Informal Report of Final Project

Introduction

Data visualization is an excellent tool to help understand and analyze data. According to Edward Tufte, there are six principles good visualization should take into account: comparison, causality, multivariate, integration, documentation, and context. Based on various modes of information like text, maps, calculations, graphs, etc., the visualization shows evidence from source data to findings. The plot should be able to present data by comparison, such as bar graphs, to describe comparisons and differences between dependent variables. It can also demonstrate how one or more independent variables affect the dependent variable. Depending description, the state before and after, the visualization shows trend lines to suggest future results. Combining a variety of data so that an otherwise complex narrative can be easily explained by the listener. Attribution, detailed headings, and measurements (scales) ought to improve credibility. Tufte believes that data scientists should avoid data results, including line plots, being scaled up disproportionately to fit a space. Or, when values are dense in one area and sparse in another, there is a desire and tendency to spread things out evenly. In each case, this can lead to a false impression of the data, and incorrect conclusions. These incorrect manipulations can undermine the credibility of the visualization and mislead the viewer. This final project will be a data visualization practice based on Tufte's six principles, using data from the United Nations' statistical data sets on population, migration, refugees, and gender worldwide.

Method

Preparation (Before creating plots):

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import plotly.express as px
from datascience import *
from tabulate import tabulate
```

```
df1 = pd.read_csv("/content/table1.csv")
df1
```

Import all function packages and datasets, which cleaned at mid-term.

```
df1.shape
```

```
(4176, 8)
```

+ 代码 + 文本

```
group= df1.groupby('continent')
print(group)
group.size()
```

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7f1fa0d9d6d0>
continent
Africa                            1044
Asia                               900
Europe                             864
Latin America and the Carribean    864
North America                       90
Oceania                            414
dtype: int64
```

Before starting to make diagrams, we need to have a general idea about the whole dataset, such as the size. Also, we can group the data frame by Continent, then see the group size, count, and let pandas aggregate data.

```
[ ] group.count()
```

continent	Unnamed: 0	country_or_area_of_destination	major_region	country_code	year	international_migrant_stock_at_mid_year	sex
Africa	1044	1044	1044	1044	1044	1044	1044
Asia	900	900	900	900	900	900	900
Europe	864	864	864	864	864	864	864
Latin America and the Carribean	864	864	864	864	864	864	864
North America	90	90	90	90	90	90	90
Oceania	414	414	414	414	414	414	414

```
[ ] df1['international_migrant_stock_at_mid_year'] = pd.to_numeric(df1['international_migrant_stock_at_mid_year'],errors='coerce')
```

Changing data type is pretty important. When we input dataset into python/Jypter Notebook, the data types will be string automatically. But if you want to create plots or do statistics, they must be float.

```
[ ] group['international_migrant_stock_at_mid_year'].agg(['max', 'min','mean'])
```

continent	max	min	mean
Africa	3142511.0	161.0	1.928769e+05
Asia	10185945.0	3299.0	7.519955e+05
Europe	12005690.0	374.0	8.707542e+05
Latin America and the Carribean	2086302.0	358.0	1.071143e+05
North America	46627102.0	470.0	5.608370e+06
Oceania	6763663.0	63.0	1.756658e+05

```
[ ] df1.groupby(by=['continent', 'sex']).size()
```

```
continent                     sex
Africa                        Both      348
                              Female    348
                              Male      348
Asia                          Both      300
                              Female    300
                              Male      300
Europe                        Both      288
                              Female    288
                              Male      288
Latin America and the Carribean Both    288
                              Female    288
                              Male      288
North America                 Both       30
                              Female     30
                              Male       30
Oceania                       Both      138
                              Female    138
                              Male      138
dtype: int64
```

We can also group data by two columns such as continent and sex, judge whether the sample size is too large.

```
df1.groupby('year')['international_migrant_stock_at_mid_year'].agg('mean')
```

```
year
1990    446091.263158
1995    470180.561404
2000    504980.435673
2005    556824.163028
2010    637109.893678
2015    700288.034483
Name: international_migrant_stock_at_mid_year, dtype: float64
```

```
[ ] df1.head()
```

	Unnamed: 0	country_or_area_of_destination	continent	major_region	country_code	year	international_migrant_stock_at_mid_year	sex
0	0	Afghanistan	Asia	Southern Asia	4	1995	39105.0	Male
1	1	Afghanistan	Asia	Southern Asia	4	2015	193445.0	Male
2	2	Afghanistan	Asia	Southern Asia	4	2000	33069.0	Female
3	3	Afghanistan	Asia	Southern Asia	4	2005	38026.0	Female
4	4	Afghanistan	Asia	Southern Asia	4	2015	188920.0	Female

```
[ ] df1.groupby(by=['continent', 'international_migrant_stock_at_mid_year']).size()
```

```
continent  international_migrant_stock_at_mid_year
Africa     161.0                                   1
           176.0                                   1
           190.0                                   1
           215.0                                   1
           218.0                                   1
                                                  ..
Oceania    4153330.0                               1
           4386250.0                               1
           4878030.0                               1
           5882980.0                               1
           6763663.0                               1
Length: 4112, dtype: int64
```

We have general reviewed the data and done some basic statistics. Now, let's create plots! Remember, we need to do this step before each plot, but I just show once in this report.

Bar plot 1:

```
[ ] df1['year'] = pd.to_numeric(df1['year'],errors='coerce')
    df1['international_migrant_stock_at_mid_year'] = pd.to_numeric(df1['international_migrant_stock_at_mid_year'],errors='coerce')
```

```
[ ] sns.set(rc={'figure.figsize':(18,20)})
    p = sns.barplot(x = 'year', y = 'international_migrant_stock_at_mid_year', hue = 'continent', data=df1)
    p.set(title = 'Six Conitents Migrant Stock Comparison from 1990 to 2015')
    p.set(xlabel='Year', ylabel='International Migrant Stock at Mid Year (Million)')
```

```
[Text(0, 0.5, 'International Migrant Stock at Mid Year (Million)'),
 Text(0.5, 0, 'Year')]
```

Still, changing the data type at first. This time, we need to use 'year' as X-axis so it should be changed from string to float. Then, we use the function from Seaborn package. First, create a canvas or image size, which can be adjusted to achieve the overall beauty and readability of the plot. Here I use 18:20. Then we input the data from df1 and use the 'year' as the x-axis, the 'international migrant stock' as the y-axis, and label the color by continents. Put those data into the sns.barplot function. We can run the code to check the plot, then edit axis and add the title.

Bar plot2:
Review the table 2 data before creating the plot, just as what I do in Preparation part. Other codes are the same as plot1.

```
[ ] df2['year'] = pd.to_numeric(df2['year'],errors='coerce')
    df2['population_as_thousands'] = pd.to_numeric(df2['population_as_thousands'],errors='coerce')
```

```
[ ] sns.set(rc={'figure.figsize':(18,20)})
    p = sns.barplot(data=df2, x="year", y="population_as_thousands", hue="continent")
    p.set(title = 'Six Conitents Population Comparison from 1990 to 2015')
    p.set(xlabel='Year', ylabel='Population (thousands)')
```

Line plot:
For this plot, we just need North America countries, so use df.loc function to select matching rows.

```
df3 = pd.read_csv("/content/table3.csv")
df3 = df3.loc[df3['continent'] == 'North America']
df3
```

Changing the data type. Then, code the sns.lineplot function. Input df3, using 'year' as x-axis and 'international migrant stock as a percentage of total' as y-axis. Label colors by countries. Err_style can change the line type, which is finest I used. Function set_ylim can change the max and min height of y-axis.

```
df3['year'] = pd.to_numeric(df3['year'],errors='coerce')
df3['international_migrant_stock_as_a_percentage_of_total'] = pd.to_numeric(df3['international_migrant_stock_as_a_percentage_of_total'], errors='coerce')
```

```
sns.set(rc={'figure.figsize':(20,14)})
p = sns.lineplot(data=df3, x = "year", y="international_migrant_stock_as_a_percentage_of_total", hue="country_or_area_of_destination", err_style = "bars")
p.set(title = 'North America Migrant Stock of Total Population Changed by year')
p.set(xlabel='Year', ylabel='Migrant Stock as a Percentage of Total Population (%)')
p.set_ylim(0,35)
```

Box plot:
Changing the data type. Using function sns.boxplot, and inputting df5. Using ‘continent’ as x-axis. Using ‘annual rate of change of the migrant stock’ as y-axis. Label the color by year. Now, this time we don’t need to change year’s data type because string can also label colors. We need to be aware that not all data are suitable for making box plots. To ensure that the plot can be easily understood, it is best to use percentage data with negative numbers.

```
[ ] #df5['year'] = pd.to_numeric(df5['year'],errors='coerce')
    df5['annual_rate_of_change_of_the_migrant_stock'] = pd.to_numeric(df5['annual_rate_of_change_of_the_migrant_stock'], errors='coerce')
```

```
[ ] sns.set(rc={'figure.figsize':(20,14)})
    p = sns.boxplot(data=df5, x = "continent", y ="annual_rate_of_change_of_the_migrant_stock", hue="year")
    p.set(title = 'Annual Rate of Change of the Migrant Stock by year for Six Continents')
    p.set(xlabel='Continents', ylabel='Annual Rate of Change of the Migrant Stock (%)')
    p.set_ylim(-40,40)
```

Violin plot:
Just as others, input dataset by pandas function and select rows contain ‘female’ and ‘male’ under ‘sex.’ Then, change the data type.

```
] df5 = df5.loc[df5['sex'].isin(['Female','Male'])]
  df5 = df5.loc[df5['major_region'].isin(['Northern America'])]
  df5
```

We want to see one violin contain two sexes, so label color by ‘sex’. Input df5, let countries as x-axis and ‘annual rate of change of the migrant stock’ as y-axis. Remember label axis and add the title.

```
sns.set(rc={'figure.figsize':(20,14)})
p = sns.violinplot(data=df5, x = "country_or_area_of_destination", y ="annual_rate_of_change_of_the_migrant_stock", hue="sex", split = True)
p.set(title = 'Distribution of Annual Rate of Change of the Migrant Stock in North America')
p.set(xlabel='North American Countries', ylabel='Annual Rate of Change of the Migrant Stock (%)')
#p.set_ylim(-40,40)
```

Bar plot3:
Just as the first and second bar plots.

```
[ ] df6 = pd.read_csv('/content/table6B.csv')
    df6
```

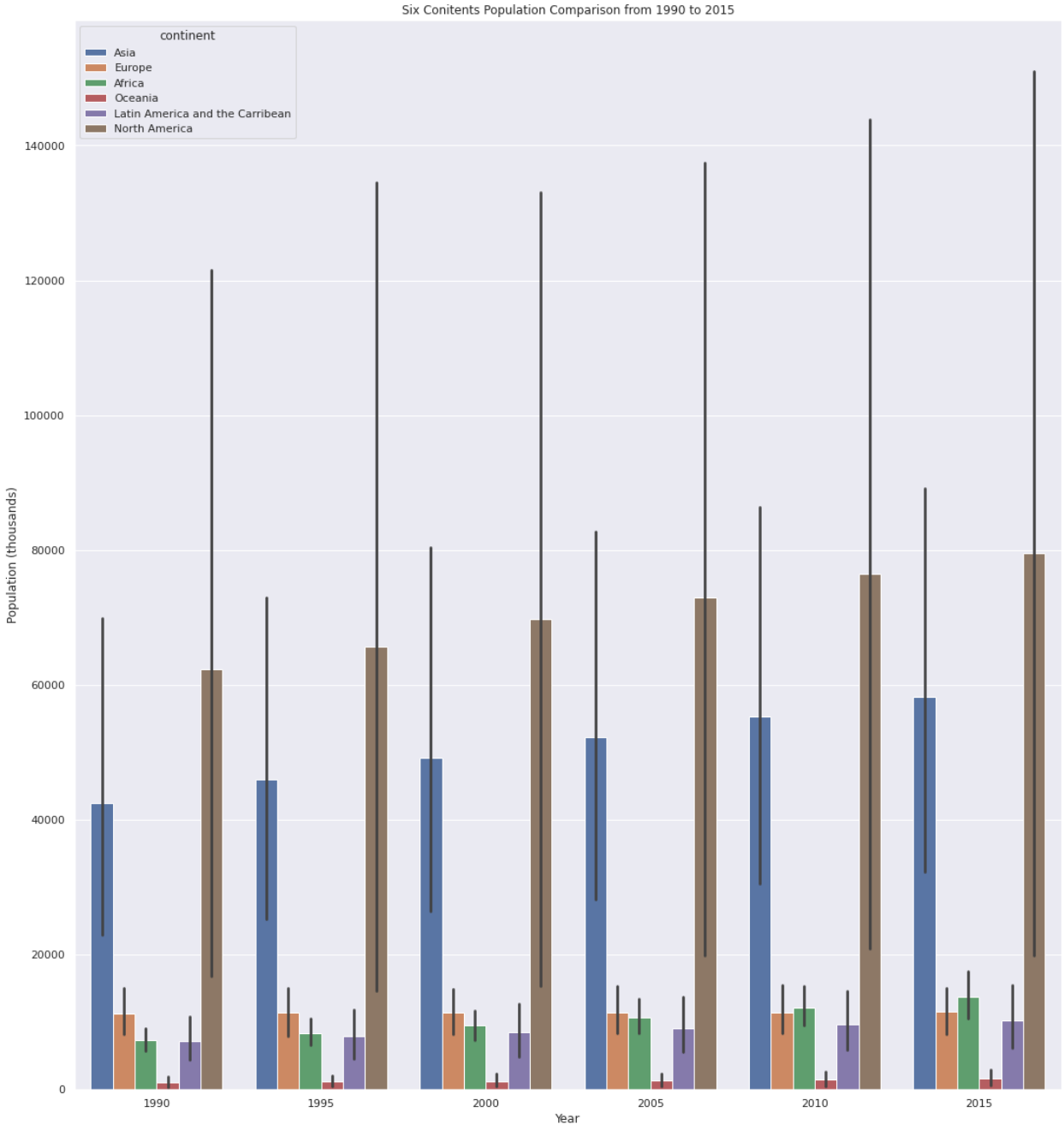
```
[ ] df6 = df6.loc[df6['country_or_area_of_destination'].isin(['Canada','United States of America'])]
    df6
```

```
[ ] df6['refugees_as_a_percentage_of_the_international_migrant_stock'] = pd.to_numeric(df6['refugees_as_a_percentage_of_the_international_migrant_stock'], error
    df6['year'] = pd.to_numeric(df6['year'],errors='coerce')
```

```
sns.set(rc={'figure.figsize':(20,14)})
p = sns.barplot(data=df6, x="year", y="refugees_as_a_percentage_of_the_international_migrant_stock", hue="country_or_area_of_destination")
p.set(title = 'Canada&USA Refugee out of Migrant Stock from 1990 to 2015')
p.set(xlabel='Year', ylabel='Refugees as a percentage of the Migrant Stock (%)')
p.set_ylim(0,4)

(0.0, 4.0)
```

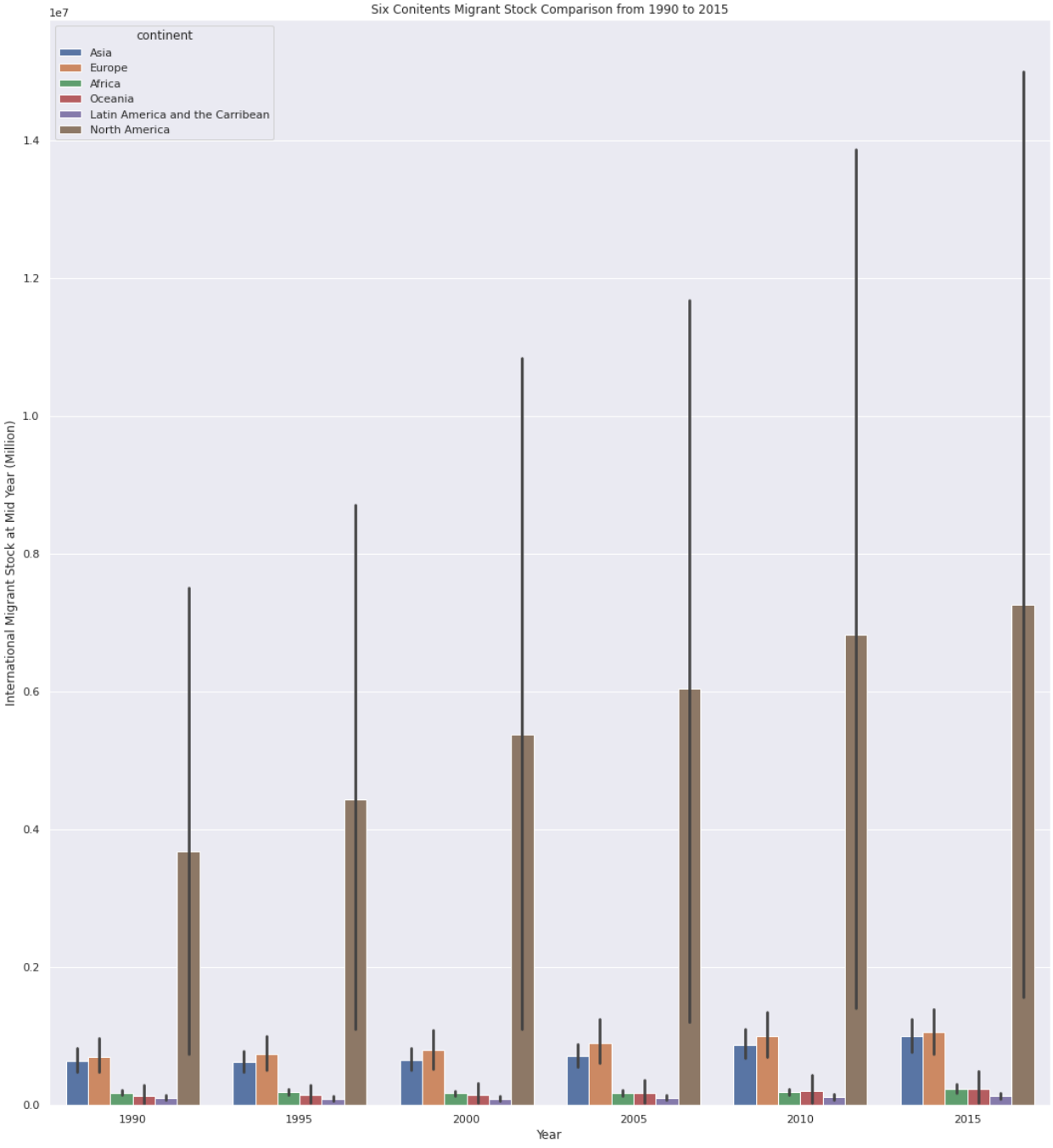
Result
Viz1: Bar plot from table2



The bars use the height of each rectangle to represent a central tendency estimate for a numeric variable and use error bars to provide some indication of the uncertainty surrounding that estimate. The bars include zeros within the quantitative axis, and they work well for comparisons when the zero is a meaningful value for the quantitative variable. The entire dataset is about the UN survey on population, gender, migration and refugees, so this project starts with the overall demographic data (table2) for data analysis and visualization. The first visualization is the

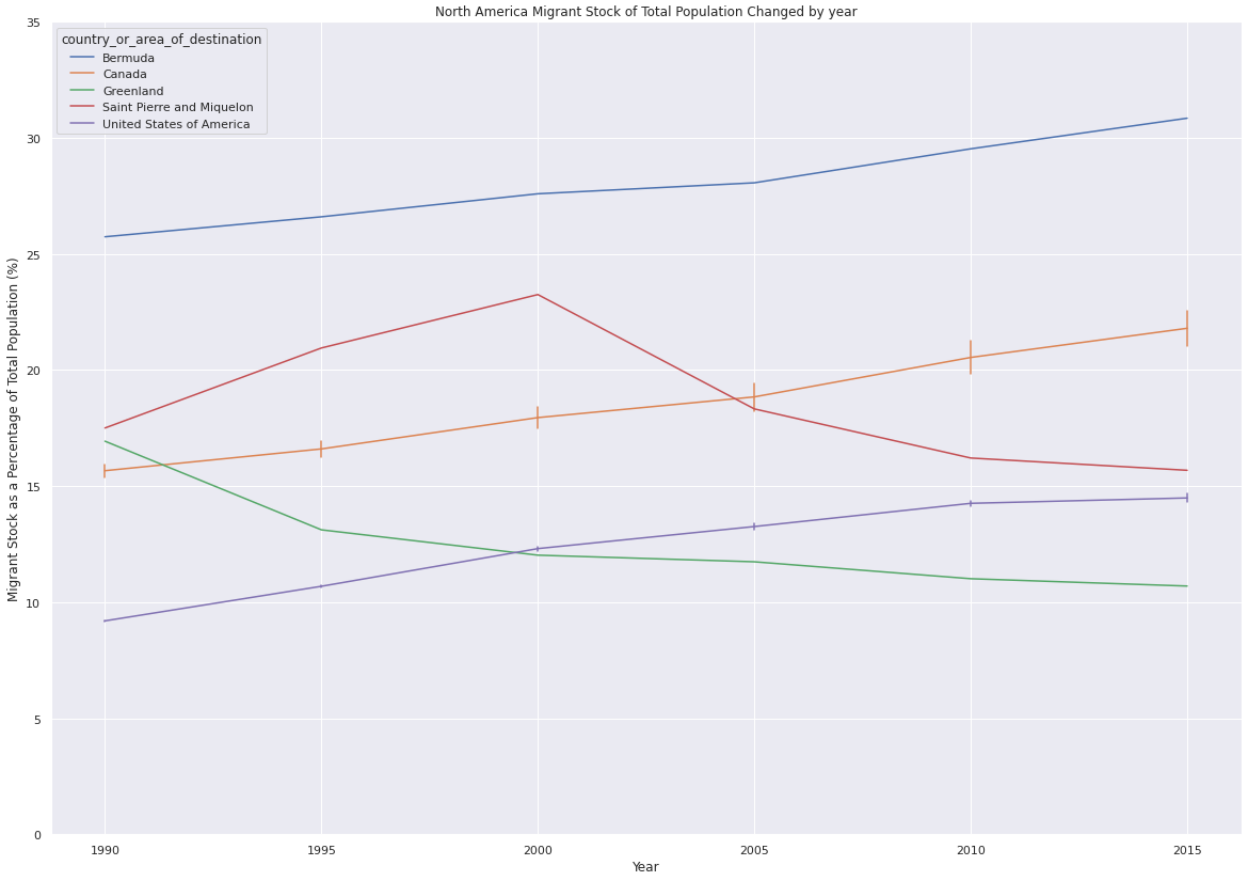
general analysis of the global population. Through color labels, this bar plot clearly presents the changes in the total population of each continent between 1990 and 2015. The plot shows that North America and Asia have a higher population base than other continents. It is still growing from 1990 to 2015. The population of Africa is also growing a little bit and then surpassed the population of Europe in 2015. For Europe, Latin America and Oceania, the population growth is very flat. This provokes thought and pointed to the next direction to explore: the composition of the population of North America, and whether migration has had an impact on demographic change.

Viz2: Bar plot from table1



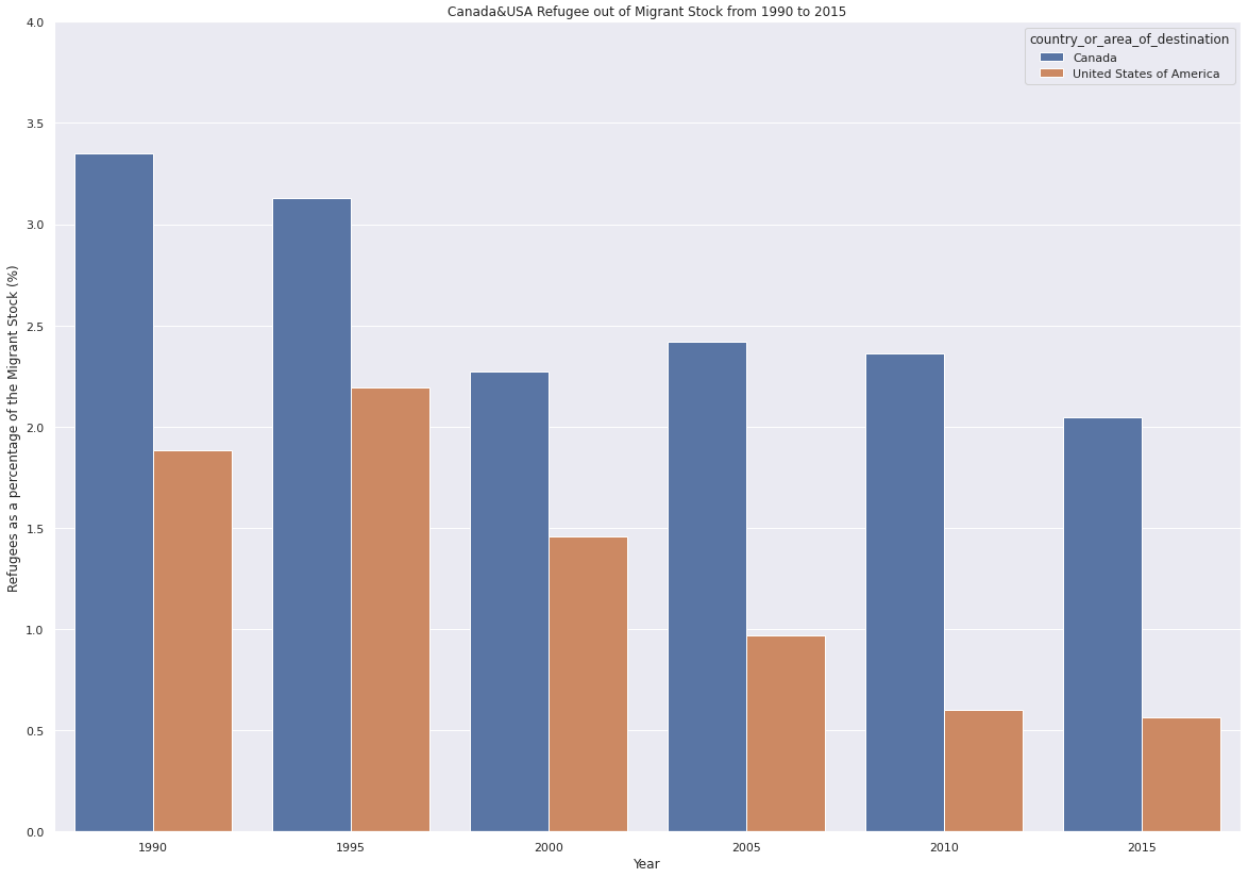
Let's continue the analysis on a larger scale and turn our attention to the migrant stock. This plot also uses the year as the x-axis and marks the regions (six continents) with colors. It is clear to see that North America has far more migrants than the others, while Asia has a large population growth (see above) but not many migrants. Combined with common sense, this is a very logical phenomenon, since the two largest immigrant countries in the world, the United States and Canada, are located in North America. The next visualization will take us into North America.

Viz3: Line plot from table3



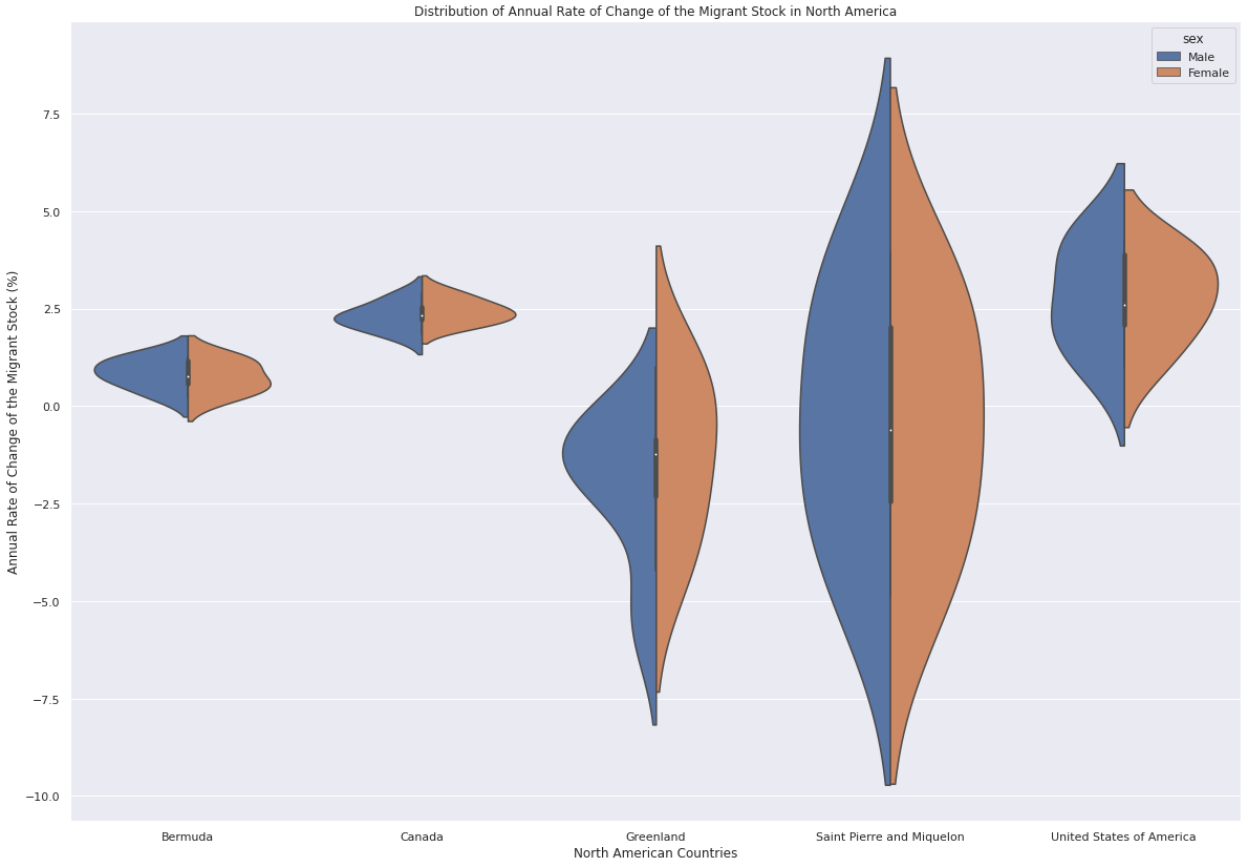
This line plot shows the ratio of immigrant population to total population for five countries within North America. Because it is a line plot, it clearly shows the trend over the years. Saint Pierre and Miquelon, Greenland, and Canada were all around 17% in 1990, and then rose slowly in Canada and fell slowly in Greenland. Upon investigation, it appears that the approval of financial independence by the French Parliament in 2000 had a negative impact on the country's already sluggish economy with a large outflow of people. This plot shows that the United States has a lower share of immigrants than Canada, and the growth trend is essentially the same. This is also consistent with the fact that the US has a stricter immigration policy compared to Canada.

Viz4: Bar plot from table6b



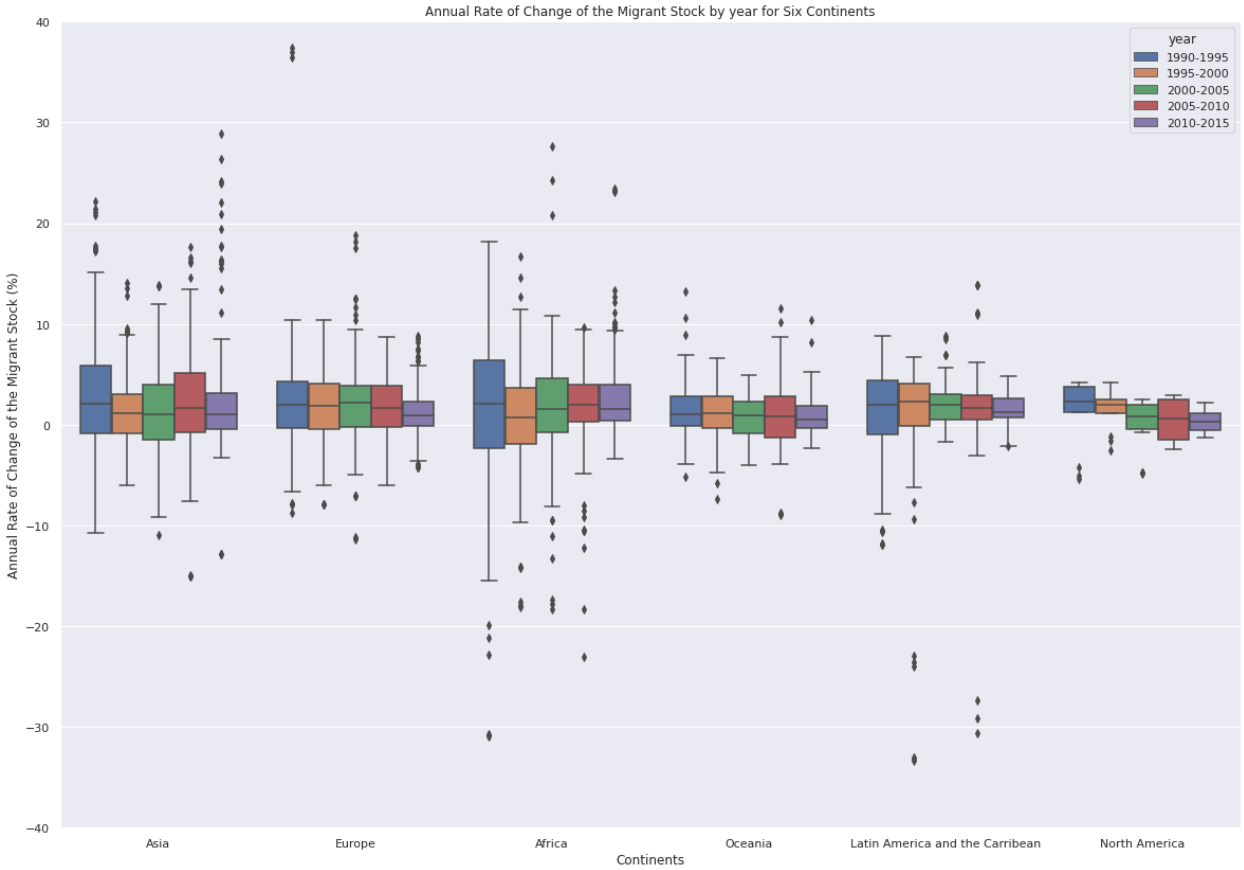
This bar chart shows the percentage of refugees to total migrants in the world's two largest countries of immigration. Canada has a higher percentage of refugees than the United States and has a downward trend between 1990 and 2000, after which it flattens out. The U.S. share of refugees declined each year after 1995, dropping from about half of Canada's share to a quarter. Interestingly, there was a paradoxical small climb in the U.S. refugee rate from 1990 to 1995. Perhaps in connection with history, one can speculate that the U.S. briefly adjusted its immigration policy, admitting a number of immigrants from the Middle East and Russia during that time period (Gulf War and the collapse of the Soviet Union). Bar charts, however, only show the mean (or other estimates), but in many cases it may be more informative to show the distribution of values at each level of a categorical variable. In such cases, other methods, such as box plots or violin plots, may be more appropriate.

Viz6: Violin plot from table5



A violin plot plays a similar role as a box and whisker plot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. Unlike a box plot, in which all of the plot components correspond to actual data points, the violin plot features a kernel density estimation of the underlying distribution. Violin diagrams can help us to observe the distribution of the data. This violin plot uses the five North American countries as the x-axis, showing their respective annual rates of change in migration, and distinguishes between men and women by color. The rate of change for Canada is clustered around 2.5, with females slightly higher than males. The annual rate of change for females in the United States is distributed around 3, but the distribution for males is not particularly concentrated, making it difficult to find specific corresponding data points. The annual rate of change for immigrants in Saint Pierre and Miquelon is very dispersed, ranging from -10 to 7.5. The images also show no outflows from Canada, trace outflows from the United States and Bermuda, and significant population movements (both losses and inflows) from Greenland and Saint Pierre and Miquelon. Violin plots may be an effective and attractive way to show multiple distributions of data simultaneously, but the estimation procedure is influenced by the sample size, and violins for relatively small samples may appear to be misleadingly smooth.

Viz5: Box plot from table5



Box plots show the distribution of quantitative data to facilitate comparisons between variables or between levels of categorical variables. Boxes show the quartiles of the data set, while whiskers extend to show the remainder of the distribution, except for points identified as "outliers" using the inter-quartile function method. After an in-depth analysis of the North American situation, back to the world again. The box plot shows the annual rate of change of the migrant population from 1990 to 2015 for six continents. box plot has the advantage that the viewer can visualize the median and interquartile range of the corresponding data, as well as its outliers. the box plot shows that the median of the annual rate of change for each continent is relatively stable and stays between 2 and 3. The difference between the upper quartile and the lower quartile is larger for Asia and Africa in 1990-1995.

Discussion

Six visualizations demonstrate the population growth of Asia and North America from 1990 to 2015 and ahead of other continents. The migrant population in North America also increased and was much higher than in the rest of the world at the same time. Considering the proportion of immigrant populations in North American countries, North America's population growth is strongly correlated with migration. In Canada, migrants contribute slightly more women than men.

In order to facilitate our visualization later when we reflect the practice, it is essential to classify the individual variables in advance. Although this dataset has six tables, the potential variables are few and all very similar. I classify them into strong continuity, weak continuity, and non-

continuity according to their properties. Define the variable with weak continuity is weak-continuous. Define the variable with strong continuity is strong-continuous. Define the variable with non-continuity is non-continuous. When the weak-continuous variable is used as the X-axis, the visualization can not only compare the data but also show the trend. Non-continuous variables can also be used as X-axis in order to describe only data comparisons. Colors can only be marked by non-continuous variables. The strong-continuous variable fits on the Y-axis. When it is used as the X-axis, the plot is essentially another form of bar chart, but usually used to show the count. Take the bar table1 as an example, the potential variables are year, international migrant stock, country_or_area_of_destination. The strong-continuous variable is only international migrant stock. The weak-continuous variable is year. Non-continuous variables are country_or_area_of_destination, continent, major_region, and sex. Observing these variables, it can be concluded that table1 is suitable for making line plots and bar plots (which require one weak continuity variable and one strong continuity variable) and not for making scatterplot because it does not have more than two strong continuity variables. It is also not suitable for box plot or violin plot because the strong continuous variables in this table show discrete data and cannot be stacked.

Moreover, Tufte's six principles are not absolutely correct. Reflecting on the six principles in light of the practice of making visualizations, it is not difficult to find that the item of causality deserves to be questioned. Visualization cannot help derive causality, only correlation. The effect of a variable on a phenomenon may well not be unique. Similarly, we cannot predict the future direction based on visualization alone, but can only summarize the existing trends. These trends can assist us in understanding or analyzing certain phenomena, but only statistical models that are constantly tested can identify causality as well as predictions.