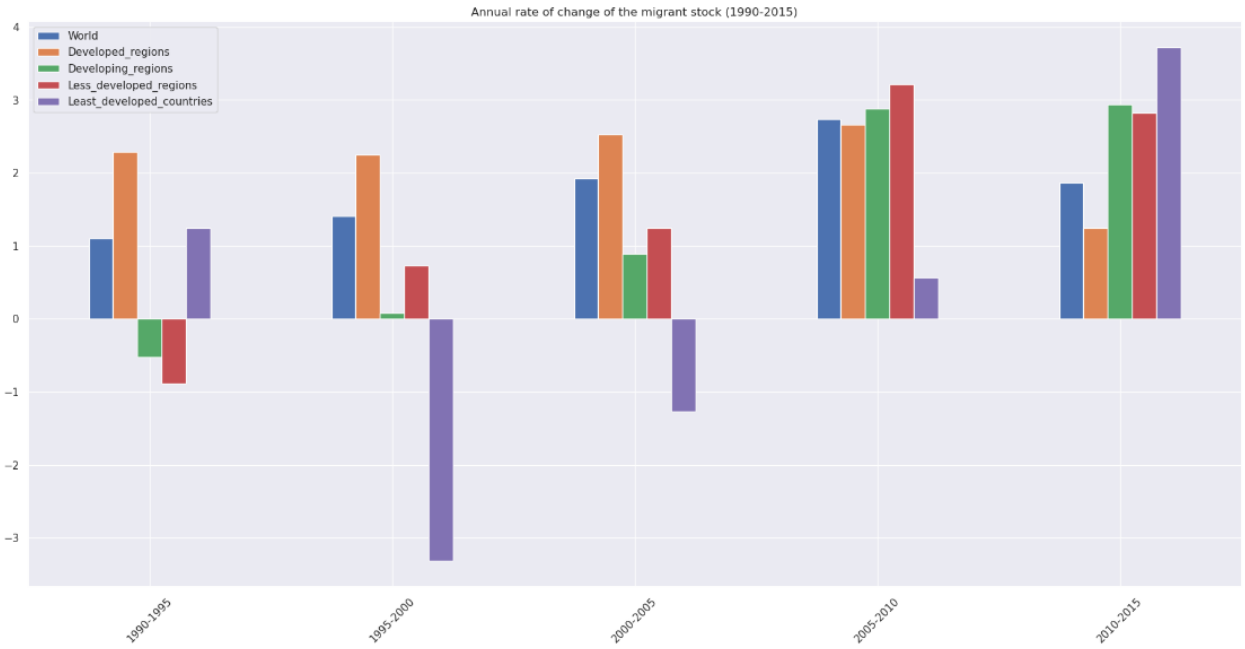


Table 5 – line chart:

The line chart of Table 4 shows that annual rate changes for female migrants experienced a significant increase during the year of 2005 to 2010. However, the year of 2000 could be seen as a watershed that the annual rate in the least developed regions keep rising after the big drop.

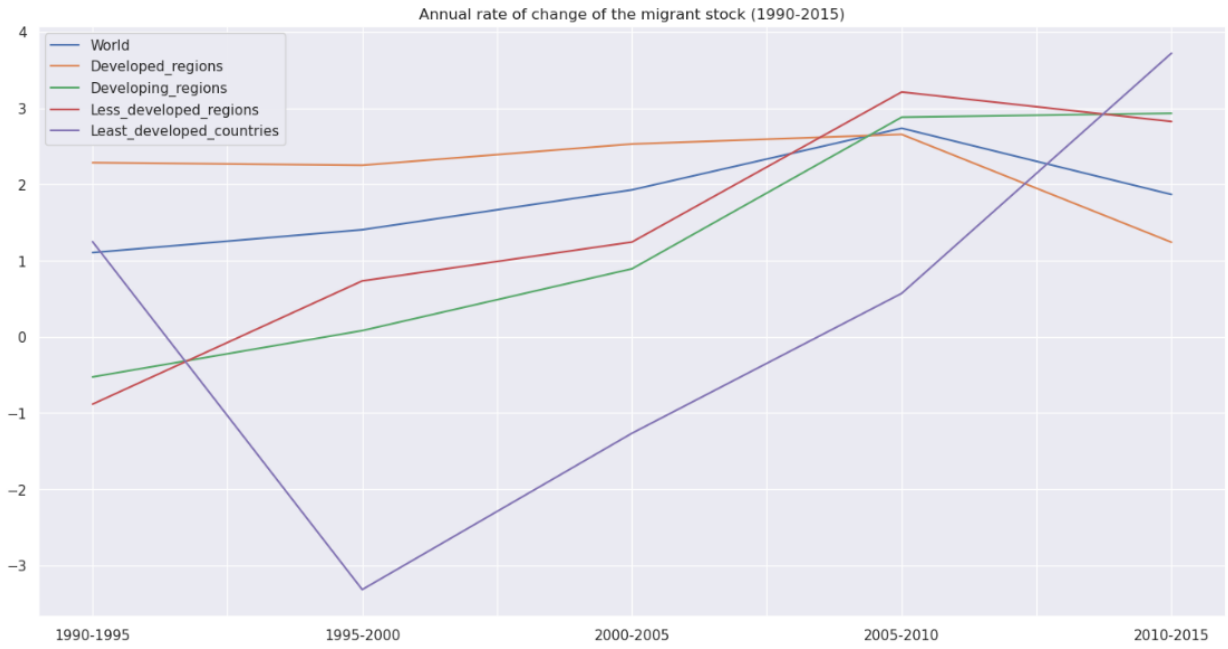
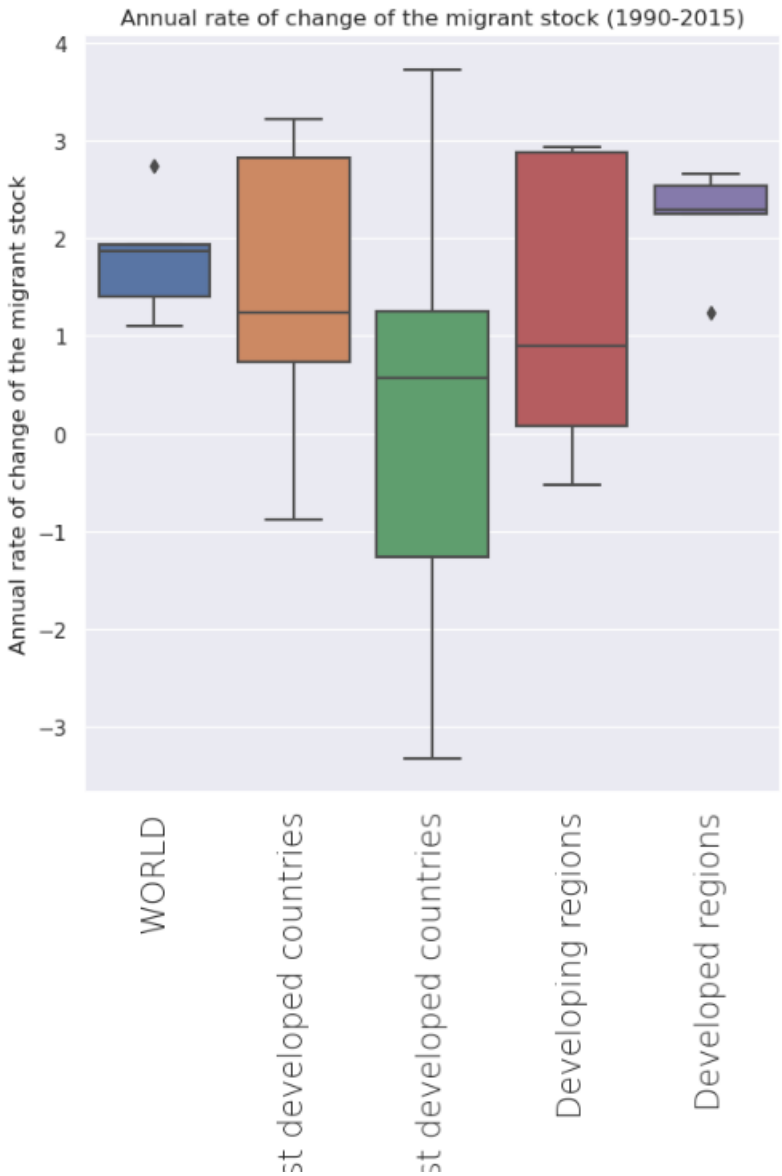


Table 5 – box plot:

The box plot of Table 5 shows that there are some outliers in the independent variables of "World" and "Developed regions". The median of annual rates do not differ greatly among all regions with a similar level (1.5%).



Discussion

The most challenge thing I found in this EDA project is to follow Tufte's principles to display the data in a befitting way. It took me almost 80% of time and effort to understand the principles and applied them in my case. On the contrast, the python methods can be easily found in the lab records and google websites.

I learned a lot from this EDA process that a good figure could save a lot of time and effort for the audience to understand what happened based on the collected dataset. Thus we should pay more attention on thinking about our audience's needs and fit them in our EDA.

Conclusion

To sum up, the female international migrants keep a raising trend during the year of 1990 to 2015. The developed countries and the least developed regions have the most greatly changes while the other regions all kept a stable level in the past 25 years. Developed countries has more than half number of international female migrants while the other regions only have a few amounts. Sine the female percentage are all near 50% in Table 4, we could conclude the gender does not differ in this case. I believe such results could explain that immigration wave as per the process of global integration that people always yearning for a better life in a developed society with better welfare, better benefits.