Qian_Assignment 2

June 10, 2021

1 INF2199 Assignment 2

1.1 A deceptive bubble chart: *The positive impact of asian immigration on GDP growth in Canada between 1980 - 2013*

Jia Lei Qian

1.2 Introduction

Recently, Immigration, Refugees, and Citizenship Canada announced an unprecedented fast-entry lottery inviting 27,332 candidates who meet Canadian experience category requirements to apply for permanent residency in Canada. The Express Entry draw dropped down to 75 points, which is the lowest point in history. **Generally, economists agree that immigration is a necessary part of achieving economic growth and keeping taxpayer-funded systems stable and balanced.** Also, according to Statistics Canada, Asia was Canada's largest source of immigration and immigration. Therefore, this study will focus on this topic to explore the relationship between the number of Asian immigrants and Canada's economic (GDP) growth, and create a **misleading** data visualization.

Two datasets have been used in this assignment: Immigration to Canada IBM Dataset retrieved from Kaggle and Canada's GDP Data from World Bank Data. Since the immigration dataset only contains immigration information from 1980 to 2013, this analysis will focus on this time range. Meanwhile, the package Plotly has been used to illustrate in this illustration. Plotly Express is an open-source data visualization for both Python and R. It is written in JavaScript to make the graphics have internal interactivity.

1.3 Design Process

The entire design process is divided into the following two parts: **Data Exploration** and **Design Iteration**.

1.3.1 Part 1. Data Exploration: Data Preparation, cleaning, speculative design thought

This analysis first generates a pie chart of immigration numbers and a time series plot of Canada's total GDP from 1980 to 2013. ***The first pie chart shows that Asian immigrants make up more than half of Canada's immigrants, while the second time series plot reveals an overall increasing trend of total GDP except a shrink in GDP in 2009 compared to the previous year. The drop is highlighted in red.***

Step 1: Import required packages and the immigration dataset

[1]:
```
# For working with the data
import pandas as pd
import numpy as np
#Data visualization
import plotly.express as px
```

[2]:
```
#read xlsx file
df3= pd.read_excel("Canada.xlsx",sheet_name='Canada by Citizenship',skiprows=range(20),skipfooter=2)

#display first 5 rows of the dataset
df3.head()
```

[2]:

	Type	Coverage	OdName	AREA	AreaName	REG	RegName	DEV	DevName	1980	…	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
0	Immigrants	Foreigners	Afghanistan	935	Asia	5501	Southern Asia	902	Developing regions	16	…	2978	3436	3009	2652	2111	1746	1758	2203	2635	2004
1	Immigrants	Foreigners	Albania	908	Europe	925	Southern Europe	901	Developed regions	1	…	1450	1223	856	702	560	716	561	539	620	603
2	Immigrants	Foreigners	Algeria	903	Africa	912	Northern Africa	902	Developing regions	80	…	3616	3626	4807	3623	4005	5393	4752	4325	3774	4331
3	Immigrants	Foreigners	American Samoa	909	Oceania	957	Polynesia	902	Developing regions	0	…	0	0	1	0	0	0	0	0	0	0
4	Immigrants	Foreigners	Andorra	908	Europe	925	Southern Europe	901	Developed regions	0	…	0	0	1	1	0	0	0	0	1	1

[5 rows x 43 columns]

[3]:
```
#select the necessary columns
Canada_df=df3.copy()
Canada_df.drop(["AREA", "REG", "DEV", "Type","Coverage"], axis = 1,inplace=True)
#rename the column
Canada_df.rename(columns = {'OdName':'Country', 'AreaName':'Continent','RegName':'Region'}, inplace= True)
```

[4]:
```
#Sum up the total number of the immigration based on country
Canada_df["Total number"]= Canada_df.sum(axis=1)
```

Step 2: Simple Data visualization: Total number of immigration in Canada grouped by Continent (1980-2013)

[5]:
```
#Draw a pie chart to show the proportion of the number of immigrants in Canada␣
↪grouped by different continent
fig = px.pie(Canada_df, values='Total number', names='Continent',␣
↪color_discrete_sequence=px.colors.sequential.RdBu,
             title="Total number of immigrants in Canada from different␣
↪continents (1980-2013)")
fig.show()
```

Step 3: Import Canada GDP data and pre-process the dataset

[6]:
```
#Resource: World Bank Data: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?
↪locations=CA
Canada_gdp= pd.read_csv("Canadagdp.csv")

#Select the required columns
Canada_gdp=Canada_gdp.drop(["Continent","Region","DevName"],axis=1)

#Restructure the dataset
Canada_gdp2 = pd.melt(Canada_gdp,
                  id_vars=['Country'],
                  var_name='year',value_name='Total GDP')

#Change the type of the "year" column
Canada_gdp2['year'] = Canada_gdp2['year'].astype(float)
```

[7]:
```
#Show the head of the dataset
Canada_gdp2.head()
```

[7]:

	Country	year	Total GDP
0	Canada	1980.0	2.740000e+11
1	Canada	1981.0	3.060000e+11
2	Canada	1982.0	3.140000e+11
3	Canada	1983.0	3.410000e+11
4	Canada	1984.0	3.550000e+11

Step 4: Vsiaulization Trend of Canada Total GDP from 1980 to 2013

[8]:
```
fig = px.line(Canada_gdp2, x="year", y="Total GDP",title = "Trend of Canada␣
↪Total GDP from 1980 to 2013")
fig.add_vrect(x0=2008,x1=2009, fillcolor="salmon", opacity=0.5,layer="below",␣
↪line_width=0,)
fig.show()
```

1.3.2 Part 2. Design Iteration

To further explore the relationship between immigration and economic growth, I decided to use **a bubble chart** with Plotly Express to create the first iteration. A bubble graph is **a variation of a scatter plot that can visualize three different measures simultaneously.** Bubble diagrams are usually used to show and compare the relationship and distribution between data. The correlation between data dimensions is analyzed by comparing the position and size of bubbles. It is **easy to read and to make relative comparisons**. However, it is also a controversial graphic which could mislead the readers. Bubble charts are **difficult to compare numerical variables**. It takes time for the reader to interpret all of the various sections of the graph. Therefore, it fits this assignment's theme of creating a deceptive chart and can make a misleading data story to depict an incorrect immigration impact on the Canadian economy. The entire design process involves four iterations.

1.3.3 *Iteration 1: Basic Scatter Plot based on immigration data*

In the first iteration, I drew a dynamic bubble chart of the total number of Asian immigrants grouped by country classification. It shows The total number of immigrants from different Asian countries to Canada from 1980 to 2013. The bubble's size represents how much it has eclipsed the number of migrants, while two colors differentiate between developing and developing countries. ****According to the graph, it is obvious to conclude that as of 2013, India had the most significant number of immigrants to Canada, followed by China and the Philippines. In contrast, the number of Japanese immigrants to Canada as a developed country is not much.****

[9]:
```
Canada_df1=Canada_df.copy()
#rename the column
Canada_df1.rename(columns = {'Total number':'Total number of asian␣
↪immigrants'}, inplace= True)
```

[10]:
```
#Plot a bubble chart
fig = px.scatter(Canada_df1[Canada_df1.Continent == 'Asia'], x='Total number of␣
↪asian immigrants', y='Country',
                 size='Total number of asian immigrants', color='DevName',
                 title = "Total number of Asian immigrants grouped by country␣
↪classification from 1980 to 2013")
fig.show()
```

1.3.4 *Iteration 2: Add timeline and facet the plot to pick apart the region*

Since the first draft does not show the difference every year and there is only one developed country in Asia, the second version adds an animated timeline. It groups the data by different Asian regions. By dividing the plot into four subplots based on country classification and setting "animation_frame='year"' and "facet_col='Region"', one can see **how the bubble chart evolved. By clicking the start button, the movement of the bubbles indicates the movement trend of immigrants from this country to Canada**. It is interesting to see that **the most considerable fluctuations in the number of migrants have occurred in Southern and Eastern Asia and the least significant in the Middle East over the period.**

[11]:
```
#Limit the region into Asia
Canada_df_reg = Canada_df.query("Continent in ['Asia']")
```

[12]:
```
#Decide to melt the dataset
Canada_df2 = pd.melt(Canada_df_reg,
                     id_vars=['Country','Continent' ,'Region', 'DevName','Total␣
 ↪number'],
                     var_name='year',value_name='Immigration number')
```

[13]:
```
#Plot a bubble chart with animated timeline
fig = px.scatter(
    Canada_df2,
    x="Immigration number",
    y="Country",
    animation_frame="year",
    animation_group="Country",
    hover_name="Region",
    facet_col="Region",
    title="Changing number of Asian immigrants grouped by country region from␣
 ↪1980 to 2013"
)
fig.update_layout(autosize=False,height=500,width=1000,
                  font=dict(size=10))
fig.show()
```

1.3.5 *Iteration 3: Combine with the GDP data and change the color and size of the bubbles*

The third version combines the immigration and GDP dataset to **mislead the viewers**. It changes the size and color of bubbles. Specifically, the larger the bubbles' size represents the higher total GDP. On the other hand, the lighter the bubbles' color means, the more immigrants from that country in the corresponding year. It is clear to find that **with the increase of Asian immigrants from different countries, the overall economy shows a trend of growth, which corresponds to the previous time series plot.**

[14]:
```
#Merge the dataset with the GDP data
Canada_df3=Canada_df2.merge(Canada_gdp2,on="year",how="left")
#Drop irrelevant column
Canada_df3=Canada_df3.drop(["Country_y"],axis=1)
#Rename the column
Canada_df3.rename(columns = {'Country_x':'Country'}, inplace= True)
#Show the head of the dataframe
Canada_df3.head()
```

[14]:

	Country	Continent	Region	DevName	Total number	\
0	Afghanistan	Asia	Southern Asia	Developing regions	58639	
1	Armenia	Asia	Western Asia	Developing regions	3310	

```
2   Azerbaijan      Asia   Western Asia  Developing regions          2649
3      Bahrain      Asia   Western Asia  Developing regions           475
4   Bangladesh      Asia  Southern Asia  Developing regions         65568

   year  Immigration number     Total GDP
0  1980                  16  2.740000e+11
1  1980                   0  2.740000e+11
2  1980                   0  2.740000e+11
3  1980                   0  2.740000e+11
4  1980                  83  2.740000e+11
```

[15]:
```
# Plor a misleading bubble chart
fig = px.scatter(
    Canada_df3,
    x="Immigration number",
    y="Country",
    animation_frame="year",
    animation_group="Country",
    size="Total GDP",
    color="Immigration number",
    facet_col="Region",
    title="Changing number of Asian immigrants sizing by Total GDP in different␣
 ↪regoins (1980-2013)"
)
fig.update_layout(autosize=False,height=500,width=1100,
                  font=dict(size=10))
fig.show()
```

It is worth noticing that there exist ***four misleading points***:

1. **Cherry picking data**: The graph did not display the immigration number's full scale. The audience cannot see the bubbles from the countries with the most immigrants such as India and China in the chart.
2. **Manipulating the Y-axis**: The graph did not list all countries. The audience is required to zoom in to see the specific country.
3. **Going against conventions**: The darker shades are used to describe lower number of immigrants, which can confuse and mislead readers.
4. **Spurious relationship**: The positive relationship between the number of Asian immigrants and Canada's total GDP is spurious. Without any proof from statistical analysis, one cannot conclude based on the above bubble chart.

1.3.6 *Iteration 4: Change the name of X-axis, remove the grouping option, switch the color scale, and adjust plot size*

Lastly, to make the misleading scientific graphic more convincing and seems to make sense, **the first three misleading points listed in iteration 3 have been corrected.** To be specific, the final version removes the grouping option and adjusts the graphic size so that readers can view the entire scale of the visualization. Also, the name of the x-axis has been changed to "Asian

Countries sizing by total GDP," which helps readers better understand the meaning of the bubble's size. Furthermore, the scale of the color has been changed to be diverging, which allows the readers to know how to interpret the data intuitively. Notice that the final version remains the spurious positive relationship between Asian immigration and the total GDP in a more compelling way.

[16]:
```
Canada_gdp3=Canada_gdp2.copy()

# Drop irrelevant column
Canada_gdp3=Canada_gdp3.drop(["Country"],axis=1)

# Merge the gdp dataset with the immigration dataset
Canada_df4=Canada_df2.merge(Canada_gdp3,on="year",how="left")

# Rename the column to help readers better understand the meaning of bubbles'␣
↪size
Canada_df4.rename(columns = {'Country':'Asian Countries sizing by total GDP'},␣
↪inplace= True)
Canada_df4.head()
```

[16]:

	Asian Countries sizing by total GDP	Continent	Region	DevName	Total number	year	Immigration number	Total GDP
0	Afghanistan	Asia	Southern Asia	Developing regions	58639	1980	16	2.740000e+11
1	Armenia	Asia	Western Asia	Developing regions	3310	1980	0	2.740000e+11
2	Azerbaijan	Asia	Western Asia	Developing regions	2649	1980	0	2.740000e+11
3	Bahrain	Asia	Western Asia	Developing regions	475	1980	0	2.740000e+11
4	Bangladesh	Asia	Southern Asia	Developing regions	65568	1980	83	2.740000e+11

[17]:
```
# Changing the color and size of the bubbles
fig = px.scatter(
    Canada_df4,
    x="Immigration number",
    y="Asian Countries sizing by total GDP",
    animation_frame="year",
    animation_group="Asian Countries sizing by total GDP",
    size="Total GDP",
    color="Immigration number",
    #the scale of the color scales has been changed to be diverging
    color_continuous_scale=px.colors.diverging.Temps,
    title="Changing number of Asian immigrants sizing by total GDP from 1980 to␣
↪2013"
)
```

```
fig.update_layout(autosize=False,height=1000,width=1100,
                  font=dict(size=10),xaxis=dict(range=[0,42584]))
fig.show()
```

1.4 Conclusion

In a nutshell, in this assignment, a deceptive bubble graph is created to depict Asian immigration impacts on economics. The diagram is based on conventional wisdom among economists that new immigrants can boost a country's economy. However, it manipulates two datasets without considering confounding variables and misleads the reader to understand the relationship between immigration and economic growth through changing the bubbles' size. It is possible that digital media without statistics backgrounds, such as business magazines, falsely use the diagram to explain why Canada's federal government plans to welcome more newcomers in the upcoming years.