Analysis of Economy of 5 developed countries

DATA 501 Project

Mahdi Aziz- 80424161

1. Abstract

Economy is one of the most important parts of the human development index of each country, which includes many factors such as production, distribution, trade, consumption of goods (Prinsipe Ybarro, 2017). In this project, we study the economy of 5 developed countries, focusing on 60 economy indicators such as Gross Domestic Product (GDP) and Consumer prices. In particular, we study the correlation between 60 economy indicators and provide a forecast for the 5 next years using regression model. The dataset we use is World development indicators published by The World Bank ("world-development-indicators," 2017). For the analysis part, we use both Python and Excel to extract useful information from the dataset.

2. Economy Definition and importance

Economy is the cornerstone in the development of any country. According to Investopedia ("economy," n.d.), "Economy is the large set of inter-related production and consumption activities that aids in determining how scarce resources are allocated. This is also known as an economic system".

Here we provided the definition of some important economy indicators.

Net national income: It refers to net national product (NNP) minus indirect taxes. It includes the income of government, households and business ("Net_national_income," n.d.).

GDP: It shows how healthy a country's economy is. It indicates the total dollar value of all goods and services for a country over a specific period of time (Investopedia Staff, 2017).

3. Dataset

The world development indicators dataset is an open dataset published by The World Bank that encompasses more than several hundred indicators of economic development from all the countries around the world. The dataset that we are using is the primary World bank collection that is compiled and collected from officially recognized international resources ("world-development-indicators," 2017). It

covers many topics ranging from agriculture and rural development to poverty and social development. It is being updated regularly and quarterly in April, July, September and December and the last update is April 27, 2017. The dataset can be download from here.

4. Tools and Methodology

For this project, for the cleaning and transforming the data we used Excel program. Because the world-bank Excel file is massive and contains more than 400 thousand rows, we used Excel program to extract 10 main indicators for each countries and 60 indicators as the secondary to perform the analysis. We stored the results on the new sheet for each country and save the results as CSV file. Then, Python was used for analyzing the data and finding the useful information about the data. In fact, we used Python to address the correlation, regression analyses on the data. More specifically, we used Numpy for providing powerful and flexible data structure provided, Scipy for the regression analysis on the dataset, Matplotlib for plotting the results of our correlation and regression analyses and finally csv for working with csv files.

5. Experimental results

In this section, we perform some analyses on the economic indicators such as correlation and regression analyses. But, first we need to justify why we used some indicators and why we applied them on these countries.

5.1. About the Dataset

The world bank file contains 1504 indicators for every country and the data were gathered from 1960 to 2016. Here, we explained why we chose the indicators and countries.

5.1.1. Selected Indicators

In this project, we study the economy of 5 developed countries based on 60 economy-related indicators. Among which are fertility rate, unemployment rate, birth rate, consumer price index, death rate, Gross Domestic Product (GDP), GINI index, inflation rate, rural population, unemployment rate. Other 50 indicators can be found in Hsheet.csv or WorldIndicators.xlsx in Hsheet sheet. It should be noted that the indicators were gathered from 1960 till 2016 (inclusive).

5.1.2. Selected Countries

The reason that we selected Canada is that it would be more interesting to know about the country that we live in. And for other four countries chosen, we chose

them because they are somehow similar in terms of economic development to Canada. The countries we chose are China, Germany, Japan and United States.

5.2. Simple Comparison

5.2.1. Inflation rate

Using python, in this section we compare the inflation rate of 5 countries. Figure 1 shows the inflation rate of five developed countries from 1960 to 2016.

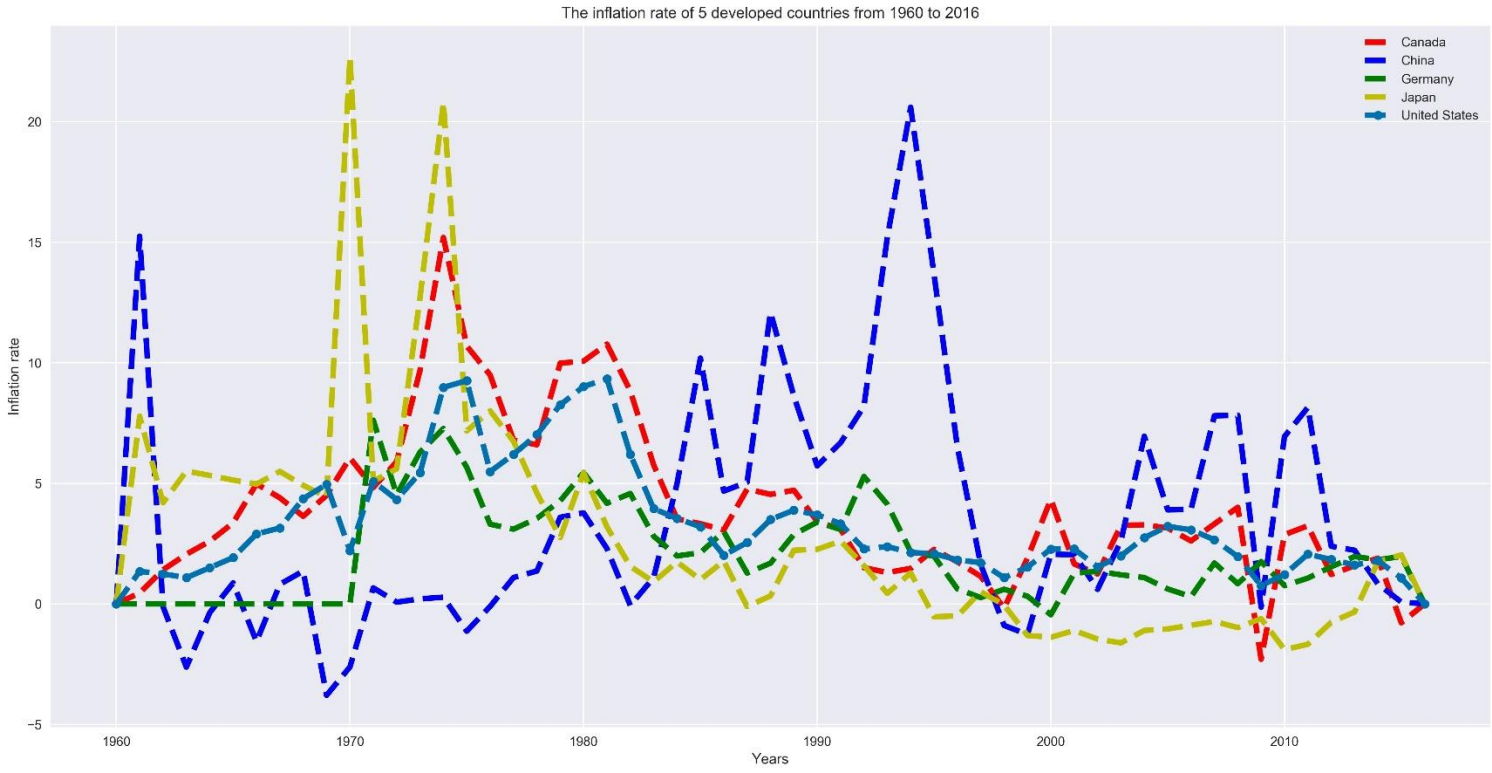


Figure 1- Inflation rate of 5 developed countries

As you can see from Figure 1, the inflation rate of United States is higher than all other countries, specifically in 1960 and 1992 to 1995.

5.2.2. Unemployment rate

Here, we study the unemployment rate of 5 developed countries. Figure 2 shows the unemployment rate of five developed countries from 1960 to 2016.

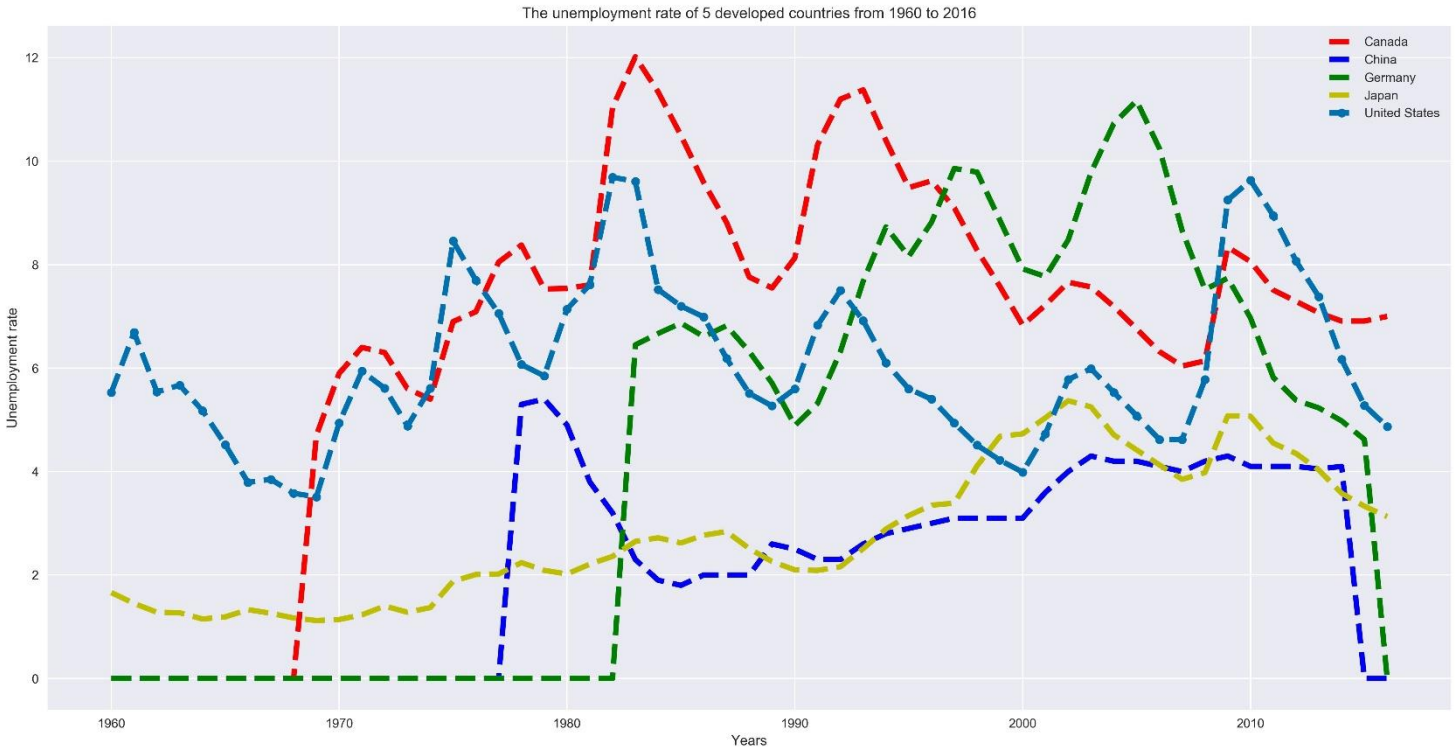


Figure 2- Unemployment rate of 5 developed countries

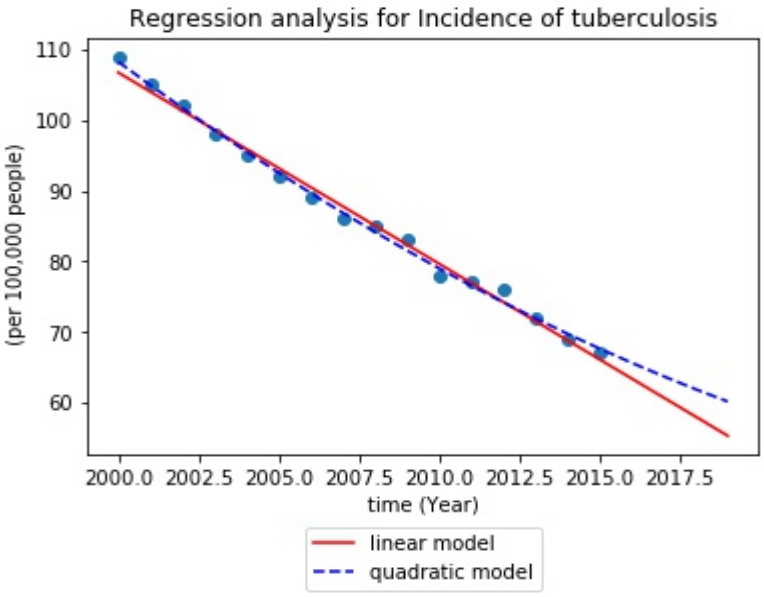
As you can see from Figure 2, the unemployment rate of Canada is the highest among these five countries, specifically from 1970 to 2000 where it is from 4 percent to 12 percent.

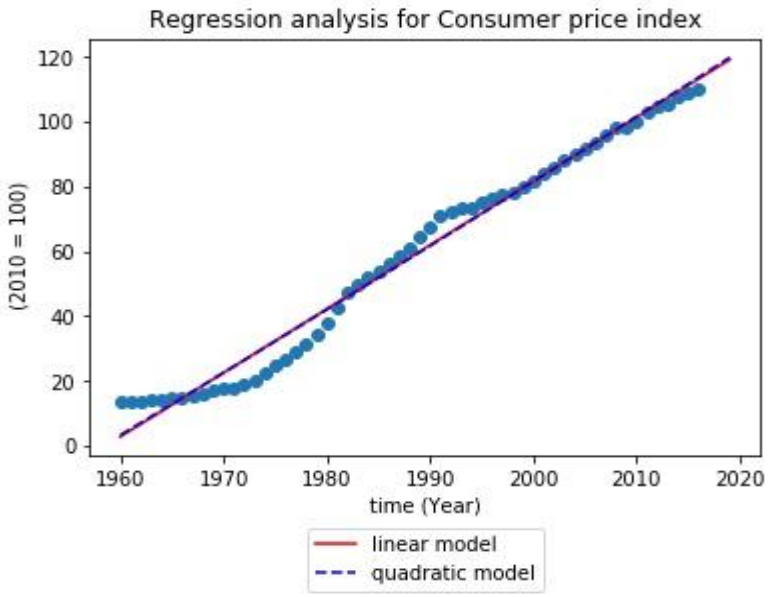
One important weakness of comparing the 5 developed countries in this manner is that for some years like the one we see in Figure 2, we do not have information for Germany before 1992. Similarly, for 2016, we only have information for United states and Canada. So, this may cause confusion in analysing the results.

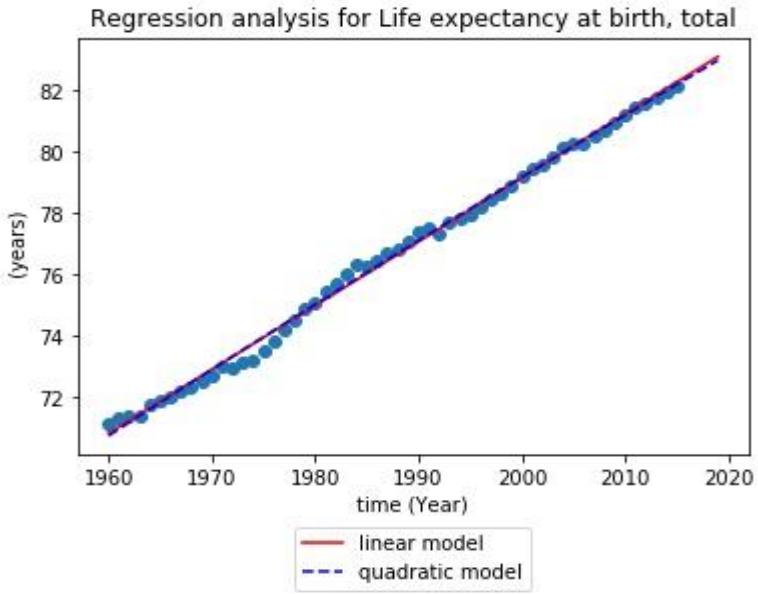
5.3. Regression analysis

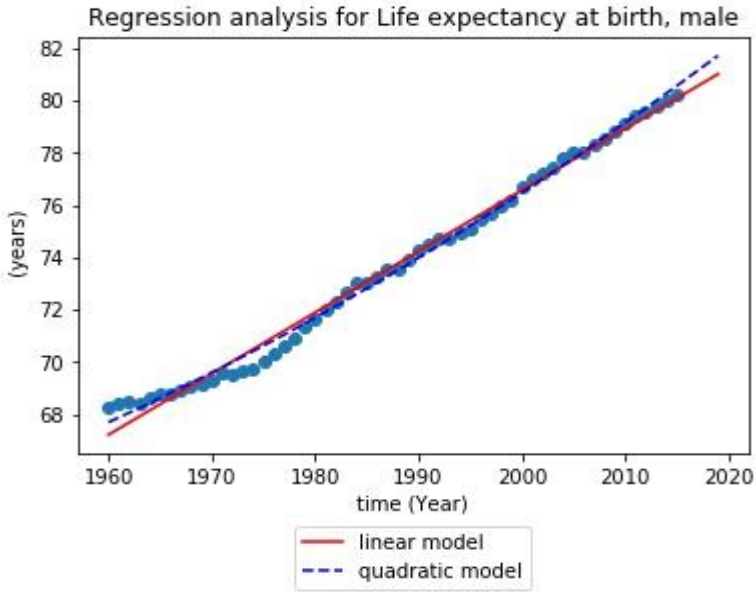
Using the regression analysis, we can forecast the future behaviour of the indicator that we study. Moreover, we can estimate the relationship between values for each indicator. In this analysis, we performed the linear and quadratic regression on all 60 economic-related indicators to forecast the behaviour of that indicator for the next 4 years, meaning till 2020. Figure 3,4,5,6,7 show the results of this analysis for some of these indicators including birth rate, GDP, Tax revenue and wage and salaries for Canada, China, Germany and Japan.

At first, we show that there are some indicators that linear models can model the sample points and can forecast the future.









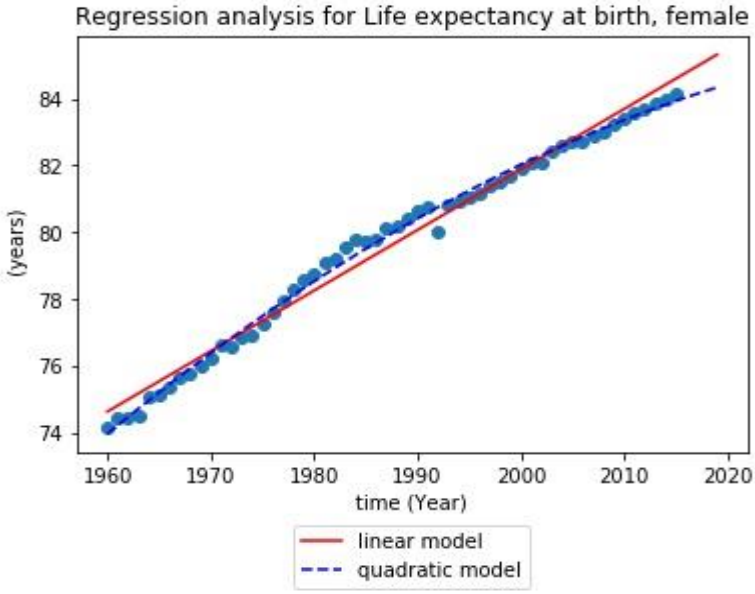
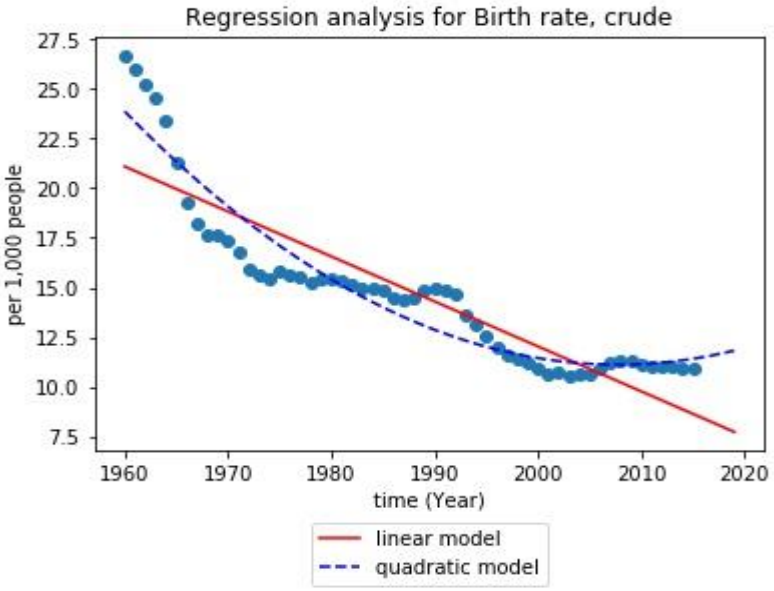
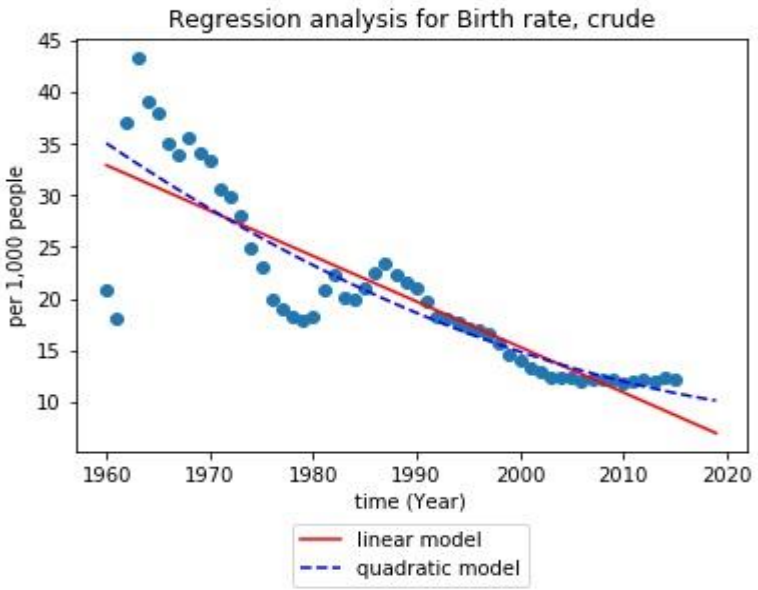


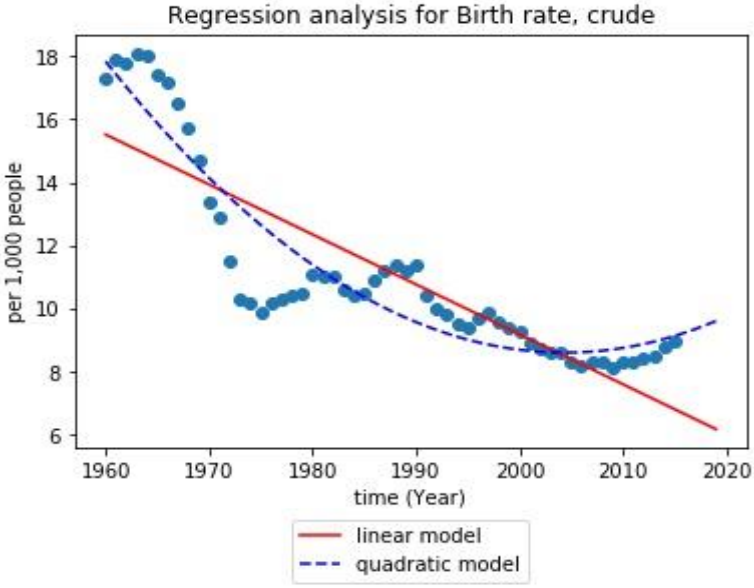
Figure 3- Some indicators that Linear model works perfect



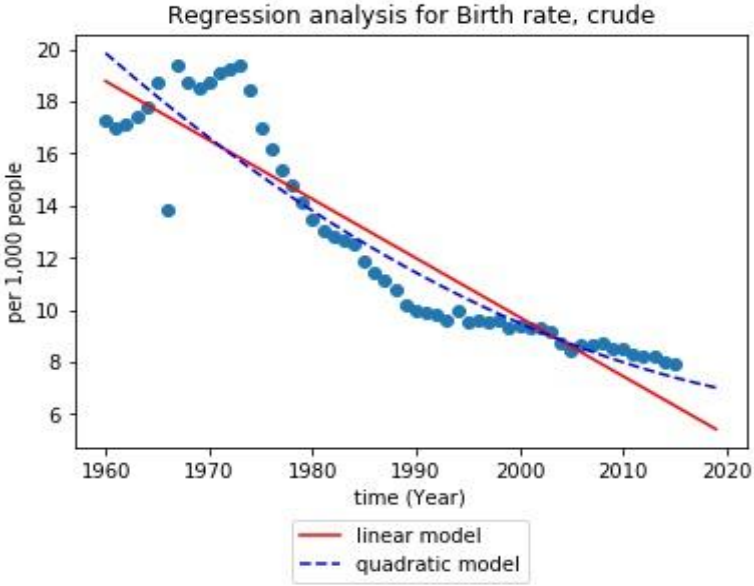
a)Canada



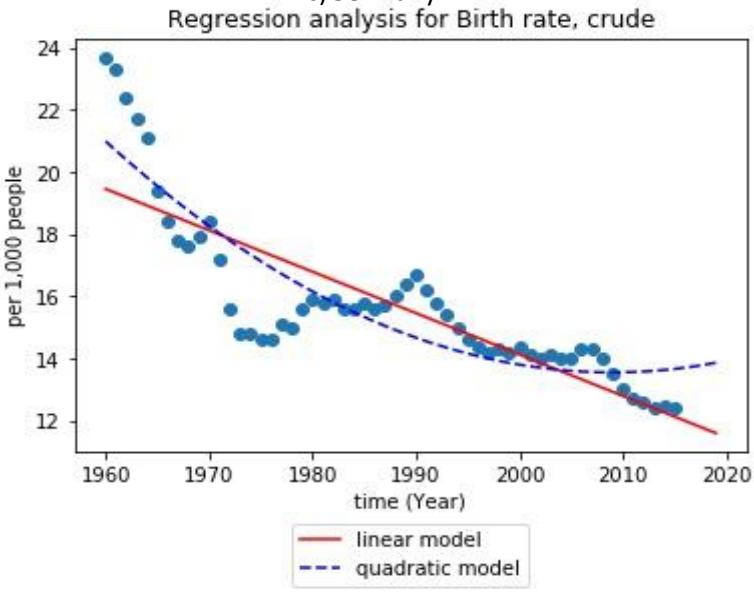
b)China



c)Germany

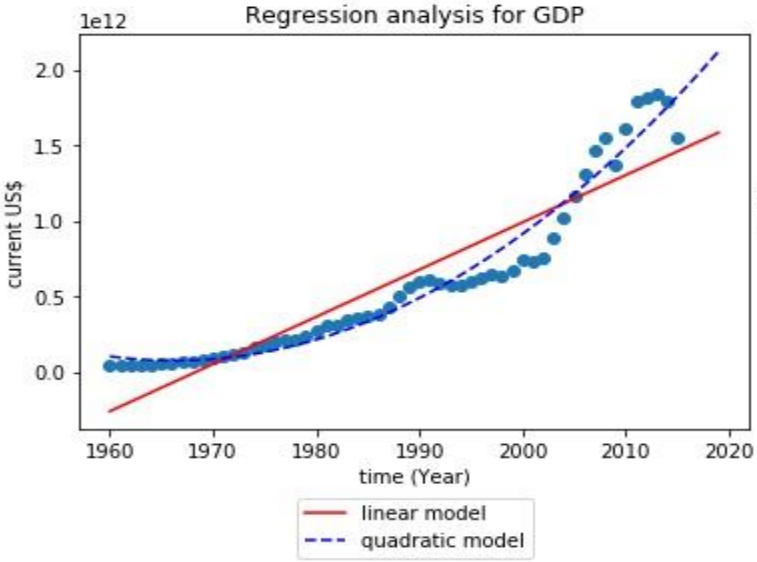


d)Japan

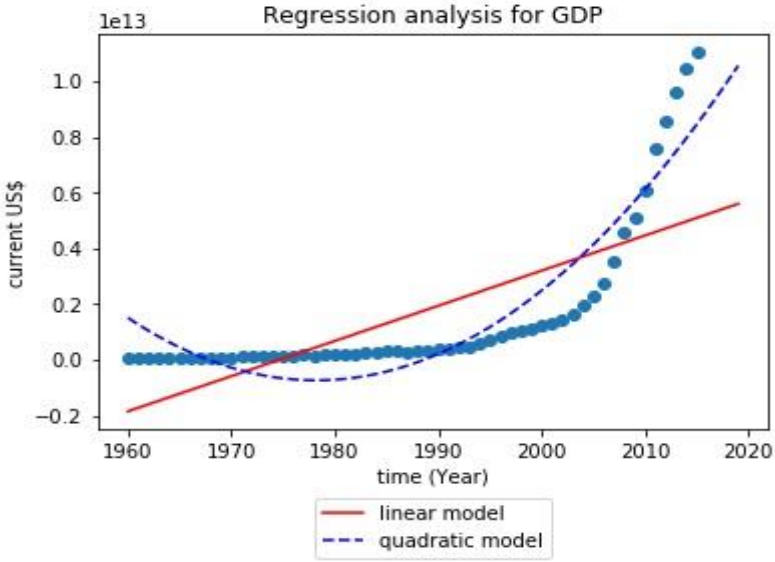


d)United States

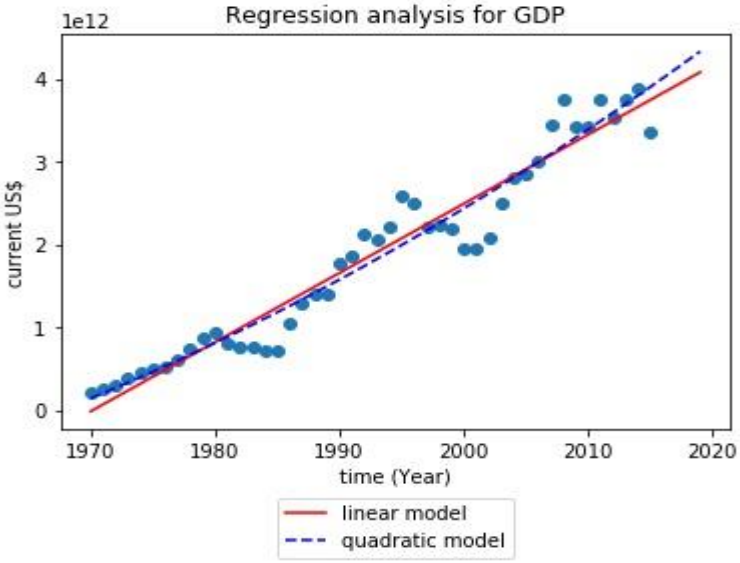
Figure 4- Regression analysis of Birth rate for five countries



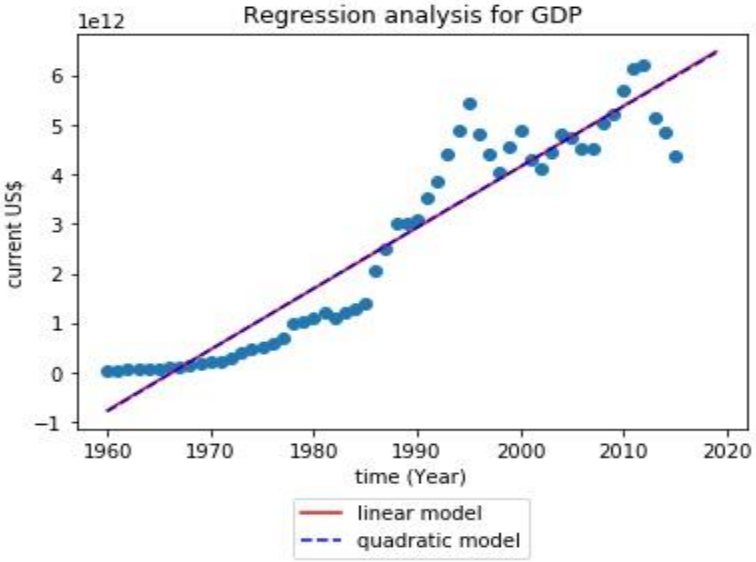
a)Canada



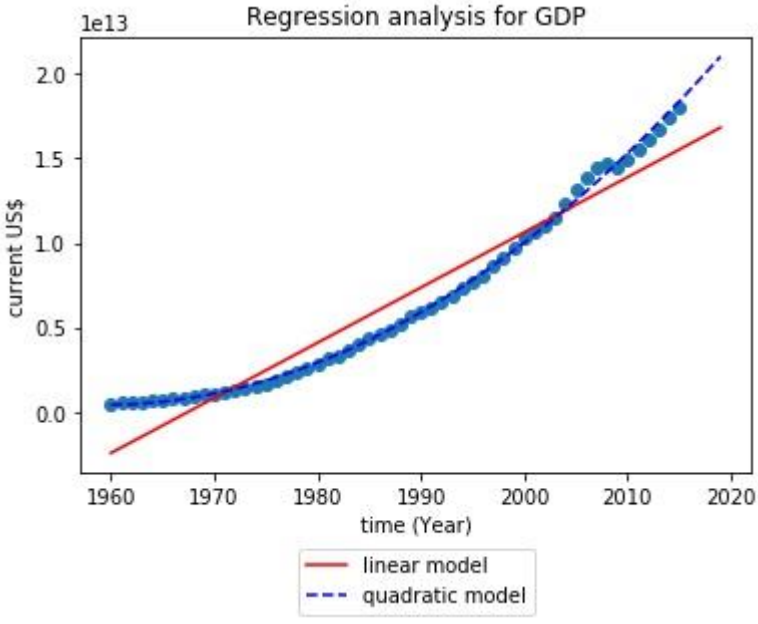
b)China



c)Germany

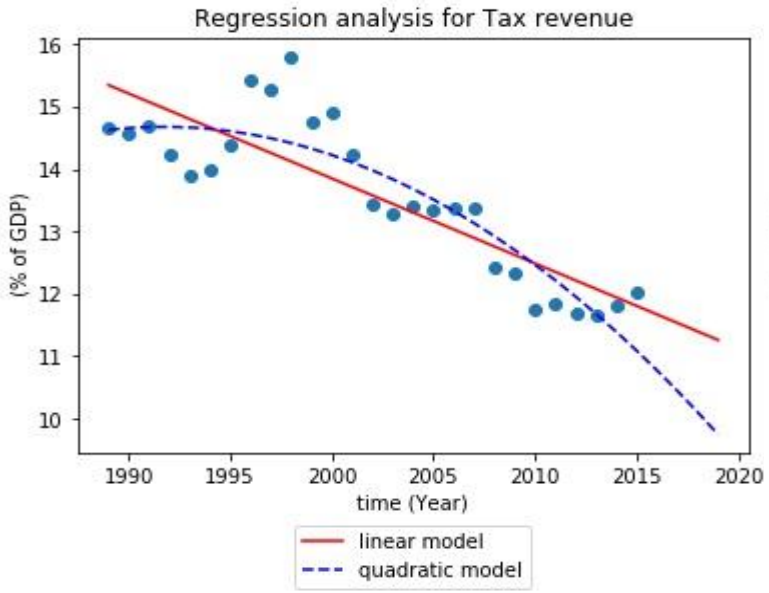


d)Japan

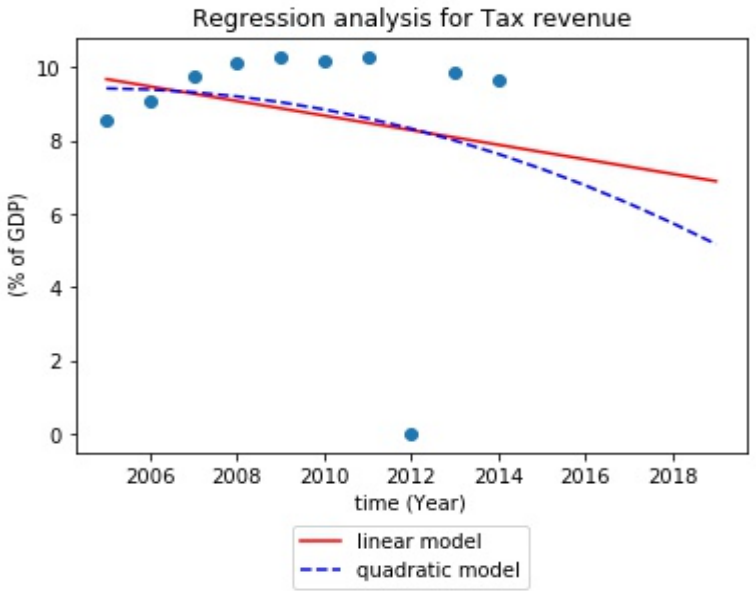


d)United States

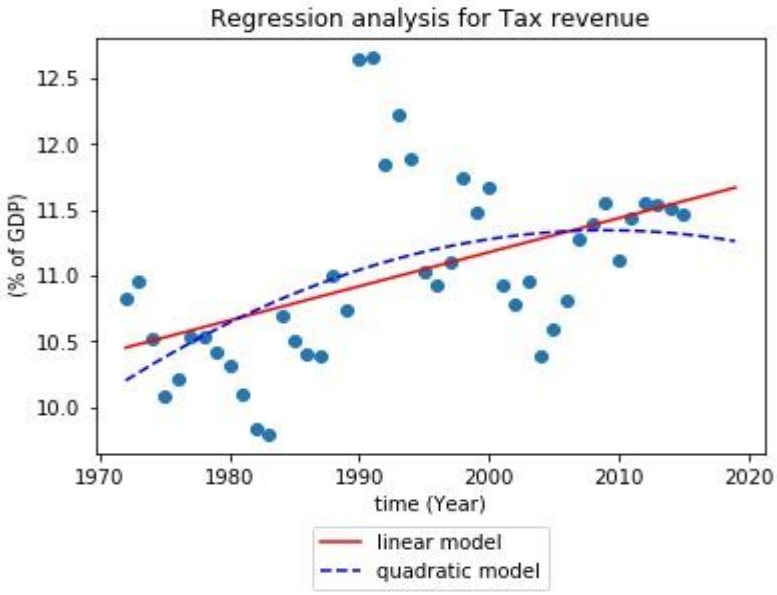
Figure 5- Regression analysis of GDP for five countries



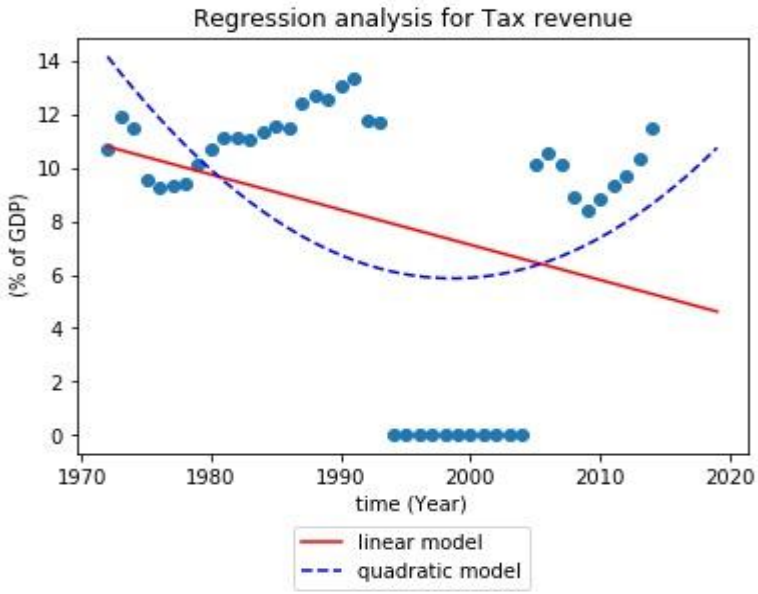
a)Canada



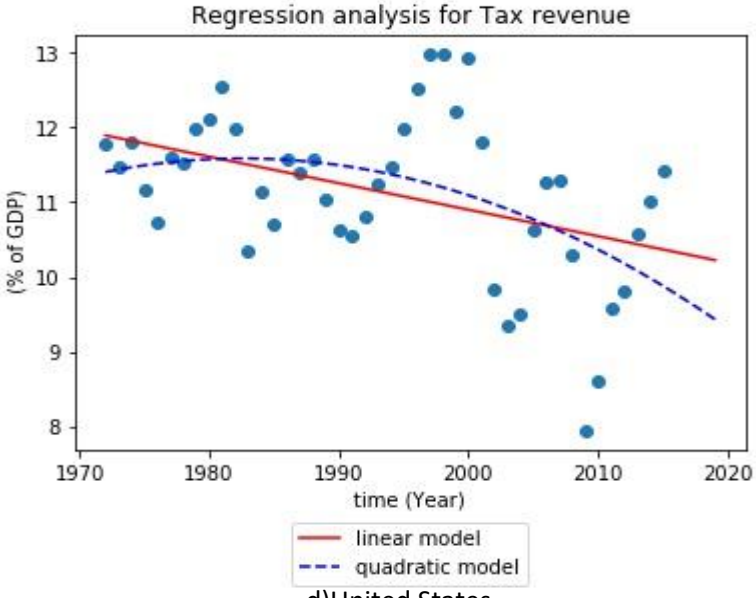
b)China



c)Germany



d)Japan



d)United States

Figure 6-The regression analysis of Tax revenue for 5 developed countries

As you can see in Figure 3, the linear model successfully predicts the behaviour of some indicators. For example, in Figure 3.a, we can see we would have less people with tuberculosis in the next 4 years. Similarly, in Figure 3.b, it can be shown that the consumer price index will increase linearly in the next 4 years based on linear model prediction. One good news is that for the next 4 years, the life expectancy will increase based on Figure 3.c, 3.d,3. e.

In contrast to Figure 3, in Figure 4,5,6 we cannot rely on the linear model as it does not provide good prediction for the sample points. Therefore, we used the quadratic modelling for that. Figure 4 shows the birth rate prediction for the next four years that is predicted by the quadratic model.

5.4. Correlation Analysis

In this part, we study the relationship between economy indicators for the five countries using Numpy library of Python. We studied the correlation among 60 economy indicators and among which we chose fertility rate, unemployment rate, birth rate, consumer price index, death rate, Gross Domestic Product (GDP), GINI index, inflation rate, rural population, unemployment rate as the main indicators. Because there is too many correlation information, we only show the ones that seem interesting.

If the absolute value of correlation between indicator values is higher than 0.66, we consider it as a correlation between indicators. For example, birth rate is correlated with 18 out of 60 indicators for Canada. Figure 7 shows the correlation between birth rate with rural population in which they are correlated with 0.95 percent. In fact, it is meaningful to consider when the birth rate decreases, the rural population diminishes as a result.

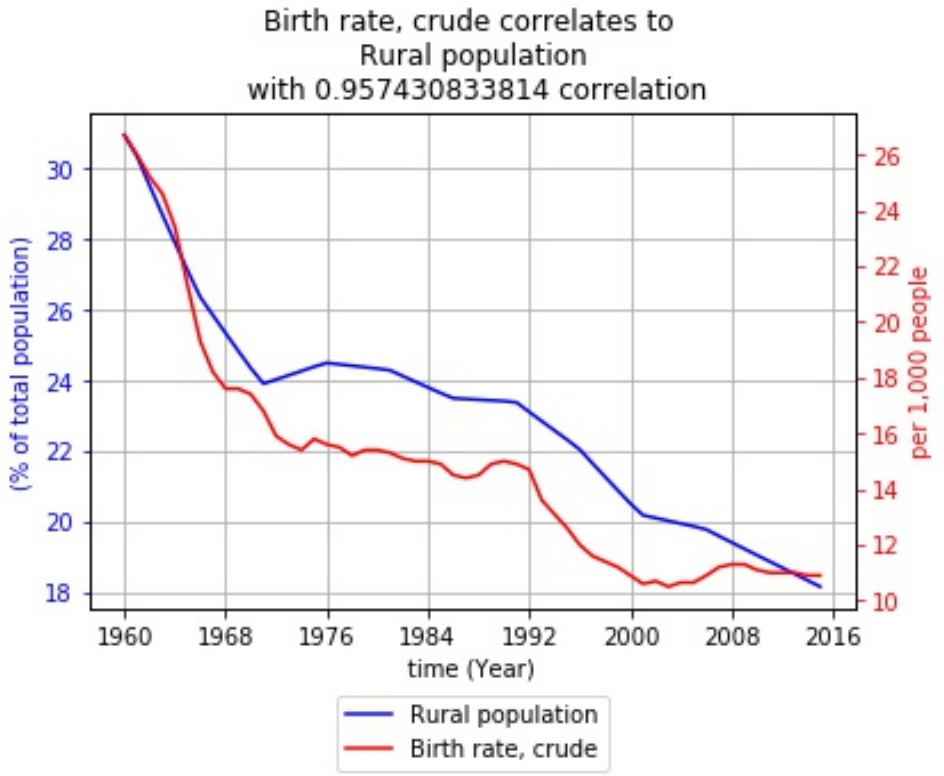


Figure 7- The correlation analysis between birth rate and rural population

More interesting fact is the relationship between birth rate and life expectancy. Figure 8 shows that when the birth rate decreases the life expectancy increases. It is counterintuitive because it means that people are happier and live longer without kids!

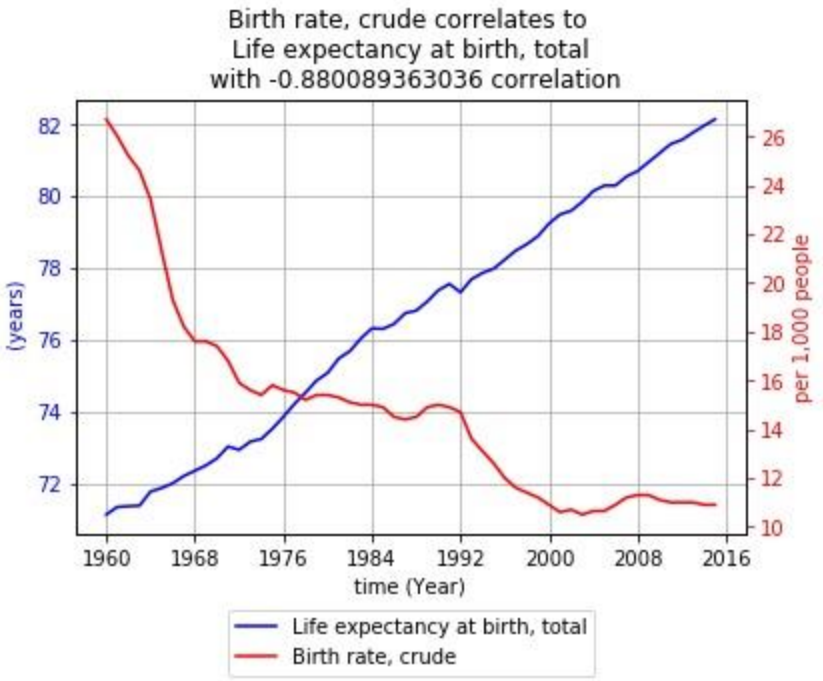


Figure 8- Correlation analysis between birth rate and life expectancy for Canadians

Another interesting fact is that when GDP increases, the birth rate decreases as it is shown in Figure 9 where the correlation between GDP and birth rate is depicted for the Canadian citizens.

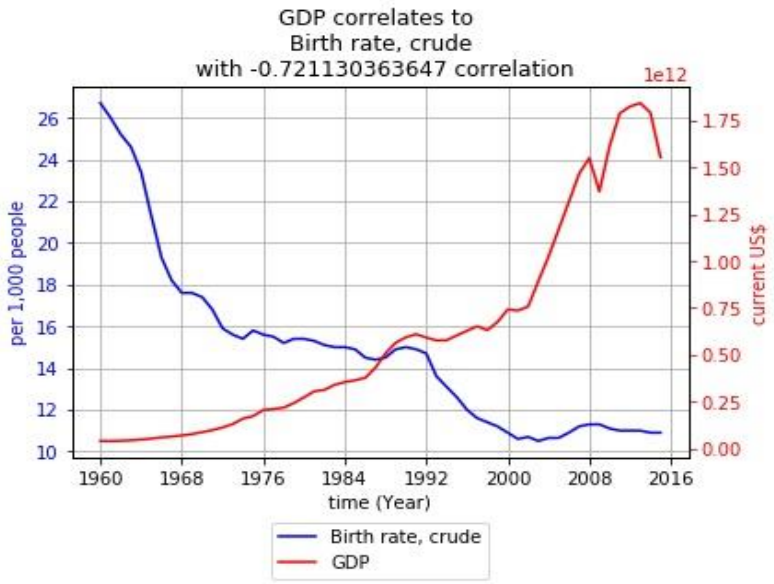


Figure 9- Correlation between GDP and birth rate for canadian people

Another interesting correlation is between GDP and life expectancy. As we expected, the more GDP a country gains, the more life expectancy it reaches. Figure 10 shows the correlation between these two indicators for Canadian people.

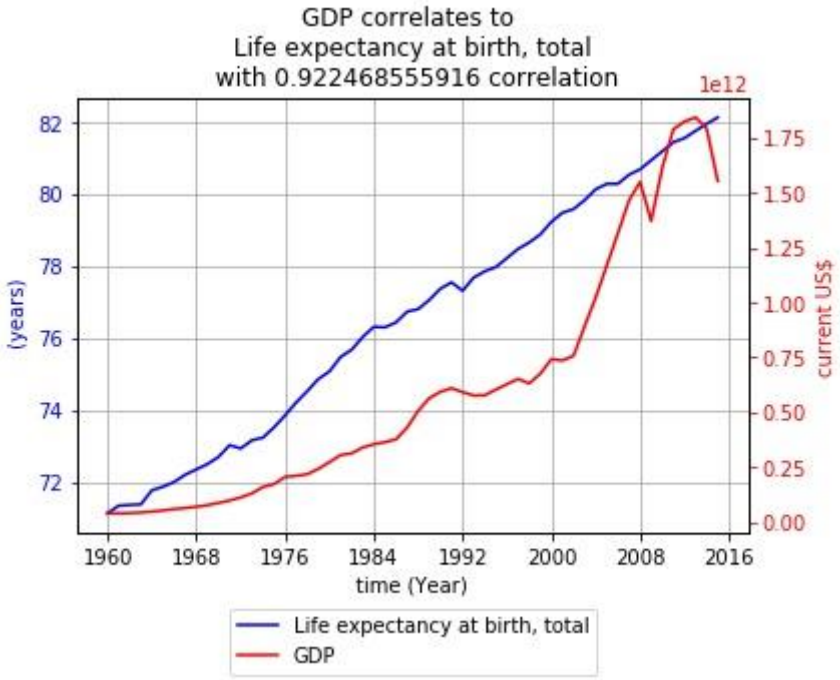


Figure 10- The correlation between GDP and life expectancy

One of the other interesting fact is the relationship between inflation and wage and salary of workers. Figure 11 shows the correlation between these two factors for Canadians.

Figure 11- The correlation between Inflation and wage and salaries of workers

As you can see from Figure 11, when hired workers decreases, the inflation will decrease too. This is not the good news for the workers as it means that if companies took less employers, the inflation rate will decrease too!

One of the other interesting or may be silly correlation is the correlation between of rural population with Air transport passenger carried. Figure 12 shows the correlation between these indicators for Canada.

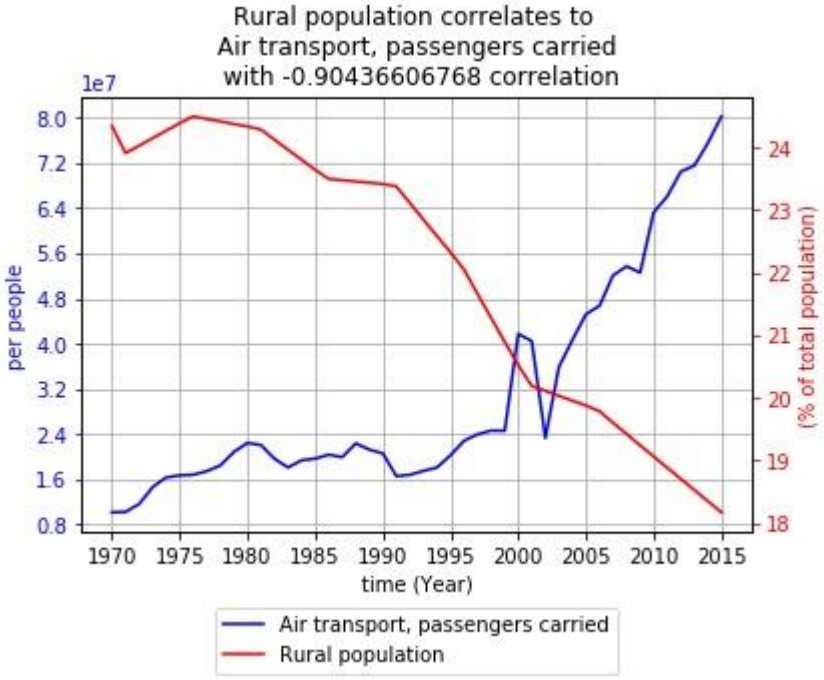


Figure 12- The correlation between rural population with air transport passengers carried

As you can be shown in Figure 12, when the rural population diminish, the number of people carried by airplane is increased.

6. Bibliography

economy. (n.d.). Retrieved from http://www.investopedia.com/terms/e/economy.asp

Gross_national_product. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Gross_national_product

Investopedia Staff. (2017). What is GDP and why is it so important to economists and investors? Retrieved from http://www.investopedia.com/ask/answers/199.asp

Mason Garrison. (2017). Misery_index_(economics). Retrieved from https://en.wikipedia.org/wiki/Misery_index_(economics)

Net_national_income. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Net_national_income

Prinsipe Ybarro. (2017). Economy. Retrieved from https://en.wikipedia.org/wiki/Economy

world-development-indicators. (2017). Retrieved from http://data.worldbank.org/data-catalog/world-development-indicators

6. Appendix

6.1. Excel:

For cleaning the data and transforming and summarizing the data, I used Excel. Here I included some scripts used:

1. For each indicator and each year:

   =INDEX(Data!E:E,MATCH(HSheet!A2&HSheet!$B2,Data!$A:$A&Data!$C:$C,0))

2. Using conditional formatting for highlighting the ones that are filled with zero

6.2. Python:

6.2.1 : Extraction information from CSV:

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jun  4 14:08:31 2017

@author: Mahdix
"""
import csv
def retrieve_Hsheet(filename):
    infile = open(filename,'r')
    csvdata =csv.reader(infile)
    Hsheet=[]
    for row in csvdata:
        Hsheet.append(row)
    infile.close()
    return Hsheet
def wanted_cols_func(filename,except_cols):
    infile = open(filename,'r')
    csvdata =csv.reader(infile)
```

```
    wanted_cols=[]
    for row in csvdata:
        if row[0] not in except_cols:
                wanted_cols.append(row[0])
    infile.close()
    return wanted_cols

def retrieve_axis(filename):
    infile = open(filename,'r')
    csvdata =csv.reader(infile)
    wanted_cols=[]
    for row in csvdata:
        wanted_cols.append(row[1])
    infile.close()
    return wanted_cols

def retrieve_cols(filename,cols,wanted_cols):
    infile = open(filename,'r')
    csvdata =csv.reader(infile)
    for row in csvdata:
        if row[0] == cols and row[2]==wanted_cols:
            result=_intStrArray(row[3:])
    infile.close()
    return result
def retrieve_col(filename,wanted_col):
    infile = open(filename,'r')
    csvdata =csv.reader(infile)
    for row in csvdata:
        if row[0] == wanted_col:
```

```
            result=_intStrArray(row[1:])
        infile.close()
        return result
    def _intStrArray(row):
        for ii in range(len(row)):
            if row[ii]=='':
                row[ii]='0'
        return row
```

6.2.2. Regression analysis:

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jun 18 09:04:59 2017

@author: Mahdix
"""

import numpy as np
import scipy as sc
import matplotlib.pyplot as py
import extraction
import os
def create_dir(dirname):
    if not os.path.exists(dirname):
        os.makedirs(dirname)
#main program
Hsheet=extraction.retrieve_Hsheet('Hsheet.csv')
Country_names=[]
for ii in range(1,6):
```

```
    Country_names.append(Hsheet[ii][0].strip())

# The name of columns that the main one is going to correlate based on
wanted_cols=[]
wanted_cols_names=[]
wanted_axis=[]
for ii in range(1,60):
    wanted_cols.append(Hsheet[ii][1])
    wanted_cols_names.append(Hsheet[ii][2].strip())
    wanted_axis.append(Hsheet[ii][3])
reg='Regression'
create_dir(reg)
for jj in range(0,len(Country_names)):
    Country_name=Country_names[jj]
    create_dir(reg+"/"+Country_name)
    for kk in range(0,len(wanted_cols)):
        col1=extraction.retrieve_col(Country_name+'.csv',wanted_cols[kk])
        new_index=col1.index(next((x for x in col1 if x!='0'), '0'))
        colr=col1[::-1]
        last = len(colr)-colr.index(next((x for x in colr if x!='0'), '0'))
        col1=col1[new_index:last]
        col1=[float(i) for i in col1]
        x=np.arange(1960+new_index, 2017-(len(colr)-last), 1)
        x2=np.arange(1960+new_index,2020)
        y=np.array(col1)
        fig = py.figure()
        p1=sc.polyfit(x,y,1)
        p2=sc.polyfit(x,y,2)
        py.plot(x,y,'o')
```

```
        lns1=py.plot(x2,sc.polyval(p1,x2),'r-')
        lns2=py.plot(x2,sc.polyval(p2,x2),'b--')
        py.title('Regression analysis for ' +wanted_cols_names[kk])
        py.xlabel('time (Year)')
        py.ylabel(wanted_axis[kk])
        lns=lns1+lns2
        lgd =py.legend(lns, ['linear model','quadratic model'], loc='upper center',
    bbox_to_anchor=(0.5, -0.15))
        py.show()
    fig.savefig(reg+"/"+Country_name+"/"+wanted_cols_names[kk].replace(",","_"
    ).replace(" ","")+'.jpg',format='jpg', bbox_extra_artists=(lgd,),
    bbox_inches='tight')
```

6.2.3. **Correlation analysis**:

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jun 11 20:13:10 2017

@author: Mahdi Aziz
"""
import numpy as np
import extraction
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import os
def create_dir(dirname):
   if not os.path.exists(dirname):
      os.makedirs(dirname)
def lens(listoflists):
  return [len(x) for x in listoflists]
```

```
def cor(a,b):
    return np.corrcoef(a,b)[0, 1]
def
plot_foursides(x,y1,y2,title,ax,ax22,Country_name,Col_except_name,Col_wanted
_name,filename):
    fig, ax1 = plt.subplots()
    for axis in [ax1.xaxis, ax1.yaxis]:
        axis.set_major_locator(ticker.MaxNLocator(integer=True))
    lns1=ax1.plot(x, y1, 'b-')
    ax1.grid()
    ax1.set_xlabel('time (Year)')
    ax1.set_ylabel(ax22, color='b')
    ax1.tick_params('y', colors='b')
    ax1.set_title(title, fontsize='large')
    ax2 = ax1.twinx()
    lns2=ax2.plot(x, y2, 'r-')
    ax2.set_ylabel(ax, color='r')
    ax2.tick_params('y', colors='r')
    lns = lns1+lns2
    lgd =ax1.legend(lns, [Col_wanted_name,Col_except_name], loc='upper center',
bbox_to_anchor=(0.5, -0.15))
    plt.show()

fig.savefig(Country_name+"/"+Col_except_name+"/"+filename.replace(",","_").re
place(" ","")+'.jpg',format='jpg', bbox_extra_artists=(lgd,), bbox_inches='tight')
## other measures vs GDP
Hsheet=extraction.retrieve_Hsheet('Hsheet.csv')
Country_names=[]
for ii in range(1,6):
    Country_names.append(Hsheet[ii][0].strip())
```

```
# The name of columns that the main one is going to correlate based on
wanted_cols=[]
wanted_cols_names=[]
wanted_axis=[]
for ii in range(1,60):
    wanted_cols.append(Hsheet[ii][1])
    wanted_cols_names.append(Hsheet[ii][2].strip())
    wanted_axis.append(Hsheet[ii][3])
#The name without axis and axis and fullname of the columns that the Correlation is based upon
except_cols=[]
except_names=[]
except_axis=[]
for ii in range(1,11):
    except_cols.append(Hsheet[ii][4])
    except_names.append(Hsheet[ii][5].strip())
    except_axis.append(Hsheet[ii][6])



#Country_name=['Canada','China','Germany','Japan','United States']
for jj in range(0,len(Country_names)):
    Country_name=Country_names[jj]
    create_dir(Country_names[jj])
    for ii in range(0,len(except_names)):
        create_dir(Country_name+"/"+except_names[ii])
        col2=extraction.retrieve_col(Country_name+'.csv',except_cols[ii])
        col2c=[float(i) for i in col2]
        for kk in range(0,len(wanted_cols)):
```

```
        col2=col2c
        col1=extraction.retrieve_col(Country_name+'.csv',wanted_cols[kk])
        new_index=col1.index(next((x for x in col1 if x!='0'), '0'))
        colr=col1[::-1]
        last = len(colr)-colr.index(next((x for x in colr if x!='0'), '0'))
        col1=col1[new_index:last]
        col2=col2[new_index:last]
        col1=[float(i) for i in col1]
        corr=cor(col1,col2)
        if abs(corr)>0.66 and corr!=1.0:
            title=except_names[ii]+' correlates to \n' + wanted_cols_names[kk]+"\n
with "+str(corr) +" correlation";
            x1 = np.arange(1960+new_index, 2017-(len(col2c)-last), 1)
            filename=except_names[ii] + "_"+wanted_cols_names[kk]

plot_foursides(x1,col1,col2,title,except_axis[ii],wanted_axis[kk],Country_name,ex
cept_names[ii],wanted_cols_names[kk],filename)
```

6.2.4. Comparison among countries (Figure 1,Figure 2)

```
    # -*- coding: utf-8 -*-
    """
    Created on Sun Jun  4 19:25:35 2017

    @author: Mahdi Aziz - SN: 8024161
    """

    import extraction
    import seaborn as sb
    import matplotlib.pyplot as plt
```

```
def plot_res(wanted_cols,xlabel,ylabel,title,picname):
    cols=["Canada","China","Germany","Japan","United States"]
    result=extraction.retrieve_cols('WDIEXCEL.csv',cols[0],wanted_cols)
    print(result)
    result2=extraction.retrieve_cols('WDIEXCEL.csv',cols[1],wanted_cols)
    result3=extraction.retrieve_cols('WDIEXCEL.csv',cols[2],wanted_cols)
    result4=extraction.retrieve_cols('WDIEXCEL.csv',cols[3],wanted_cols)
    result5=extraction.retrieve_cols('WDIEXCEL.csv',cols[4],wanted_cols)
    sb.set_palette("colorblind")
    sb.set_style("darkgrid")
    sb.set_context("notebook",rc={"lines.linewidth":4.3})
    fig = plt.figure(figsize=(20,10),dpi=300);
    x=range(1960,2017)
    plt.plot(x,result,'r--',x,result2,'b--',x,result3,'g--',x,result4,'y--',x,result5,'--o')
    plt.ylabel(ylabel)
    plt.xlabel(xlabel)
    plt.legend(cols)
    plt.title(title)
    fig.savefig(picname,format='jpg')


def plot_res_all():
    #plot_res("Inflation, GDP deflator (annual %)",'Years','Inflation rate','The inflation rate of 5 developed countries from 1960 to 2016','inflation.jpg')
    #plot_res("Inflation, consumer prices (annual %)",'Years','Inflation rate','The inflation rate of 5 developed countries from 1960 to 2016','inflation2.jpg')
    plot_res('Unemployment, total (% of total labor force) (national estimate)','Years','Unemployment rate','The unemployment rate of 5 developed countries from 1960 to 2016','unemployment.jpg')
plot_res_all()
```