Inf1340 final writeup
Jiawei Liu 1004803028
2022/12/14

Introduction:

Data visualization could help people to understand key information presented by a complex table. In the midterm project, I reorganized and cleaned the data on international immigrants offered by the United Nations. Now, I try to explore what kind of information this data brings to us. Based on Tufte's visualization principles we learned, I created several plots to mainly compare international immigrant data of developed regions and developing regions.

Methods:

The process of creating plots is similar.

1.Reading the cleaned tables from mid-term file

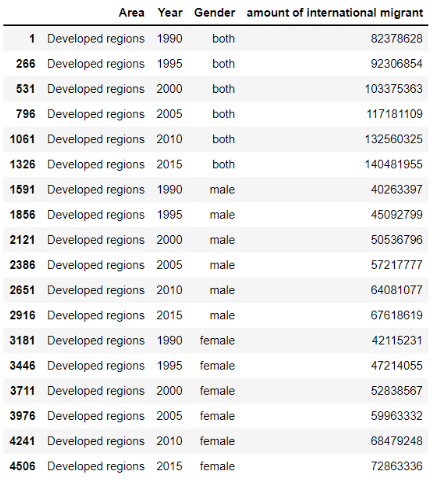
2.Selecting the certain rows and columns which are used for the plots.

3.To create a pivot-table

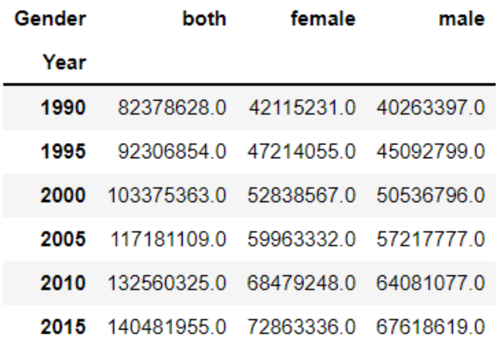
4.Creating the plots based on pivot-table.

I will show you one example for creating the plot "Amount of international migrant in developed regions from 1990-2015" to describe the specific process.

After reading the cleaned table, I extracted all the rows for developed regions and dropped the unnecessary columns for this plot. (See the left table below.)



	Area	Year	Gender	amount of international migrant
1	Developed regions	1990	both	82378628
266	Developed regions	1995	both	92306854
531	Developed regions	2000	both	103375363
796	Developed regions	2005	both	117181109
1061	Developed regions	2010	both	132560325
1326	Developed regions	2015	both	140481955
1591	Developed regions	1990	male	40263397
1856	Developed regions	1995	male	45092799
2121	Developed regions	2000	male	50536796
2386	Developed regions	2005	male	57217777
2651	Developed regions	2010	male	64081077
2916	Developed regions	2015	male	67618619
3181	Developed regions	1990	female	42115231
3446	Developed regions	1995	female	47214055
3711	Developed regions	2000	female	52838567
3976	Developed regions	2005	female	59963332
4241	Developed regions	2010	female	68479248
4506	Developed regions	2015	female	72863336



Gender	both	female	male
Year			
1990	82378628.0	42115231.0	40263397.0
1995	92306854.0	47214055.0	45092799.0
2000	103375363.0	52838567.0	50536796.0
2005	117181109.0	59963332.0	57217777.0
2010	132560325.0	68479248.0	64081077.0
2015	140481955.0	72863336.0	67618619.0

Then I created a pivot table by pandas, setting “Year” as index, “Gender” as column, and the value is “amount of international migrant” (See the right table above). Therefore, based on the pivot-table, I can create the plot I want. The plots will be showed on next part.

And these six data visualization principles from Tufte tell me what is a good plot. (ISDI,2020)

1. **Comparisons:** Show data by comparisons (bar charts and the like) to depict contrasts and differences between dependent variables.
2. **Causality:** Demonstrate how one or more independent variables impact or influence dependent variables.
3. **Multivariate:** Various data are combined so an audience can easily interpret an otherwise complex narrative.
4. **Integration:** Incorporate various modes of information (texts, maps, calculations, diagrams, etc.), to show evidence of source data-to-findings.
5. **Documentation:** For credibility, include attribution, detailed titles, and measurements (scales).
6. **Context:** Describe or depict the before and after state. Show trend lines to hint at results in the future.

Results and Discussion

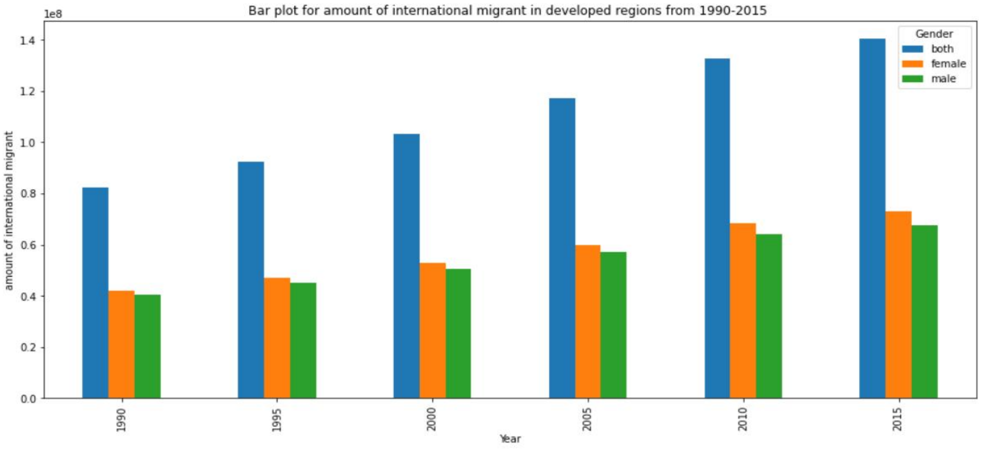


***Figure1**: Bar plot for amount of international migrant in **developed regions** from 1990-2015
Data from United Nations*

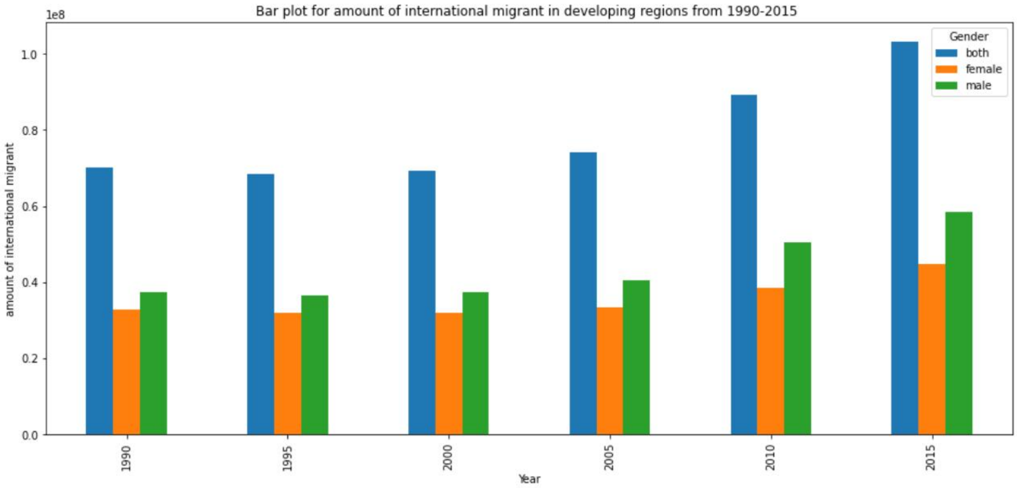


***Figure2**: Bar plot for amount of international migrant in **developing regions** from 1990-2015*
Data from United nations

Comparing Figure 1 and Figure 2, both of the plots show the change in the number of international migrants from the year 1990-2015. No matter whether in developed or developing regions, as time goes on, the amount of international immigrants gradually increases. In contrast, the amount of immigrants in developed regions grows faster than that in developing regions. (80million to 140 million vs. 70million to100 million) We can conclude that developed regions are more popular for immigrants. Furthermore, amount of female migrants in developed regions is slightly larger than male migrants, while in developing regions, the situation is the opposite.

These two figures meet the majority of data visualization principles, I compare the number of immigrants in developed regions with that in developing regions and then demonstrate how the amount change with time goes on. These are the principles of “Comparisons” and “Causality”. I combine the variables “Gender”, “Year”, “Area” and “amount of international migrants”, this is called the principle “Multivariate”. Finally, the plots include appropriate title and scales, which meets the principle called “Documentation”.

I also created a box plot (Figure 3) to check the data distribution for the number of international migrants in developed and developing regions. I wanted to see the interquartile range (IQR) of the data and to check whether there are some outliers.

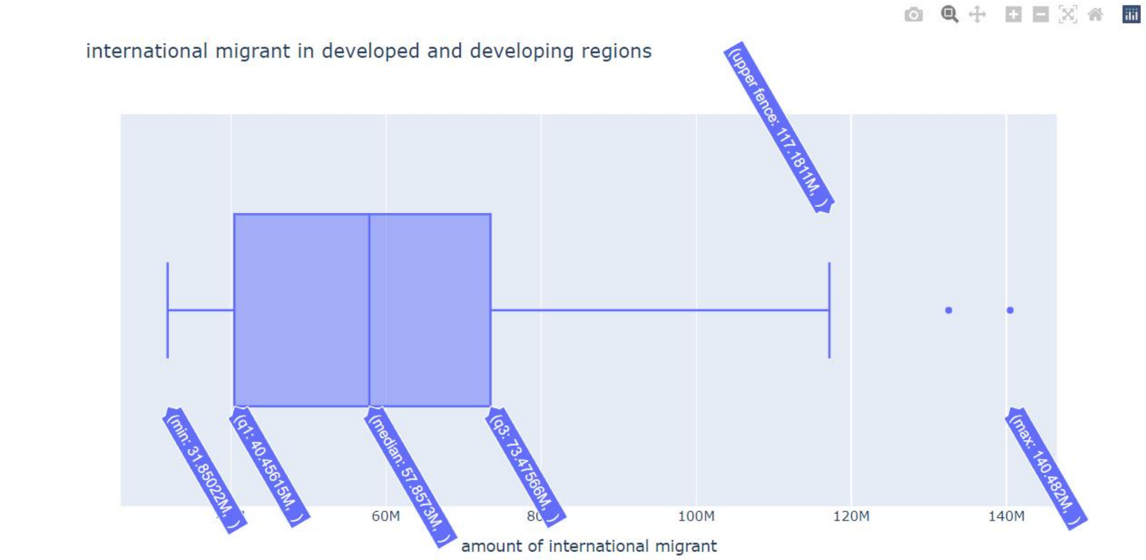


***Figure3**: Box plot for amount of international migrant in developed and developing regions. Data from United Nations*

From the box plot, there are some descriptions of the data set, including min, max, and median. Then, IQR can be calculated although it is not quite important for this data frame. There are two outliers, they are the migrants' amount in 2010 and 2015 in developed regions. We should keep these two outliers since they are parts of useful data.

To integrate a comparison of developed and developing regions, I also created a barplot for the immigrant percentage in the total population of developed and developing regions from 1990-2015. (Figure 4)

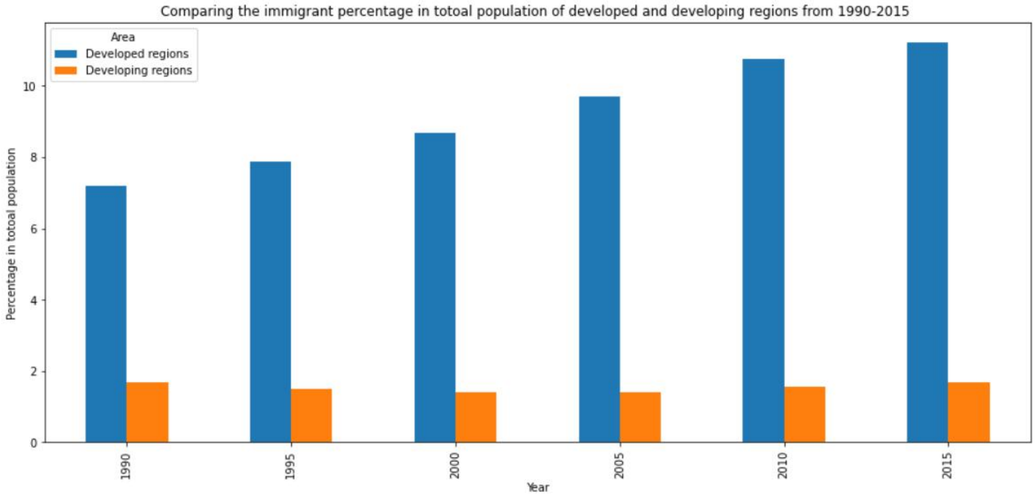


***Figure4**: Barplot for percentage of international migrant in total population in developed and developing regions.*

Now, I directly put both developed and developing regions together in one plot, but only for both genders. From this plot, we can see the increasing trend of immigrant percentage in the total population in developed regions, but for the developed region, it almost keeps the same from 1990-2015. We can conclude that the growth of the number of immigrants is faster than the growth of the population in developed regions, while in developing regions, the growth rate of them almost equal.

In Figure 4, the immigrant percentage in the total population in developed regions is much higher than that in developed regions. I guess the total population in developing regions is large and also growing quickly. In developing regions, the immigrant percentage in the total population has not change from 1990-2015 but the number of immigrants increase a lot. Therefore, Figure 5 is created.

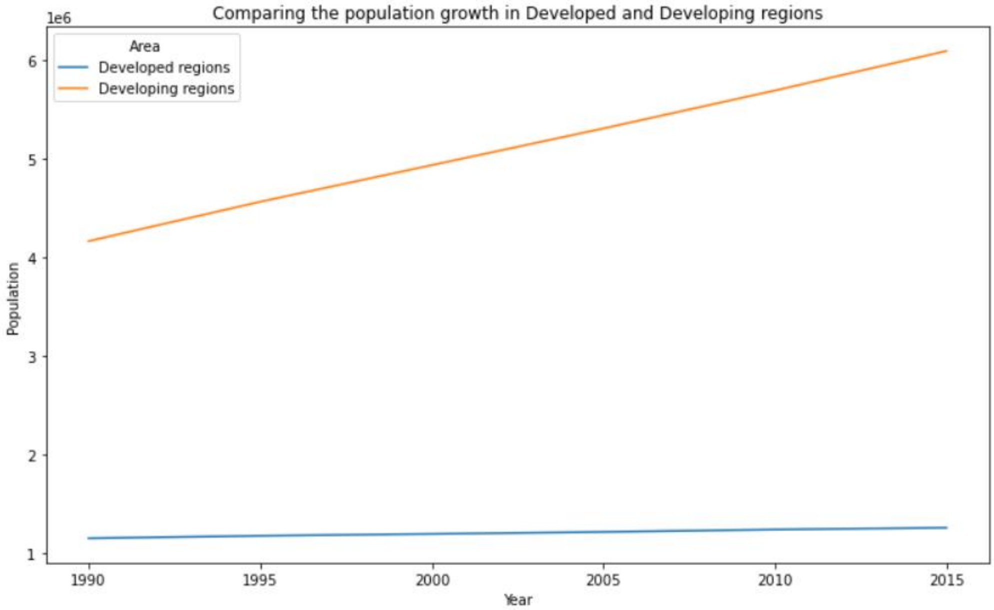


Figure5*: Line plot comparing the population growth in Developed and Developing regions*
Data from United Nations

This line plot proves my hypothesis before. The population of developing regions is much higher than developed regions. Furthermore, it still grows at a high speed. The future trend will still be like the graph, with less increase in developed regions but an increase fast in developing regions.

Now we have 3 kind of plots and describe the trend that data tells to us, the data visualization project meets the last two principle called “Integration” and “Context”. The journey of data visualization is successful but not finish yet.

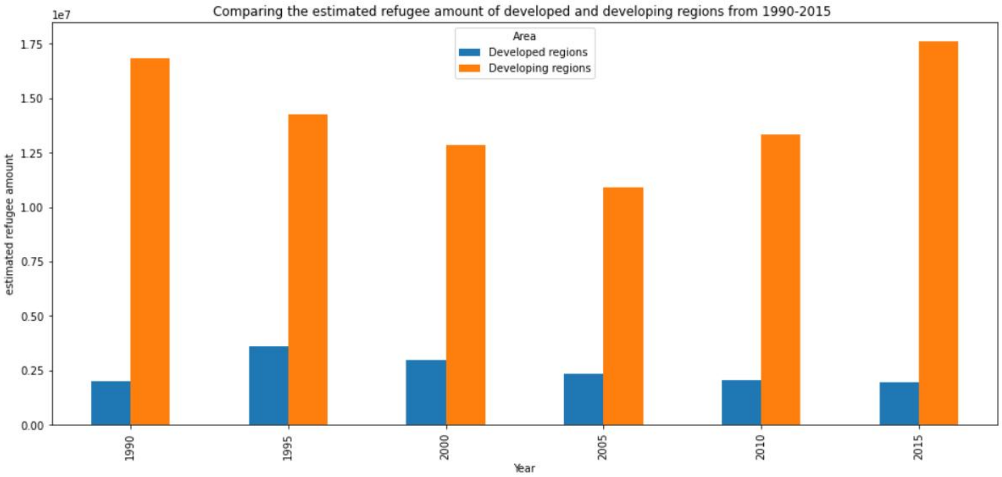


***Figure6**: Bar plot comparing the estimate refugee ion developed and developing regions Data from United Nations*

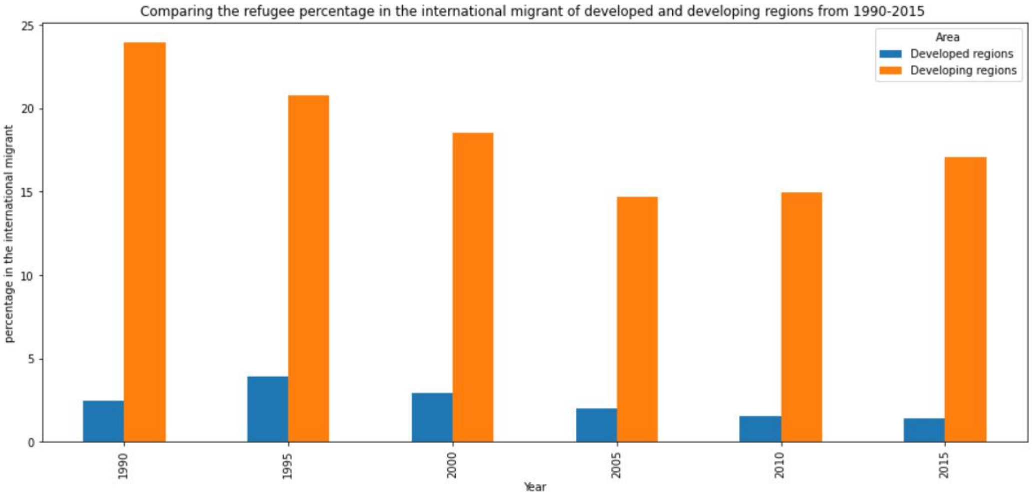


***Figure7**: Bar plot comparing the percentage of refugee in the international migrant in developed and developing regions. Data from United Nations*

Figure 7 tells us that refugee is an important part of migrants in developing regions, the percentage is much higher than that in developed regions. It is easy to understand since in Figure6, the estimate amount of refugee in developing regions is also high. One thing I am confused is the trend of estimate refugees in developing regions. Why the amount firstly decreases from 1990 – 2005 but bounces to increase from 2005-2015? From my perspective, it will gradually decrease since the development of world’s economy.

Therefore, predicting the future trend based on the plot is difficult.

Conclusion:

When I create plots, I realized the importance of tidy data again. A clean data set is the basic thing and then we can visualize. These seven plots could explain a part of information which included in dataset. They still have a lot of limitations. For example, the dataset includes two hundreds of different areas but I only focus on two of them, developed and developing regions. I want to learn more about data visualization, to create kinds of different plots and know how to explain these plots.

Reference:

Admin. (2020, September 3). *Tufte's 6 principles for graphical integrity (adopted for service design)*. International Service Design Institute. Retrieved December 14, 2022, from https://internationalservicedesigninstitute.com/tuftes-6-principles-graphical-integrity-adopted-service-design/