Analysis and Visualization of Social Factors Impacting Fertility Rate Which Leads to Population Aging in Canada

Team 13： Yuxiang Wang, Huaien Gu

Introduction

Population aging has been identified as a high-priority problem in many developed countries. It describes a phenomenon where the country's **population distribution is being shifted towards the elderly population with a decline in the proportion of the younger population**. This could be resulting in overburdening of the welfare system, shortage in labor supplies, decline in productivity and many other serious consequences that could ultimately lead to economical hardship.

Currently, the only feasible solution is to **improve birth rates**, which has been heavily incentivized in many developed countries like Korea and New Zealand, in order to combat the problem of Population Aging.

In this paper, we will investigate Canada's birth rates by province, which is a key determinant for whether Canada is subjected to population aging. By analyzing the birth rate by province, we can further investigate whether the provincial birth rate is associated with other influencing social factors such as provincial GDP per capita, unemployment rate, crime rate, and education level. Therefore, the paper will reach a conclusion on the effect of the potential contributing factors by analyzing the correlation between these factors and the birth rate. From our result, the policy makers can visualize which practice is more effective in terms of increasing the birth rate and to effectively combat the problem with population aging.

Guiding questions

The following Guiding questions aim to establish the existance of the problem of Population Aging in Canada, and to visualize the severity of the problem of low birth rate in each province. Then, We will analyze the correlation between each proposed social factor and birth rate. By cross-analyzing each province, we can have enough data to accurately determine whether the said social factors is impacting the birth rate.

We will only be using the year range from 2000 to 2020 due to data set limitations, also because earlier data is less indicative for analyzing recent trends.

1. Is Canada subjected to population aging?

2. If Canada is subjected to population aging, What is the severity of the problem with low birth rate for each province?

3. What are the education levels in each Canadian province, and how will they affect the birth rate?

4. Will GDP and crime rate have an impact on the birth rate?

5. How much the unemployment affect the birth rate?

Dataset

- “Population ages 65 and above - Statistics Canada” [10] This dataset is derived from “The World Bank". We have the permission to use their data. The dataset is excel tabular data with 271 rows by 61 columns .
- “Population ages 0-14 – Statistics Canada” [9] This dataset is derived from “The World Bank". We have the permission to use their data. The dataset is excel tabular data with 271 rows by 61 columns.
- “Yearly Canadian population estimation - Statistics Canada” [8] This dataset is derived from “Statistics Canada". We have the permission to use their data. The dataset is excel tabular data with 21 rows by 61 columns (customizable year range from source to align with rest of data column range.)
- “Crude birth rate and total fertility rate- Statistics Canada” [5] This dataset is derived from “Statistics Canada". We have the permission to use their data. The dataset is grouped excel tabular data with many subgroups. Overall size is 145 rows by 22 columns.
- “Educational attainment in the population aged 25 to 64” [1] This dataset is derived from “open.canada.ca”. We have permission to use their data. The dataset is grouped excel tabular data. Overall size is 616 rows by 16 columns, we will focus on provinces, date, education level and the percentage.
- “Unemployment rate, participation rate and employment rate by sex, annual” [2] This dataset is derived from “Statistics Canada”. We have permission to use their data. The dataset is grouped excel tabular data. Overall size is 33 rows by 35 columns, we will focus on provinces, unemployment rate and percentage.
- “Gross domestic product (GDP) at basic prices, by industry, provinces and territories” [6] This dataset is derived from “Statistics Canada”. We have permission to use their data. The dataset is grouped excel tabular data. Overall size is 13 rows by 12 columns, we will focus on provinces, date and dollars.
- “Incident-based Crime Statistics. - Statistics Canada” [7] This dataset is derived from “Statistics Canada”. We have permission to use their data. The dataset is grouped excel tabular data. Overall size is 50 rows by 23 columns

Analysis

We will puch forward our project through two parts, firstly... Secontly...

In [1]:
```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import geopandas as gpd
import plotly as plotly
import plotly.offline as py
import plotly.graph_objs as go
import plotly.express as px
import csv
import warnings
warnings.filterwarnings("ignore") # to hide/ignore warnings
from urllib.request import urlopen
import json
```

Guiding Question #1

1. Is Canada subjected to population aging?

In this question, we will explore the trend of the elder population (age 65 and above) and the younger population (age 14 and below) from year 2000-2020, and the change in population proportion. By visualizing the proportion and the trend of the young and old population, we can easily contrast whether Canada is subjected to population aging.

Data Source: *The World Bank, Statistics Canada*

Data wrangling for Q1

(This is a general guideline, please refer to more detailed procedures from in-line comments)

- Import Data - Data on Young Population (age 14 and below) from *The World Bank.*
- Drop columns that are not needed.
- Scoop out only Year 2000 to 2020.
- Eliminate data for other countries, keep only data for Canada.
- Same procedure for Data on Elder Population (age 65 and above) from *The World Bank.*

Then

- Import and Wrangle Overall Canadian Popolation Data from *Statistics Canada*.
- Here, we need to discard the age groups preset by Statistics Canada, keep only overall Canadian Population
- Break the comma-seperated thousands, fix bad column names
- Convert all population data from string to numeric dtypes for calculation.

Finally

- Build a new datafram contains useful columns from the cleaned datasets.
- Use the new dataframe to plot line charts that illustrate trends.
- Make new column 'proportion' by calculation from age group population divided by overall population.
- Plot trend of proportion change to illustrate the problem

In [2]:
```
#importing dataset of young'population (age 15 and below)
#Data source from The World Bank. Accessesible at :https://data.worldbank.org/
indicator/SP.POP.0014.TO?end=2020&locations=CA&start=1 960&view=chart
young_pop=pd.read_csv('pop_young.csv' , index_col=[0])
print('\n')
print('Before wrangling the data looks like this:')
display(young_pop.head())

#Filter out the interested year range: 2000-2019
cols=list(young_pop.columns)
young_pop= young_pop[cols[0:1]+cols[43:]]
#filter out for ONLY the canadian population
young_pop=young_pop.loc[young_pop['Country Code']== 'CAN']
young_pop.drop("Country Code", axis=1, inplace=True)
print('\n')
print('After wrangling the Data Looks like this')
print('\n')
display(young_pop.head(5))
```

```
Before wrangling the data looks like this:
```

	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963
Country Name							

Aruba	ABW	Population ages 0-14, total	SP.POP.0014.TO	23769.0	24035.0	24139.0	24091.0	2
Africa Eastern and Southern	AFE	Population ages 0-14, total	SP.POP.0014.TO	57144288.0	58943932.0	60748348.0	62553938.0	6434
Afghanistan	AFG	Population ages 0-14, total	SP.POP.0014.TO	3791398.0	3892774.0	3987207.0	4079604.0	417
Africa Western and Central	AFW	Population ages 0-14, total	SP.POP.0014.TO	40179920.0	41258443.0	42322255.0	43381248.0	4443
Angola	AGO	Population ages 0-14, total	SP.POP.0014.TO	2298278.0	2366950.0	2439505.0	2504062.0	254

5 rows × 64 columns

```
After wrangling the Data Looks like this
```

	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2**
Country Name									
Canada	5880513.0	5843483.0	5816282.0	5781391.0	5742896.0	5699388.0	5667703.0	5635757.0	56160

1 rows × 21 columns

In [3]:

```
#Importing Older population dataset, data from The World Bank
#Available at: https://data.worldbank.org/indicator/SP.POP.65UP.TO?end=2020&lo
cations=CA&start=1960&view=chart
old_pop=pd.read_csv('pop_old.csv')
print('\n')
print('Before wrangling the data looks like this:')
display(old_pop.head(5))
#Filter out the interested year range: 2000-2019 for Canada only
colso=list(old_pop.columns)
old_pop= old_pop[colso[0:1]+colso[-21:]]
old_pop=old_pop.loc[old_pop['Country Name']== 'Canada']
print('\n')
print('After wrangling the Data Looks like this')
print('\n')
display(old_pop)
```

```
Before wrangling the data looks like this:
```

	Country Name	**Country Code**	**Indicator Name**	**Indicator Code**	**1960**	**1961**	**1962**	**1963**	**19**
0	Aruba	ABW	Population ages 65 and above, total	SP.POP.65UP.TO	1346.0	1433.0	1513.0	1588.0	167
	Africa		Population						

1	Africa Eastern and Southern	AFE	ages 65 and above, total	SP.POP.65UP.TO	4043770.0	4151048.0	4251472.0	4347316.0	444130
2	Afghanistan	AFG	Population ages 65 and above, total	SP.POP.65UP.TO	251763.0	257489.0	262225.0	265890.0	26839
3	Africa Western and Central	AFW	Population ages 65 and above, total	SP.POP.65UP.TO	2829296.0	2912796.0	2989600.0	3058187.0	311679
4	Angola	AGO	Population ages 65 and above, total	SP.POP.65UP.TO	164027.0	168439.0	171872.0	174434.0	17610

5 rows × 65 columns

```
After wrangling the Data Looks like this
```

	Country Name	2000	2001	2002	2003	2004	2005	2006	2007	
35	Canada	3855655.0	3931204.0	4003899.0	4070291.0	4144518.0	4229591.0	4321765.0	4419182.0	4!

1 rows × 22 columns

In [4]:

```python
#Importing and cleaning Overall Canadian population, data from Statistics Cana
da
#Available at: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=171000050
1&pickMembers%5B 0%5D=1.1&pickMembers%5B1%5D=2.1&cubeTimeFrame.startYear=1995&
cubeTimeFra me.endYear=2020&referencePeriods=19950101%2C20200101.

#This dataset consists of age groups which could potentially used to replace t
he above datasets, but we chose to use the above dataset because this one is o
nly the estimated population
#so, the above two datasets provide more accuracy for the age group populatio
n.
all_pop=pd.read_csv("Canadian Population.csv",thousands=",")
print('\n')
print('Before wrangling the data looks like this:')
display(all_pop.head())


#Filter out All ages population which is the overall Canadian population at ea
ch year.
all_pop=all_pop.loc[all_pop['Age group 3 5']== 'All ages']

#change the sturborn comma seperated value to float.
all_pop['2000'] = all_pop['2000'].str.replace(',', '').astype(float)
#Change the default name to a readable name
all_pop.rename(columns={'Age group 3 5': 'All Canadian Population'},inplace=Tr
ue)
```

```
print('\n')
print('After wrangling the Data Looks like this')
display(all_pop)
```

Before wrangling the data looks like this:

	Age group 3 5	2000	2001	2002	2003	2004	2005	2006	2007	200
0	All ages	30,685,730	31020902	31360079	31644028	31940655	32243753	32571174	32889025	3324711
1	0 to 4 years	1,790,699	1754354	1723292	1705276	1705030	1708245	1727509	1753475	179355
2	5 to 9 years	2,036,949	2017049	1988754	1947657	1905050	1864782	1824535	1801934	179048
3	10 to 14 years	2,055,843	2079739	2114746	2139150	2141832	2124530	2096117	2065911	203229
4	15 to 19 years	2,095,909	2115915	2128264	2129918	2144151	2176159	2211529	2232083	225094

5 rows × 22 columns

After wrangling the Data Looks like this

	All Canadian Population	2000	2001	2002	2003	2004	2005	2006	2007	
0	All ages	30685730.0	31020902	31360079	31644028	31940655	32243753	32571174	32889025	33

1 rows × 22 columns

In [5]:

```
#Combining the 3 processed datasets together for future work.
print('\n')
print('Concatenating datasets without cleaning looks like this')
df=pd.concat([young_pop,old_pop,all_pop])
display(df.head())

#Cleaning the datasets, setting appropriate index, transform the dataset into
 more readable frame.
df.reset_index(inplace=True)
df= df.transpose()
df.reset_index(inplace= True)
df.columns=df.iloc[0]
#Renaming columns into more readable names
df=df.rename({'index':'Year', 'Canada': "Young population (15 and below)"},axi
s=1)
df=df.rename({35:'Old Population (65 and above)', 0: "Overall"},axis=1)
#skip the first row which was used to rename column names
df = df.iloc[1: , :]
df.dropna(inplace=True)

print('\n')
```

```
print('After cleaning the dataset looks like this')
display(df.head())
```

```
Concatenating datasets without cleaning looks like this
```

	2000	2001	2002	2003	2004	2005	2006	200
Canada	5880513.0	5843483.0	5816282.0	5781391.0	5742896.0	5699388.0	5667703.0	5635757.
35	3855655.0	3931204.0	4003899.0	4070291.0	4144518.0	4229591.0	4321765.0	4419182.
0	30685730.0	31020902.0	31360079.0	31644028.0	31940655.0	32243753.0	32571174.0	32889025.

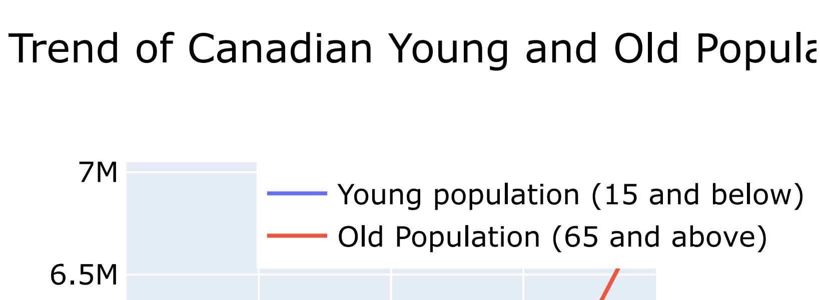
3 rows × 23 columns

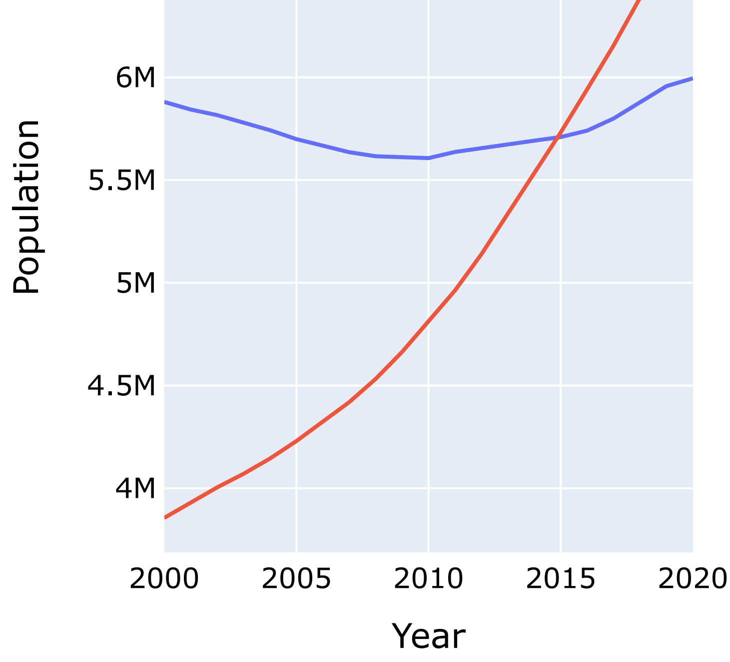
```
After cleaning the dataset looks like this
```

	Year	Young population (15 and below)	Old Population (65 and above)	Overall
1	2000	5.88051e+06	3.85566e+06	3.06857e+07
2	2001	5.84348e+06	3.9312e+06	3.10209e+07
3	2002	5.81628e+06	4.0039e+06	3.13601e+07
4	2003	5.78139e+06	4.07029e+06	3.1644e+07
5	2004	5.7429e+06	4.14452e+06	3.19407e+07

In [6]:

```
#Creating line chart to show the trend of the old and young population from ye
ar 2000 to 2020
fig1 = go.Figure()
fig1.add_trace(go.Scatter(
    x=df['Year'],
    y=df['Young population (15 and below)'],
    name="Young population (15 and below)" ))
fig1.add_trace(go.Scatter(
    x=df['Year'],
    y=df['Old Population (65 and above)'],
    name="Old Population (65 and above)" ))
fig1.update_layout(
    title="Trend of Canadian Young and Old Population from 2000 to 2020",
    xaxis_title="Year",
    yaxis_title="Population",
    font=dict(
        size=14,
        color="black"))
fig1.show()
```





At this point, we can see a clear upward trend for the elder population since 2000.

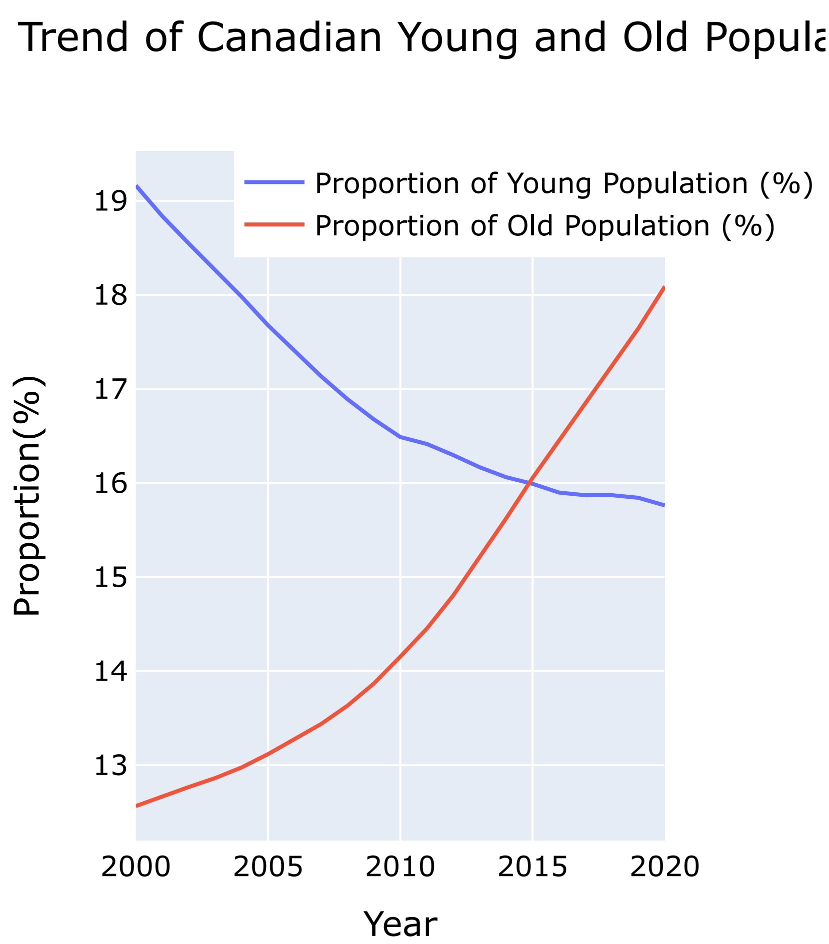
However, the young population almost stayed as a flatline, does it mean everything's good? The answer is no, we will see why it is when we translate it into proportion.

In [7]:
```
#Calculate population proportion and add them to the datafram
df['Overall']=pd.to_numeric(df['Overall']) #converting populatin from str-type to float
df['Young_proportion']=df['Young population (15 and below)'].div(df['Overall'].values)*100 #dividing population group by overall population, *100 to get percentage
df['Old_proportion']=df['Old Population (65 and above)'].div(df['Overall'].values)*100
display(df.head(5))
```

	Year	Young population (15 and below)	Old Population (65 and above)	Overall	Young_proportion	Old_proportion
1	2000	5.88051e+06	3.85566e+06	30685730.0	19.1637	12.565
2	2001	5.84348e+06	3.9312e+06	31020902.0	18.8372	12.6728
3	2002	5.81628e+06	4.0039e+06	31360079.0	18.5468	12.7675
4	2003	5.78139e+06	4.07029e+06	31644028.0	18.2701	12.8627
5	2004	5.7429e+06	4.14452e+06	31940655.0	17.9799	12.9757

In [8]:
```
#Plotting the trend of proportion of the two age groups from 2000 to 2019
fig2 = go.Figure()
fig2.add_trace(go.Scatter(
    x=df['Year'],
    y=df['Young_proportion'],
    name="Proportion of Young Population (%)" ))
fig2.add_trace(go.Scatter(
    x=df['Year'],
    y=df['Old_proportion'],
    name="Proportion of Old Population (%)" ))
fig2.update_layout(
```

```
    title="Trend of Canadian Young and Old Population Proportion from 2000 to 2020",
    xaxis_title="Year",
    yaxis_title="Proportion(%)",
    font=dict(
        size=14,
        color="black"))
fig2.show()
```



Importance of analysis from Q1

Here, we can clearly see the **proportion of elder population has been climbing up, while the young population proporiton has been going stright downwards.**

Since 2015, we witnessed the proportion of elder population surpassed the young population, and the difference of proportion of the two age groups is getting larger every year.

This is important because the chart clearly shows the country is subjected to population aging. Without intervention, the aforementioned serious consequences from introduction can very likely come true in Canada.

Guiding Question #2

2. What is the severity of low birth rate for each province?

Since we can conclude Canada is subjected to Population Aging from Q1, and we know from introduction that the only feasible solution to combat the problem is to improve birth rate to increase proportion of young population to fill the labor demands. Therefore, In Q2, we hope to illustrate live birth rate from each province to further visualize the severity of the problem with birth rate in each province.

Data Source: Github Open Source Dataset, Statistics Canada

Data wrangling for Q2

(This is a general guideline, please refer to more detailed procedures from in-line comments)

- Import and Wrangle 'Crude Birth Rate in Canada' from *Statistics Canada*.
- Drop columns that are not needed.
- Eliminate empty cells.
- Drop the unavailable data points denoted as '..' by Statistics Canada
- numerate str-type numbers for calculation.
- Clean the column names, reaarange dataframe to produce understandable and presentable data.

Then

- Import and Wrangle raw geojson data for Canada only.
- Create a dictionary to store provincial ID from geojson for mapping purpose.
- Match provincial ID to the individual province from 'Crude Birth Rate in Canada' dataset,

Finally

- Get the average of yearly difference in birth rate from each province since 2015 (variable to be mapped out).
- Use Plotly Express.choropleth to map out the average birth rate in each province.

In [9]:

```python
#Importing Birth Rate dataset from Statistics Canada
#Available at https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310041801
br=pd.read_csv('Crude_Birth_rate.csv', index_col=[0])
print('\n')
print('raw data looks like this')
display(br.head())

br=br.dropna() #remove the empty space that was intended for visual appeal in Excel.
br.replace("..", np.nan, inplace=True) # remove unavailable data denoted as '..' by Statistics Canada
br.dropna(inplace=True) #Deleting the empty cells
br.reset_index(inplace=True)
br['Canada, place of residence of mother 12']=br['Canada, place of residence of mother 12'].apply(lambda x: x.split(',')[0]) #Getting the province name, without the tailing description.
br.rename(columns={'Canada, place of residence of mother 12':'Province'},inplace=True)
br=br.iloc[1:,:] # skip 1st row which was used to name column.

#tweak the dataframe for line graph only,dedicated for graph Figure 3.
br_line_data= br.transpose().reset_index()
br_line_data.columns=br_line_data.iloc[0]
br_line_data= br_line_data.iloc[1:, :]
br_line_data.rename(columns={'Province':'Year'},inplace=True)
print('\n')
print('processed data for line plot for figure 3')
display(br_line_data.head())
```

raw data looks like this

raw data looks like this

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	...	2011	2012	2013	201
Canada, place of residence of mother 12															
Canada, place of residence of mother 12	10.7	10.8	10.5	10.6	10.6	10.6	10.9	11.2	11.4	11.3	...	11	11	10.8	10.
Newfoundland and Labrador, place of residence of mother	9.2	9	9	8.9	8.7	8.8	8.9	8.9	9.6	9.5	...	8.5	8.3	8.6	8.
Prince Edward Island, place of residence of mother	10.6	10.1	9.7	10.3	10.1	9.7	10.2	10.1	10.7	10.4	...	10	9.1	9.8	9.
Nova Scotia, place of residence of mother	9.8	9.6	9.3	9.2	9.3	9.1	9	9.5	9.8	9.6	...	9.4	9.3	9	9.
New Brunswick, place of residence of mother	9.8	9.6	9.4	9.5	9.3	9.2	9.4	9.6	9.9	9.9	...	9.4	9.3	9.2	9.

5 rows × 21 columns

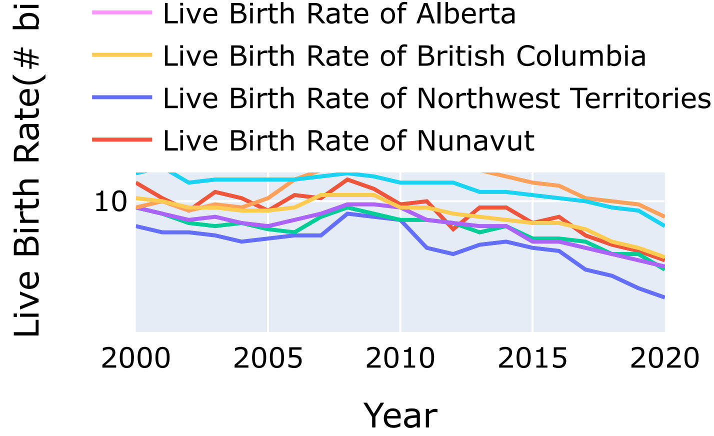
processed data for line plot for figure 3

	Year	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec	Ontario	Manitoba	Saskatchewan	Alberta
1	2000	9.2	10.6	9.8	9.8	9.8	10.9	12.3	12.1	12.3
2	2001	9	10.1	9.6	9.6	10	11.1	12.2	12.3	12.3
3	2002	9	9.7	9.3	9.4	9.7	10.6	12	11.8	12.4
4	2003	8.9	10.3	9.2	9.5	9.9	10.7	12	12.1	12.7
5	2004	8.7	10.1	9.3	9.3	9.8	10.7	11.8	12	12.6

In [10]:

```
#Plotting the trend of live birth rate
fig3 = go.Figure()
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Newfoundland and Labrador'],
    name="Live Birth Rate of Newfoundland and Labrador" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Prince Edward Island'],
    name="Live Birth Rate of Newfoundland and Labrador" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Nova Scotia'],
```

```
    name="Live Birth Rate of Nova Scotia" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['New Brunswick'],
    name='Live Birth Rate of New Brunswick'))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Quebec'],
    name="Live Birth Rate of Quebec" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Ontario'],
    name="Live Birth Rate of Ontario" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Manitoba'],
    name="Live Birth Rate of Manitoba" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Saskatchewan'],
    name="Live Birth Rate of Saskatchewan" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Alberta'],
    name="Live Birth Rate of Alberta" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['British Columbia'],
    name="Live Birth Rate of British Columbia" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Northwest Territories'],
    name="Live Birth Rate of Northwest Territories" ))
fig3.add_trace(go.Scatter(
    x=br_line_data['Year'],
    y=br_line_data['Nunavut'],
    name="Live Birth Rate of Nunavut" ))
fig3.update_layout(
    title="Live Birth Rate of Each Canadian Province from 2000 to 2020",
    xaxis_title="Year",
    yaxis_title="Live Birth Rate(# birth per 1000 women)",
    font=dict(
        size=14,
        color="black"))
fig3.show()
```



From the graph, we can see a **downward trend for live birth rate in almost all provinces in Canada, especially since the year 2015.** It is possible for experts to investigate what happened in year 2015 that caused the all-around decline in the country's birth rate.

In [11]:

```
#Creating new dataset for illustrating average live birth rate since year 2015
with map.
#we chose year 2015 - 2020 because we established 2015 is the year when young-
population was surpassed by elder population, and where a clear decline in liv
e birth rate in observed.
#Also, we want a more recent live birth rate to make our finding more up-to-da
te. Thus the year range 2015-2020.

br_for_plot=br[['Province','2015','2016','2017','2018','2019','2020']]
cols = br_for_plot.columns.drop('Province') #get rid of string-typed province
 for calculation.
br_for_plot[cols] = br_for_plot[cols].apply(pd.to_numeric)

br_diff=br_for_plot[cols].diff(axis=1) # calculate the difference of live birt
h rate from each year
br_diff.dropna(axis=1,inplace=True)
br_diff['Average']=br_diff.mean(axis=1)# Calculate the average of the live bir
th rate from year 2015 to 2020.
br['Average BR Since 2015']=br_diff['Average']# attaching the calculated mean
 birth rate to the main dataframe.


#Loading Canadian geojson data for mapping.
with urlopen('https://raw.githubusercontent.com/codeforgermany/click_that_hoo
d/main/public/data/canada.geojson') as response:
    provs = json.load(response)

#Create a dictionary to store provincial ID from geojson for mapping purpose.
prov_id_map={}
for feature in provs['features']: #use for-loop to avoid writing dozens of 'i
f' statements.
    feature['id']=feature['properties']['cartodb_id']
    prov_id_map[feature['properties']['name']]=feature['id']

br['id']=br['Province'].apply(lambda x: prov_id_map[x]) #creating a column 'i
d' to match the provincial id to each province from the main dataframe.

warnings.filterwarnings("ignore") #skip warning for version-compatibility.

print('\n')
print('After wrangling and processing, this is the dataframe that are to be gr
```

```
aphed')
display(br.head())
```

After wrangling and processing, this is the dataframe that are to be graphed

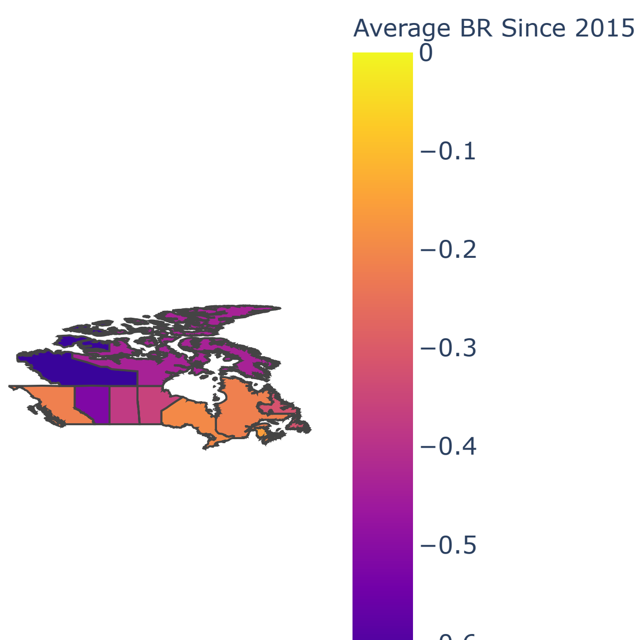
	Province	2000	2001	2002	2003	2004	2005	2006	2007	2008	...	2013	2014	2015	2016	2
1	Newfoundland and Labrador	9.2	9	9	8.9	8.7	8.8	8.9	8.9	9.6	...	8.6	8.7	8.5	8.4	
2	Prince Edward Island	10.6	10.1	9.7	10.3	10.1	9.7	10.2	10.1	10.7	...	9.8	9.8	9.3	9.5	
3	Nova Scotia	9.8	9.6	9.3	9.2	9.3	9.1	9	9.5	9.8	...	9	9.2	8.8	8.8	
4	New Brunswick	9.8	9.6	9.4	9.5	9.3	9.2	9.4	9.6	9.9	...	9.2	9.2	8.7	8.7	
5	Quebec	9.8	10	9.7	9.9	9.8	10.1	10.7	11	11.3	...	11	10.8	10.6	10.5	1

5 rows × 24 columns

In [12]:
```
#Using plotly.express.choropleth to map out the Canadian provinces with averag
e live birth rate.

fig4=px.choropleth(br,locations='id',
geojson=provs,
color= 'Average BR Since 2015', #this is the average live birth rate we calcua
lted from year 2015 to 2020.
hover_name="Province",
range_color=(-0.7, 0), #upper bound set to 0 so all provinces with positive bi
rht rate will be catagorized to same color
color_continuous_scale=px.colors.sequential.Plasma)

fig4.update_geos(fitbounds='locations',visible=False)
fig4.show()
```



Important finding from the analysis in Q2

From the map above, we can see that all canadian provinces are having an average negative birth rate. Negative birth rate means each year we are seeing less and less new borns compared to the previous year.

Combined with the overall declining birth rate trend(line graph fig 3) since year 2000, we can further impressionate provincial governments about the severity of low birth rate in their province, so that every province government can realise the localzied problem of population aging in their jurisdiction, thus be motivated to make changes.

Guiding Question #3

In this quesion, we will use education levels dataset from Government of Canada to find is there relationshp between education level and birth rate. Firstly, we analyze data from Canada as a whole, and then using plotly_express to visulize each province's situation, finially, calculate correlation coefficient and give our conclusion.

Data Source: Educational attainment in the population aged 25 to 64，Government of Canada

Data process manually

- Download education dataset (csv format) from Government of Canada.
- Get birth rate dataset (birth rate.csv) from question one and two, which has already been processed.

Data wrangling through programs

- Import education and birth rate dataset
- Select 10 years data
- Drop some rows which has no meaning

Then

- reconstruct birth rate dataset
- reconstruct education dataset
- merge two above datasets into one table for comparation

In [12]:
```
br=pd.read_csv('Crude_Birth_rate.csv', index_col=[0])
print('\n')
print('raw birth data ')
display(br )

provinceList =  [ 'Canada','Newfoundland and Labrador','Prince Edward Island',
'Nova Scotia','New Brunswick',
                  'Quebec','Ontario','Manitoba','Saskatchewan','Alberta','Briti
sh Columbia',
                  'Yukon','Northwest Territories','Nunavut']
```

```
#br_raw = pd.DataFrame("REF_DATE": "2000","GEO":provinceList "VALUE": )
```

```
raw birth data
```

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	...	2011	2012	2013	20
Canada, place of residence of mother 12															
Canada, place of residence of mother 12	10.7	10.8	10.5	10.6	10.6	10.6	10.9	11.2	11.4	11.3	...	11	11	10.8	10
Newfoundland and Labrador, place of residence of mother	9.2	9	9	8.9	8.7	8.8	8.9	8.9	9.6	9.5	...	8.5	8.3	8.6	8
Prince Edward Island, place of residence of mother	10.6	10.1	9.7	10.3	10.1	9.7	10.2	10.1	10.7	10.4	...	10	9.1	9.8	9
Nova Scotia, place of residence of mother	9.8	9.6	9.3	9.2	9.3	9.1	9	9.5	9.8	9.6	...	9.4	9.3	9	9
New Brunswick, place of residence of mother	9.8	9.6	9.4	9.5	9.3	9.2	9.4	9.6	9.9	9.9	...	9.4	9.3	9.2	9
Quebec, place of residence of mother	9.8	10	9.7	9.9	9.8	10.1	10.7	11	11.3	11.3	...	11.1	11	11	10
Ontario, place of residence of mother	10.9	11.1	10.6	10.7	10.7	10.7	10.7	10.8	10.9	10.8	...	10.6	10.6	10.3	10
Manitoba, place of residence of mother	12.3	12.2	12	12	11.8	12	12.3	12.9	12.9	13.2	...	12.7	13.1	13	1
Saskatchewan, place of residence of mother	12.1	12.3	11.8	12.1	12	12	12.4	13.2	13.5	13.8	...	13.4	13.7	13.5	1
Alberta, place of residence of mother	12.3	12.3	12.4	12.7	12.6	12.7	13.2	14	14.1	14.1	...	13.5	13.6	13.4	13
British Columbia, place of residence of mother	10.1	10	9.8	9.8	9.7	9.7	9.8	10.2	10.2	10.2	...	9.8	9.6	9.5	9
Yukon, place of residence of mother	12.2	11.4	11.2	10.8	11.6	10	11.3	10.9	11.3	11.4	...	12.2	12	10.8	10
Northwest Territories including Nunavut, place of residence of	..	..	..	..	..	..	..	..	..	..	...	..	..	..	

of residence of mother 13															
Northwest Territories, place of residence of mother 14	16.6	15	15.2	16.5	16.1	16.4	15.9	16.7	16.6	16.5	...	15.9	15.8	15.3	15
Nunavut, place of residence of mother	26.4	25.2	25.2	25.9	25	23	24.2	25.3	25.2	26.9	...	24.5	24.3	25.9	2

15 rows × 21 columns

In [13]:

```
#importing dataset of education levels
#Data source from Government of Canada https://open.canada.ca/data/en/dataset/
c9c59a8f-ebe9-4444-a543-63261372c648
# step1 : read and wrangling Educational file
rdata_edu_r = pd.read_csv("./Education-raw.csv")
display("Education Raw Data:",rdata_edu_r.head(3),len(rdata_edu_r))

rdata_edu_r['Population characteristics'] = rdata_edu_r['Population characteri
stics'].str.strip()
rdata_edu_r2 = rdata_edu_r.loc[ (rdata_edu_r['REF_DATE'] > 2009 )
                                & (rdata_edu_r['Educational attainment level']
!= 'Trades')
                                & (rdata_edu_r['Educational attainment level']
!= 'Total, all levels')
                                & (rdata_edu_r['Population characteristics'] ==
'Total population') ]

display("After wrangling Education Data:",rdata_edu_r2.head(3) ,len(rdata_edu_
r2))

# step2 : compare canada's birth rate with education level
# define some common variables
fig       = plt.figure(figsize=(10,5))
ax1       = fig.add_subplot(121)
ax2       = fig.add_subplot(122)
colPar    = ['c','g','b','y']     # line color
eduLevel  = ['Less than high school','High school','College','University'] # e
ducation levels
rdata_edu = rdata_edu_r2

# plot canada's education
axnum = 0
for i,j in zip(colPar,eduLevel):
    rdata_edu2  = rdata_edu.loc[(rdata_edu['GEO'] == 'Canada') & (rdata_edu['E
ducational attainment level'] == j)  &
                                 (rdata_edu['REF_DATE'] > 2009 )& (rdata_edu['R
EF_DATE'] <2020)]
    xdata     = rdata_edu2['REF_DATE']
    ydata     = rdata_edu2['VALUE']
    if axnum < 2:
        ax1.plot(xdata,ydata,color=i, marker='+',label=j)
    else:
        ax2.plot(xdata,ydata,color=i, marker='+',label=j)
    axnum += 1

# plot canada's birthrate
rdata_br  = pd.read_csv("./birth rate.csv")
display("Birth Raw Data:",rdata_br.head(5))
```

```
rdata_br2 = rdata_br.loc[(rdata_br['GEO'].str.contains('Canada')) &
                         (rdata_br['Characteristics'] == 'Total fertility rate
per 1,000 females')&
                         (rdata_br['REF_DATE'] > 2009 )& (rdata_br['REF_DATE']
<2020)]
birthnum = rdata_br2['VALUE']/100  # adjust scale of Y-axis for birthrate
ax2.plot(xdata,birthnum,color='r', marker='*',label="Birth Rate")
ax2.legend(loc='best')
ax1.plot(xdata,birthnum,color='r', marker='*',label="Birth Rate")
ax1.legend(loc='best')
print('\n')
```

'Education Raw Data:'

	REF_DATE	GEO	DGUID	Educational attainment level	Population characteristics	UOM	UOM_ID	SCALAR_FACTOI
0	2007	Canada	2016A000011124	Total, all levels	Total population	Percent	239	unit
1	2007	Canada	2016A000011124	Total, all levels	Off-reserve Indigenous population	Percent	239	unit
2	2007	Canada	2016A000011124	Total, all levels	Non-Indigenous population	Percent	239	unit

3528

'After wrangling Education Data:'

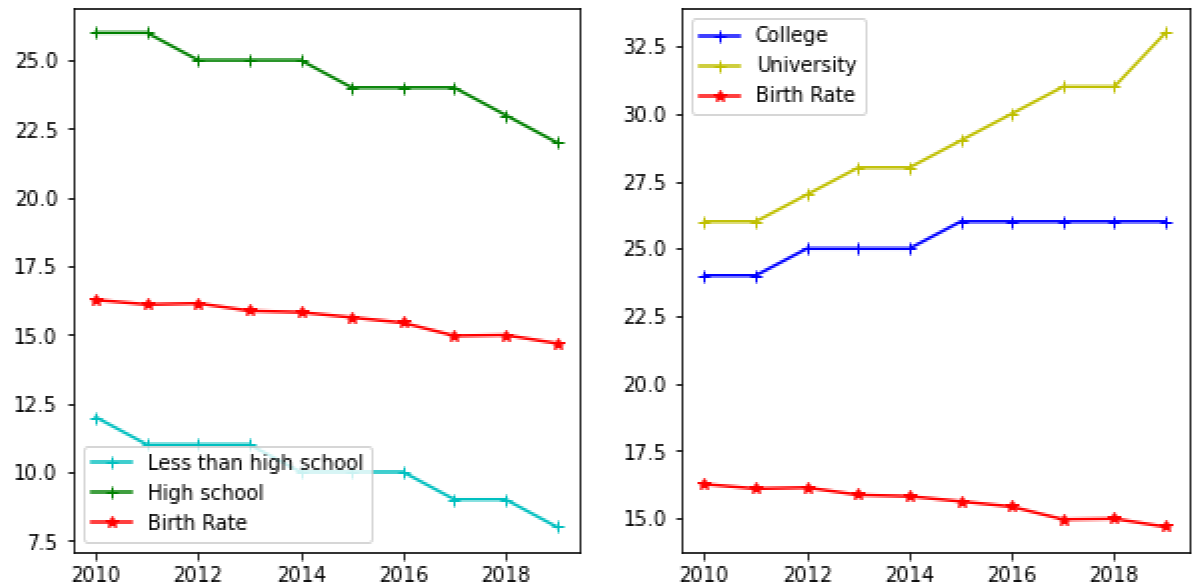
	REF_DATE	GEO	DGUID	Educational attainment level	Population characteristics	UOM	UOM_ID	SCALAR_FACT
759	2010	Canada	2016A000011124	Less than high school	Total population	Percent	239	u
762	2010	Canada	2016A000011124	High school	Total population	Percent	239	u
768	2010	Canada	2016A000011124	College	Total population	Percent	239	u

616

'Birth Raw Data:'

	REF_DATE	GEO	DGUID	Characteristics	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID
0	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 15 to 19 ...	Rate	257	units	(
1	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 20 to 24 ...	Rate	257	units	(
2	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 25 to 29 ...	Rate	257	units	(
		Canada		Age-specific				

3	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 30 to 34 ...	Rate	257	units	(
4	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 35 to 39 ...	Rate	257	units	(



From the picture above about Canada's birth rate and education level, we can observe an **inverse relationship of higher-education and birth rate.** From the figure on the right, we can see birth rate is getting lower with higher percentage of higher-education level, which is again confirmed by the figure on the left, where birth rate dips with lower-education dips(aka more higher education)

But we are currently only using data from Canada as a whole. In order to verify whether this trend is correct, we need to compare from each province To get the correlation, we pick "university" education level as variable against birth rate.

In [14]:

```
# step3 : compare each province's birth data with higher education level
# reconstruct birthrate dataset, only keep date,geo and value,and add new column Name = 'Birth Level'
display("Raw Birth Dataset:",rdata_br.head(3))
rdata_br3 = rdata_br.loc[ (rdata_br['Characteristics'] == 'Total fertility rate per 1,000 females')
                        & (rdata_br['GEO'] != 'Northwest Territories including Nunavut')  #this is not a standard province, discard it
                        & (rdata_br['REF_DATE'] > 2009 )& (rdata_br['REF_DATE'] <2020)]   #get all provinces' birth rate dataset
rdata_br3['GEO'] = rdata_br3['GEO'].str.replace(str(', place of residence of mother'), '')  #modify raw data GEO column
rdata_br4 = rdata_br3.loc[:,['REF_DATE','GEO','VALUE' ]]
rdata_br4['Name'] = 'Birth Level'
rdata_br4['VALUE'] = rdata_br4['VALUE']/100
rdata_br5 = rdata_br4.loc[ (rdata_br4['GEO'] != 'Northwest Territories including Nunavut')]
display("Constructed Birth Data:",rdata_br5.head(3))

# reconstruct  education dataset, only keep date,geo and value ,and add new column Name = 'university'
```

```
rdata_edu3  = rdata_edu.loc[(rdata_edu['Educational attainment level'] == 'University')
                              &  (rdata_edu['REF_DATE'] > 2009 )& (rdata_edu['REF_DATE'] <2020)]
rdata_edu4 = rdata_edu3.loc[:,['REF_DATE','GEO','VALUE' ]]
rdata_edu4['Name'] = 'University'
display("Constructed Education Data:",rdata_edu4.head(3))

# combine new dataset
rdata_combine = rdata_br5.append(rdata_edu4)
display("Combined Dataset(Education and Birth) :",rdata_combine.head(3) )

# using ploy express lib to draw scatter graphs,  to compare education with birth rate, for each province through 10 years
fig = px.scatter(rdata_combine,x ="REF_DATE",y ="VALUE",animation_frame = "GEO" ,color = "Name",width=800, height=400,
        title = 'Compare Birth with Education, base on  all provinces and 10 years scope'  )

fig.layout.updatemenus[0].buttons[0].args[1]["frame"]["duration"] = 500    # control animation speed
fig.show()

# step4: calculate correlation base on Canada as a whole
X = rdata_br2['VALUE']
Y = rdata_edu2['VALUE']
result = np.corrcoef(X, Y)
print("The correlation value is {}, that is, Higher-education and Birth Rate in Canada has are strongly negatively correlated\n".format(result[0,1]))
```

'Raw Birth Dataset:'

	REF_DATE	GEO	DGUID	Characteristics	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID
0	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 15 to 19 ...	Rate	257	units	(
1	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 20 to 24 ...	Rate	257	units	(
2	2000	Canada, place of residence of mother	2016A000011124	Age-specific fertility rate, females 25 to 29 ...	Rate	257	units	(

'Constructed Birth Data:'

	REF_DATE	GEO	VALUE	Name
1358	2010	Canada	16.269	Birth Level
1367	2010	Newfoundland and Labrador	15.836	Birth Level
1376	2010	Prince Edward Island	16.169	Birth Level

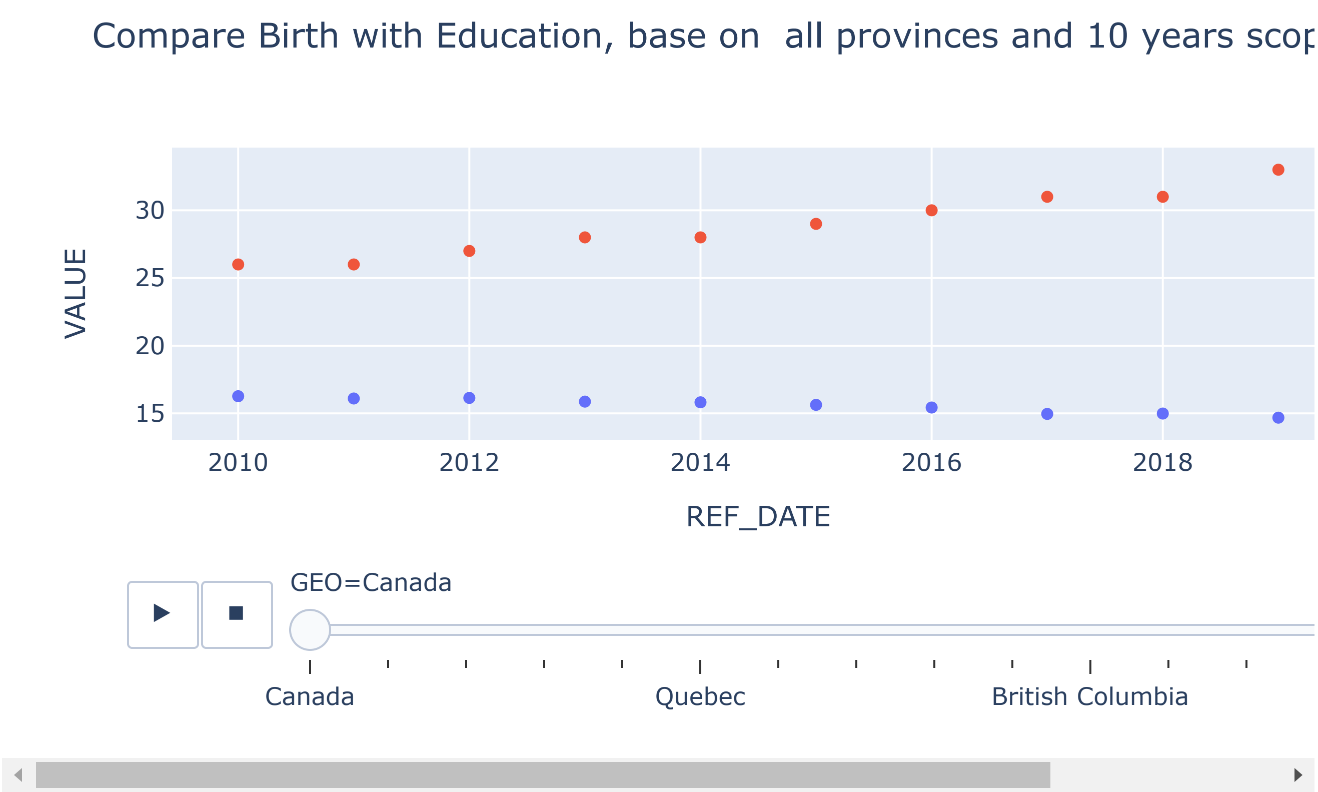
'Constructed Education Data:'

	REF_DATE	GEO	VALUE	Name

	REF_DATE	GEO	VALUE	Name
771	2010	Canada	26.0	University
789	2010	Newfoundland and Labrador	15.0	University
807	2010	Prince Edward Island	22.0	University

```
'Combined Dataset(Education and Birth) :'
```

	REF_DATE	GEO	VALUE	Name
1358	2010	Canada	16.269	Birth Level
1367	2010	Newfoundland and Labrador	15.836	Birth Level
1376	2010	Prince Edward Island	16.169	Birth Level




```
The correlation value is -0.9829818469114102, that is, Higher-education and
Birth Rate in Canada has are strongly negatively correlated
```

Importance of findings from Guiding Question #3

Base on data from the past 10 years from all provinces, we can see all provinces share the same correlation. The calculated correlation between birth rate and higher education level('University') is -0.98. So, we can conclude that higher education and birth rate are closely and negatively correlated. Which means higher education is strongly associated with lower birth rate.

Guiding Question #4

In this quesion, we will use GDP,crime two datasets from statistics canada, to find whether they could affect birth rate. Firstly, we use matplotlib.pyplot to visulize each province's comparation situation, and then calculate correlation coefficient for each province, finally we will use chart to support our conclusion.

Data Source1: Gross domestic product (GDP) at basic prices，Statistics Canada

- **Data process manually**
  - Download GDP dataset from Statistics Canada, the parameters selected online are : the period(2000-2020), value(chained(2012)dollars) and industry parameters(All industries) online
- **Data wrangling through programs**
  - Import GDP dataset
  - Discard all non-data from dataframe
  - Convert string number to float number
  - reconstruct GDP dataset

Data Source2: Crime Severity Index，Statistics Canada

- **Data process manually**
  - "Data table for Chart 11","Data table for Chart 12","Data table for Chart 13" are dataset we want, but they are data list on page
  - Copy them into csv file directly
- **Data wrangling through programs**
  - Import Crime dataset
  - Slice to get data for comparison
  - reconstruct Crime dataset

In [15]:
```
#importing dataset of GDP
#Data source from Statistics Canada. Accessesible at :https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610040201
#importing dataset of crime
#Data source from Statistics Canada. Accessesible at :https://www150.statcan.gc.ca/n1/pub/85-002-x/2021001/article/00013-eng.htm
# step1 : read GDP file
rdata_gdp_r  = pd.read_csv("./GDP-raw.csv" ,header = 1,delimiter="\t",encoding= "ISO-8859-1"  )
display("GDP Raw Data:",rdata_gdp_r.head(3))     # the orignal file header has lots notes , so the following will show NaN mostly

col_name = ['Geography','2000','2001','2002','2003','2004','2005','2006','2007','2008','2009','2010',
                        '2011','2012','2013','2014','2015','2016','2017','2018','2019','2020']
rdata_gdp_r2 = rdata_gdp_r[10:23]
rdata_gdp_r2.columns = col_name   # change column name

# change  all column from string to number
i = 1
while i < 21 :
  rdata_gdp_r2[col_name[i]] = rdata_gdp_r2[col_name[i]].str.replace(',', '').astype(float)
  i += 1

rdata_gdp =  rdata_gdp_r2
display("After wrangling GDP Data:",rdata_gdp.head(3),len(rdata_gdp) )

#read Crime file
rdata_crime  = pd.read_csv("./crime.csv"  )
display("crime Raw Data:",rdata_crime.head(5))

# step2 : use matplotlib plot  to compare GDP , crime and birth rate for each province among 10 years, in seperate chart
def Plotbypro(ProName,Birthrate ):
    xdata  =  np.array([  2010 , 2011 , 2012 , 2013 , 2014 , 2015 , 2016 , 2017 , 2018 , 2019 ])
```

```python
    # 10 years GDP data for one province ProName
    rdata_gdp_p1 = rdata_gdp.loc[rdata_gdp['Geography'] == ProName]
    ydata_gdp    =  np.array([rdata_gdp_p1['2010'],rdata_gdp_p1['2011'],rdata_
gdp_p1['2012'],rdata_gdp_p1['2013'],
                        rdata_gdp_p1['2014'],rdata_gdp_p1['2015'],rdata_gdp_p1
['2016'],rdata_gdp_p1['2017'],
                        rdata_gdp_p1['2018'],rdata_gdp_p1['2019']])

    # 10 years crime data for one province ProName
    ydata_crime     =  np.array(rdata_crime[ProName][12:22]*300).tolist()

    ax1.plot(xdata,ydata_gdp,  color='g', marker='+',label='GDP')
    ax1.plot(xdata,ydata_crime,color='b', marker='+',label='Crime')
    ax1.plot(xdata,Birthrate*3000,color='r', marker='*',label="birth rate")
    plt.title(ProName)
    plt.xticks(rotation = 60)
    plt.legend(loc='best')

    #calculate correlation between crime,gdp with birth rate for each province
    ydata_gdp_l = np.transpose(ydata_gdp)
    corgdp  = np.corrcoef(Birthrate, ydata_gdp_l)
    corcrime = np.corrcoef(Birthrate, ydata_crime)

    return (corgdp[0,1],corcrime[0,1])

# call function and use loop to draw all charts for 10 provinces
fig = plt.figure(figsize=(20,16))
j = 1
provinceList =  [ 'Newfoundland and Labrador','Prince Edward Island','Nova Sco
tia',
                 'New Brunswick','Quebec','Ontario','Manitoba','Saskatchewan',
                 'Alberta','British Columbia']
                 # discarded Yukon, 'Northwest Territories', 'Nunavut' because
less data
cor_gdp   = []*10
cor_crime = []*10
for i in provinceList:
    ax1 = fig.add_subplot(3,4,0+j)
    br_onepro  = rdata_br5.loc[rdata_br5['GEO']==i]
    br_onepro2 = np.array(br_onepro['VALUE'])
    gdp,crime = Plotbypro(i,br_onepro2)
    cor_gdp.append(gdp)
    cor_crime.append(crime)
    j += 1
```

'GDP Raw Data:'

	Frequency: Annual	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	U
0	Table: 36-10-0402-01 (formerly CANSIM 379-0030)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	Release date: 2021-05-03	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	Geography: Province or territory	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

3 rows × 22 columns

'After wrangling GDP Data:'

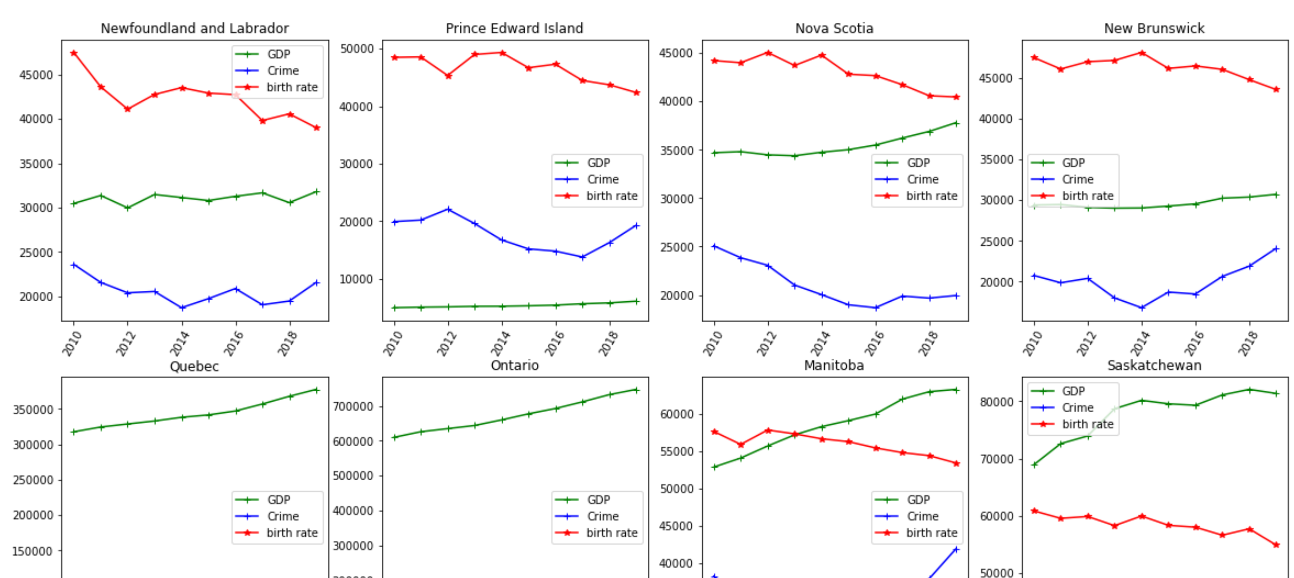
	Geography	2000	2001	2002	2003	2004	2005	2006	2007	2008	...
10	Newfoundland and Labrador	21945.8	22429.7	26019.3	27753.1	27485.0	28185.9	29220.9	32536.6	31947.8	...
11	Prince Edward Island	4191.4	4139.5	4340.2	4415.8	4545.9	4690.6	4804.3	4781.6	4836.1	...
12	Nova Scotia	29571.6	30377.3	31608.4	32018.4	32360.0	32761.9	32920.8	33235.4	33797.3	...

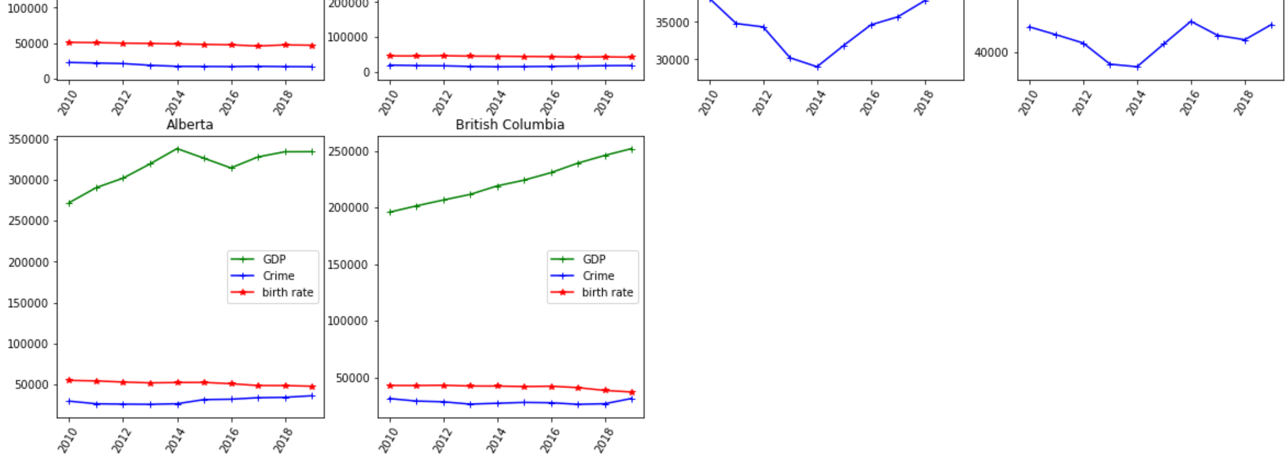
3 rows × 22 columns

13

'crime Raw Data:'

	Year	Quebec	Ontario	British Columbia	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Manitoba	Saskatchew
0	1998	112.7	100.7	166.9	76.4	73.3	105.4	90.0	154.5	176
1	1999	104.3	92.3	155.8	69.2	79.0	104.6	90.0	152.6	167
2	2000	101.8	89.0	144.7	70.1	76.3	95.3	84.8	149.5	169
3	2001	96.6	86.5	146.6	69.1	75.4	92.5	83.4	152.5	176
4	2002	93.5	84.5	148.1	71.4	85.2	93.9	84.6	148.3	175

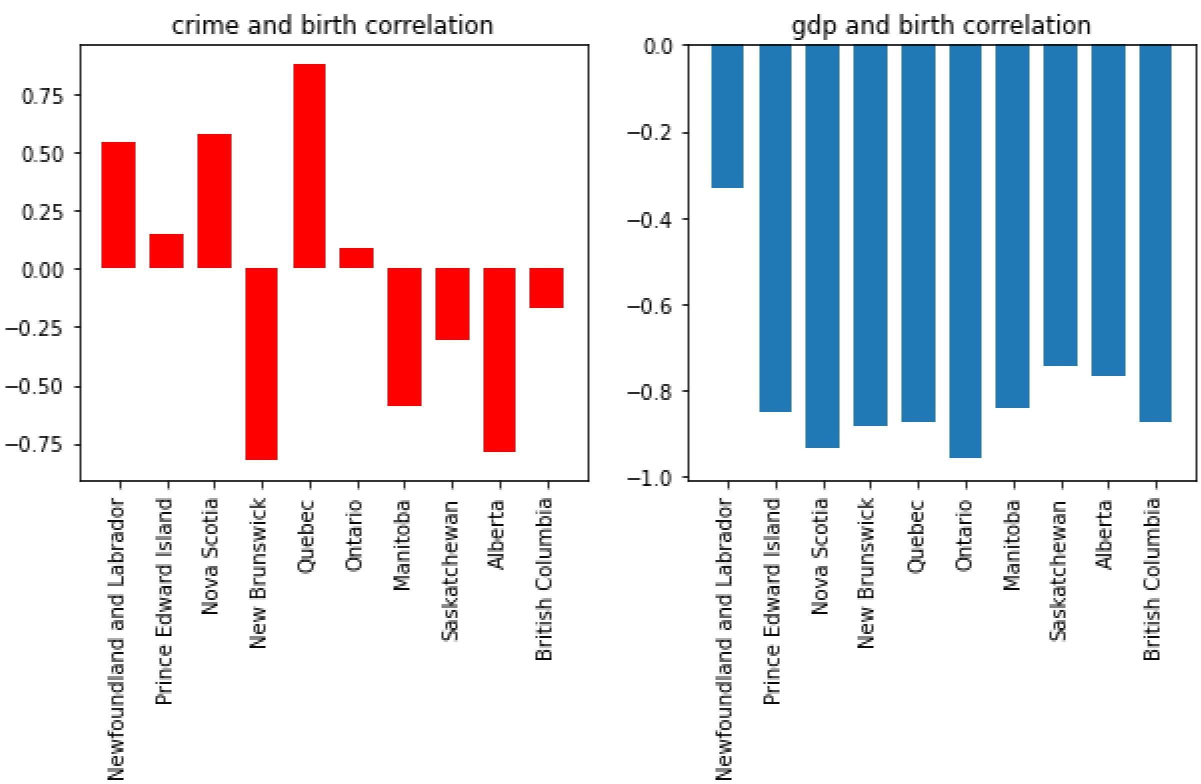




Base on data from the past 10 years from all provinces, we can see there is **matching trend between GDP and birth rate, but we can't observe any pattern between birth rate and crime rite.**

In order to further prove this observation, we calulate and plot the correlation between crime, gdp and birth rate below.

In [16]:
```
# step3 : plot correlation for each province
fig = plt.figure(figsize=(10,4))
ax1 = fig.add_subplot(1,2,1)
plt.bar(provinceList, cor_crime,width = 0.7,color='r')
plt.title("crime and birth correlation")
plt.xticks(rotation=90)
ax1 = fig.add_subplot(1,2,2)
plt.bar(provinceList, cor_gdp,width = 0.7)
plt.title("gdp and birth correlation")
plt.xticks(rotation=90)
plt.show()
```



Importance of findings from Q4

At this point, we can conclude that there is no relationship between crime rate and birth rate from the cross-province correlation analysis because we don't observe a uniform pattern.

However, we can see that birth rate and gdp has opposite correlation. Especially in Ontario, Nova scotianowe and British Columbia, where the correlation has breached -0.8, showing strong negative correlation between GDP and birth rate.

So, we can conclude that, GDP and Birth rate are negatively correlated, and there is no relationship between crime and birth rate.

Guiding Question #5

In this quesion, we will use unemployment dataset from Statistics Canada to find is there relationshp between unemployment and birth rate. Firstly, we will use sklearn.linear_model to calculate R-square value for each province, and then visulizing the result as well as describing our analyzing result.

R-squared (R2), is a statistical measure that explains what extent the variance of one variable(independent) explains the variance of the second variable(dependent) . R-squared value means: [12]

- if R-squared value < 0.3 this value is generally considered a None or Very weak effect ,
- if R-squared value 0.3 < r < 0.5 this value is generally considered a weak or low effect ,
- if R-squared value 0.5 < r < 0.7 this value is generally considered a Moderate effect ,
- if R-squared value r > 0.7 this value is generally considered strong effect .

Data Source: Unemployment rate，Statistics Canada

Data process manually

- Download dataset from Government of Canada.The parameter selected online are: Reference period(2000 to 2020),Age group(25 years and over)

Data wrangling through programs

- Import unemployment dataset
- Slice to discard some columns and rows
- Rename columns

Then

- Reconstruct unemployment dataset,keep necessary columns for comparation

In [17]:

```
#importing dataset of unemployment
#Data source from Statistics Canada. Accessesible at :https://www150.statcan.g
c.ca/t1/tbl1/en/tv.action?pid=1410032702
# step1 : read  file
f = open(r'./unemployment-raw.csv','r')
reader = csv.reader(f)
rdata_un_r = pd.DataFrame(reader,dtype=str)
display("unemployment Raw Data:",rdata_un_r.head(5))    # orignal file has lots
of non data rows

rdata_un_r2 = rdata_un_r.loc[12:45,0:22]                # slice column and row
col_name = ['Geography 3','Labour force characteristics',
           '2000','2001','2002','2003','2004','2005','2006','2007','2008','20
09','2010',
```

```
          '2011','2012','2013','2014','2015','2016','2017','2018','2019','20
20']
rdata_un_r2.columns = col_name                          # rename sliced dataset
rdata_un = rdata_un_r2
rdata_un2 = rdata_un.loc[rdata_un['Labour force characteristics'] == 'Unemploy
ment rate 4']
display("After Wrangling unemployment Data:",rdata_un.head(5) )


#get main provinces' birth rate dataset  , from 2000 to 2019
rdata_br  = pd.read_csv("./birth rate.csv")
rdata_br['GEO'] = rdata_br['GEO'].str.replace(str(', place of residence of mot
her'), '')
rdata_br6  = rdata_br.loc[ (rdata_br['Characteristics'] == 'Total fertility ra
te per 1,000 females')]
rdata_br7  = rdata_br6.loc[ (rdata_br6['GEO'] != 'Northwest Territories includ
ing Nunavut')
                          & (rdata_br6['GEO'] != 'Nunavut') & (rdata_br6['GEO'
] != 'Northwest Territories')
                          & (rdata_br6['GEO'] != 'Yukon')  ] # unemployment da
taset has no these provinces
display("Modified birth rate Data :",rdata_br7.head(3) )

# step2 : define function, use sklearn package to calculate linear regression
r Square
def getRSQ(ProName) :
    #get birth rate for one province
    x_br  = np.array(rdata_br6.loc[rdata_br6['GEO'] ==  ProName ]['VALUE']/100
).tolist()

    #get unemployment data for one province
    rdata_un3 = rdata_un2.loc[rdata_un['Geography 3'] ==  ProName ]
    y_un = []
    ListYear = np.arange(2000,2020 )
    for i in ListYear:
        a = np.array(rdata_un3[str(i)])      # unemployment file is horizontal
 table, read all value from year 2000 to 2020
        #y_un.extend(a)
        y_un.append(float(a))


    #calculate linear regression r square
    x_br2 = np.array(x_br).reshape((-1, 1))
    model = LinearRegression()
    model = LinearRegression().fit(x_br2, y_un)
    r_sq  = model.score(x_br2, y_un)
    return r_sq


#step 3: use look to calculate r-square value for each province
provinceList =  [ 'Newfoundland and Labrador','Prince Edward Island','Nova Sco
tia',
                 'New Brunswick','Quebec','Ontario','Manitoba','Saskatchewan',
                 'Alberta','British Columbia']
sq_un=[]
for i in provinceList :
    r_sq_un = getRSQ(i)
    sq_un.append(r_sq_un)

# step4 :   visulize r square ( independent variable is unemploy , dependent v
ariable is birth rate; for each province and 20 years)
fig   = plt.figure(figsize=(9,5))
```

```
plt.bar( provinceList, sq_un, width=0.9,color=['r', 'g', 'b','y'])
plt.xticks(rotation = 90)
plt.ylim((0,1))
plt.ylabel("R-squared value")
index = np.arange(len(sq_un))
for a,b in zip(index,sq_un):
    plt.text(a, b+0.05, '%.2f'%b, ha='center', va= 'bottom',fontsize=7)
plt.title("R-square (unemployment and birthrate)")
fig.show()
```

'unemployment Raw Data:'

	0	1	2	3	4	5	6	7	8	9	...	55	56
0	"Unemployment rate	participation rate and employment rate by sex	annual 1 2"	None	None	None	None	None	None	None	...	None	None
1	Frequency: Annual	None	None	None	None	None	None	None	None	None	...	None	None
2	Table: 14-10-0327-02	None	None	None	None	None	None	None	None	None	...	None	None
3	Release date: 2021-01-27	None	None	None	None	None	None	None	None	None	...	None	None
4	Geography: Canada, Province or territory	None	None	None	None	None	None	None	None	None	...	None	None

5 rows × 65 columns

'After Wrangling unemployment Data:'

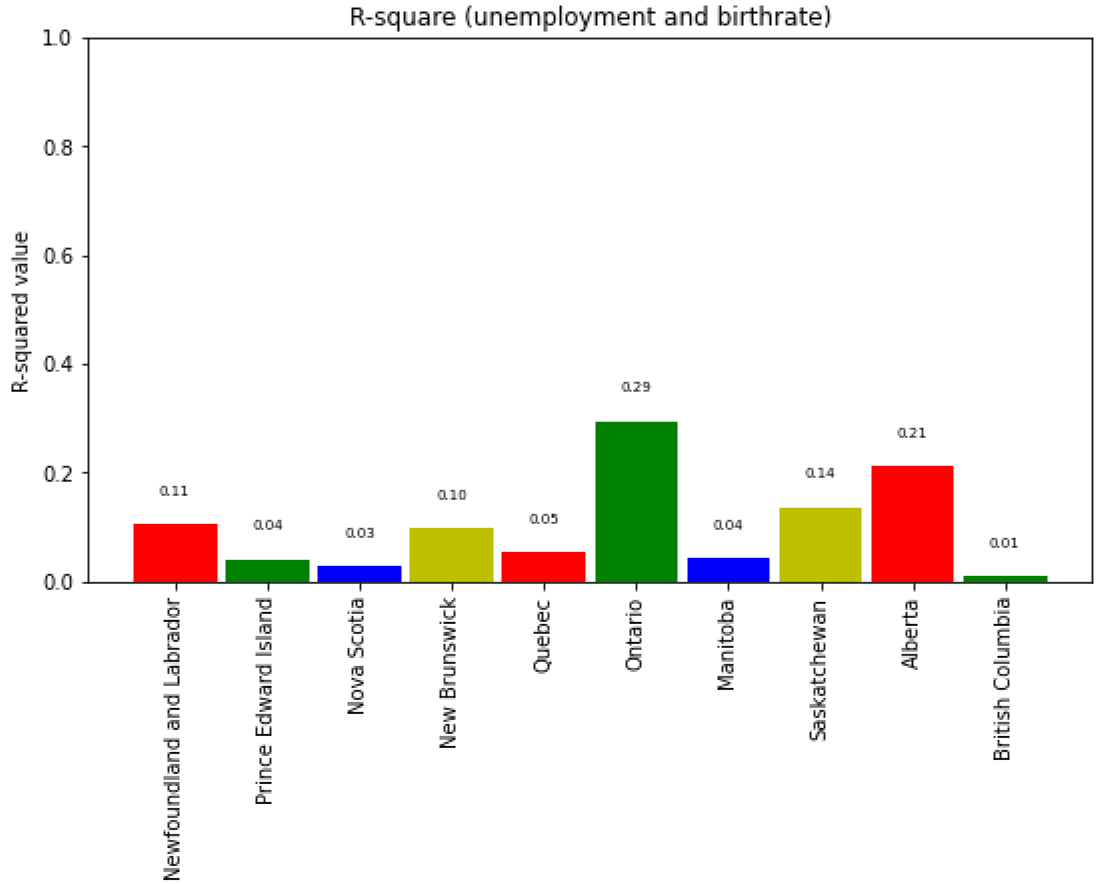
	Geography 3	Labour force characteristics	2000	2001	2002	2003	2004	2005	2006	2007	...	2011	2012	201
12		Unemployment rate 4	5.7	6.1	6.5	6.4	5.9	5.7	5.3	5.1	...	6.4	6.1	6.
13		Participation rate 5	66.0	66.2	66.9	67.6	67.6	67.4	66.9	67.2	...	66.8	66.8	66.
14		Employment rate 6	62.3	62.1	62.6	63.3	63.6	63.6	63.4	63.7	...	62.6	62.7	62.
15	Newfoundland and Labrador	Unemployment rate 4	15.0	14.4	15.6	15.1	14.4	14.1	13.5	12.4	...	11.6	11.5	11.
16		Participation rate 5	57.3	58.2	59.2	60.0	59.9	59.6	60.0	59.9	...	60.2	61.6	61.

5 rows × 23 columns

'Modified birth rate Data :'

	REF_DATE	GEO	DGUID	Characteristics	UOM	UOM_ID	SCALAR_FACTOR	SCAL
8	2000	Canada	2016A000011124	Total fertility rate per 1,000 females	Rate	257	units	
17	2000	Newfoundland and Labrador	2016A000210	Total fertility rate per 1,000 females	Rate	257	units	

26	2000	Prince Edward Island	2016A000211	Total fertility rate per 1,000 females	Rate	257	units



Importance of finding from Q5

From above we can see, the R-square are all less than 0.3, that is, only very little variance of birth rate can be explained by variance of unemployment. We can conclude that <u>employment rate has very weak effect with birth rate </u>. Thus this should be in low-priority regards to improving birth rate.

Conclusion

In this project, we succesfully visualized and analyzed the topic of aging in Canada, especially the underlying social factors that are contributing to the low birth rate as the root problem.

In summary, our conclusions are as following:

1. Canada is subjected to the problem of population aging, and it has exabercated over the years.
2. All provinces in Canada are netting negative birth rate each year, meaning there are less children born each year throught Canada. This points out the importance of imminent changes.

From the analysis of our proposed contributing social factors to a low birth rate, our analysis conclude:

- The main factors affecting the birth rate are higher-education level and GDP per capita. Both higher-education and GDP are strongly and negatively associated with birth rate. Meaning with higher education and more finiancial freedom, people are less likely to have children.
- The unemployment rate has very little impact on birth rate, and is too weak to produce meaningful statistical significance.

- The crime rate has no or negligible contribution to birth rate.

Based on our findings, we hope to inspire government officals to start implementing changes to combat the problem with population aging and low birth rate. Furthermore, future studies can prioritize the focus based on our result, such as elaborating on the specifics on effects of GDP per capita and higher-education level when making changes.

In [ ]:

References

1. Government of Canada a, , Educational attainment in the population aged 25 to 64. Available: https://open.canada.ca/data/en/dataset/c9c59a8f-ebe9-4444-a543-63261372c648 [2020, Dec.12th,].
2. Statistics Canada b, , Unemployment rate, participation rate and employment rate by sex, annual. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410032702 [2021, .09.27].
3. Open Source Project a, , Geopandas. Available: https://geopandas.org/docs.html [2021, .09.27].
4. Open Source Project b, , Shapely. Available: https://pypi.org/project/Shapely/ [2021, .09.27].
5. Statistics Canada a, , Crude birth rate and total fertility rate. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310041801 [2021, .09.27].
6. Statistics Canada b, , Gross domestic product (GDP) at basic prices, by industry, provinces and territories. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610040201 [2021, .09.27].
7. Statistics Canada c, , Incident-based crime statistics. Available: https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=3510017701 [2021, .09.27].
8. Statistics Canada d, , Yearly Canadian population estimation . Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.1&cubeTimeFrame.startYear=1995&cubeTimeFrame.endYear=2020&referencePeriods=19950101%2C20200101.
9. The World Bank a, , Population ages 0-14 - Canada. Available: https://data.worldbank.org/indicator/SP.POP.0014.TO?end=2020&locations=CA&start=1960&view=chart [2021, .09.27].
10. The World Bank b, , Population ages 65 and above - Canada. Available: https://data.worldbank.org/indicator/SP.POP.65UP.TO?end=2020&locations=CA&start=1960&view=chart [2021, .09.27].
11. Statistics Canada c, ,Crime Severity Index: https://www150.statcan.gc.ca/n1/pub/85-002-x/2021001/article/00013-eng.htm [2021,.10.01]
12. Source: Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company. Page (138).
13. Github, , Canada.geojson. Available: https://raw.githubusercontent.com/codeforgermany/click_that_hood/main/public/data/canada.geojson

In [ ]: