Project: Gapminder Data Analysis Project

Table of Contents

Questions being investigated in the analysis:

```
1. Analyze the population growth rate for each country and compare the trends of al
l 3 countries.
2. Analyze and compare the per capita health expenditure data for all 3 countries.
3. Analyze the number of mobile phones per 100 people data for each country and com
pare the trends of all
   3 countries.
4. Is there any correlation between the each 2 of the 3 indicators selected?
```

Introduction

For this project, I have selected 3 indicators from the Gapminder data site: Population Growth (annual %), Mobile Cellular Subscriptions (per 100 people) and Per Capita Total Expenditure on Health at average exchange rate (US$). For each indicator, the yearly data is available for all the countries for different time periods. The available data has been analysed for variation of each indicator across the given time period for USA, Canada amd Mexico.

In [67]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
% matplotlib inline
```

Data Wrangling

In [47]:
```
#reading CSV files
df_pop=pd.read_csv('./Gapminder Data/population_growth.csv', delimiter=',')
df_phones=pd.read_csv('./Gapminder Data/cell_phones_per_100.csv', delimiter=
',')
df_health=pd.read_csv('./Gapminder Data/health_spending.csv', delimiter=',')
```

Data Cleaning

From each table, I need data to be extracted for USA, Mexico and Canada only and then arrange data in 1 dimensional format for analysis like [Country, Year, Indicator Value].

In all the three tables, data is present in a 2 dimensional format as can be seen above. The step by step process that was followed for cleaning data in each of the three tables is below. Step 1: Rename the first column to 'Year' Step 2: Split the Population Growth Rate data into 3 tables wrt to the countries.

```
    1. Set the index to 'Year' column inorder to locate the country name.
    2. Create another dataframe for the selected country by selecting the entire ro
w for that country.
    3. Transpose data to get all the years in one column and the corresponding  ind
icator value in the next column.
    4. Create another column for the country name.
    5. Repeat the above steps for other 2 countries.
    6. Append the data for all 3 countries into 1 table for analysis
```

Step 3: Follow Step 2 for Per Capita Helath Expenditure Data and Mobile Phones/100 people data.

FINALLY, AT THE END OF DATA CLEANING STEP, 3 DIFFERENT DATASETS ARE CREATED, 1 FOR EACH INDICATOR WITH THE DATA OF ALL 3 COUNTRIES IN EACH OF THEM.

In [48]:

```
#renaming columns
df_pop=df_pop.rename(index=str, columns={"Population growth (annual %)":"Year"
})
df_phones=df_phones.rename(index=str, columns={"Mobile cellular subscriptions
 (per 100 people)":"Year"})
df_health=df_health.rename(index=str, columns={"Per capita total expenditure o
n health at average exchange rate (US$)":"Year"})

#df_POP['per_capita_health_expenditure']=pd.Series(df_HEA['per capita health e
xpenditure'])
#df_POP.tail()
df_health.head()
```

Out[48]:

	Year	1995	1996	1997	1998	1999	2000	2
0	Abkhazia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Akrotiri and Dhekelia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Albania	27.910805	43.045818	36.135184	47.102142	65.024024	75.236623	79.862
4	Algeria	62.055538	61.769883	66.893742	65.983195	62.521470	62.607389	67.814

In [49]:

```
#DATA WRANGLING FOR POPULATION GROWTH DATA
df_pop.set_index('Year',inplace=True) #set index
```

In [50]:

```
#extracting population growth data data for USA
df_pop1=df_pop.loc[['United States']] #create another df for USA population gr
owth data
df_pop1.head()
df_pop1=df_pop1.T #transpose data
df_pop1=df_pop1.rename(index=str, columns={'United States': 'population growth
 rate(%)'}) #rename column
country_USA=np.repeat('USA', 52) #creating 'country' column for USA dataframe
 to join this data with that of other 2 countries
df_pop1['country']=country_USA ##adding array to dataframe
df_pop1.head()
```

Out[50]:

Year	population growth rate(%)	country
1960	1.701993	USA
1961	1.657730	USA
1962	1.537997	USA
1963	1.439165	USA
1964	1.389046	USA

In [51]:

```
#extracting population growth data data for Mexico
df_pop2=df_pop.loc[['Mexico']]
df_pop2=df_pop2.T
df_pop2=df_pop2.rename(index=str, columns={'Mexico': 'population growth rate
(%)'})
country_Mexico=np.repeat('Mexico', 52)
df_pop2['country']=country_Mexico
df_pop2.head()
```

Out[51]:

Year	population growth rate(%)	country
1960	3.294347	Mexico
1961	3.240361	Mexico
1962	3.180950	Mexico
1963	3.117918	Mexico
1964	3.053644	Mexico

In [52]:

```
#extracting population growth data data for Canada
df_pop3=df_pop.loc[['Canada']]
df_pop3=df_pop3.T
df_pop3=df_pop3.rename(index=str, columns={'Canada': 'population growth rate
(%)'})
country_Canada=np.repeat('Canada', 52)
df_pop3['country']=country_Canada
df_pop3.head()
```

Out[52]:

Year	population growth rate(%)	country
1960	2.298627	Canada
1961	2.001122	Canada
1962	1.859888	Canada
1963	1.862846	Canada
1964	1.885715	Canada

In [53]:

```
#combining the 3 population dataframes
df_POP = df_pop1.append([df_pop2,df_pop3])
df_POP.head()
df_POP.dtypes
df_POP.shape
df_POP.isnull().values.any()
df_POP.head()
```

Out[53]:

Year	population growth rate(%)	country
1960	1.701993	USA
1961	1.657730	USA
1962	1.537997	USA
1963	1.439165	USA
1964	1.389046	USA

In [54]:

```
#DATA WRANGLING FOR MOBILE PHONES DATA
df_phones.set_index('Year',inplace=True)
```

In [55]:

```
#extracting mobile phones/100 people data data for USA
df_phones1=df_phones.loc[['United States']]
df_phones1=df_phones1.T
df_phones1=df_phones1.rename(index=str, columns={'United States': 'mobile phones/100 people'})
phones_USA=np.repeat('USA', 11)
df_phones1['country']=phones_USA
df_phones1.head(11)
```

Out[55]:

Year	mobile phones/100 people	country
2001	45.001698	USA
2002	49.156350	USA
2003	55.146605	USA
2004	62.850112	USA
2005	68.627383	USA
2006	76.644603	USA
2007	82.471958	USA
2008	85.675203	USA
2009	89.149116	USA
2010	89.856451	USA
2011	105.913601	USA

In [56]:

```
#extracting mobile phones/100 people data data for Mexico
df_phones2=df_phones.loc[['Mexico']]
df_phones2=df_phones2.T
df_phones2=df_phones2.rename(index=str, columns={'Mexico': 'mobile phones/100 people'})
phones_Mexico=np.repeat('Mexico', 11)
df_phones2['country']=phones_Mexico
```

In [57]:

```
#extracting mobile phones/100 people data data for Canada
df_phones3=df_phones.loc[['Canada']]
df_phones3=df_phones3.T
df_phones3=df_phones3.rename(index=str, columns={'Canada': 'mobile phones/100
 people'})
phones_Canada=np.repeat('Canada', 11)
df_phones3['country']=phones_Canada
df_phones3.head()
```

Out[57]:

Year	mobile phones/100 people	country
2001	34.387958	Canada
2002	37.951869	Canada
2003	42.048734	Canada
2004	47.020366	Canada
2005	52.710040	Canada

In [58]:

```
#combining the 3 mobile phones dataframes
df_MOB = df_phones1.append([df_phones2,df_phones3])
df_MOB.head()
df_MOB.dtypes
df_MOB.shape
df_MOB.isnull().values.any()
```

Out[58]:

```
False
```

In [59]:

```
#DATA WRANGLING FOR HEALTH EXPENDITURE DATA
df_health.set_index('Year',inplace=True)
```

In [60]:

```
#extracting health expenditure data for USA
df_health1=df_health.loc[['United States']]
df_health1=df_health1.T
df_health1=df_health1.rename(index=str, columns={'United States': 'per capita
 health expenditure'})
health_USA=np.repeat('USA', 16)
df_health1['country']=health_USA
df_health1.head()
```

Out[60]:

Year	per capita health expenditure	country
1995	3747.692121	USA
1996	3899.976933	USA
1997	4054.627219	USA
1998	4235.837199	USA
1999	4450.044994	USA

In [61]:

```
#extracting health expenditure data for Mexico
df_health2=df_health.loc[['Mexico']]
df_health2=df_health2.T
df_health2=df_health2.rename(index=str, columns={'Mexico': 'per capita health
 expenditure'})
phones_Mexico=np.repeat('Mexico', 16)
df_health2['country']=phones_Mexico
```

In [62]:

```
#extracting health expenditure data for Canada
df_health3=df_health.loc[['Canada']]
df_health3=df_health3.T
df_health3=df_health3.rename(index=str, columns={'Canada': 'per capita health
 expenditure'})
health_Canada=np.repeat('Canada', 16)
df_health3['country']=health_Canada
df_health3.head()
```

Out[62]:

Year	per capita health expenditure	country
1995	1820.609018	Canada
1996	1828.837371	Canada
1997	1873.002802	Canada
1998	1848.571848	Canada
1999	1936.587877	Canada

In [63]:

```
#combining the 3 health expenditure dataframes
df_HEA = df_health1.append([df_health2,df_health3])
#df_HEA.set_index('Year')
df_HEA.head()

df_HEA.dtypes
df_HEA.shape
df_HEA.isnull().values.any()
```

Out[63]:

```
True
```

There is 1 null value in this dataframe for Mexico for year 2010. Since this value comes at the end of the curve in the plot and it does not make any sense in this case to replace with an average of values for all other years, it has been ignored.

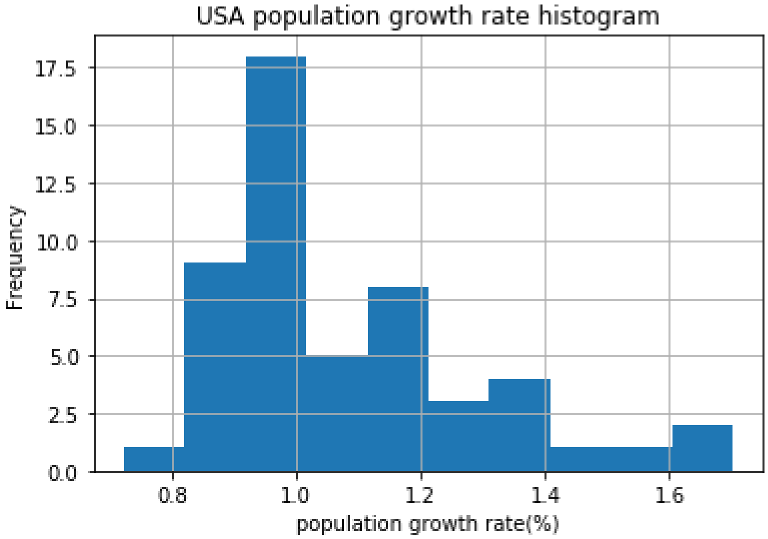
Exploratory Data Analysis

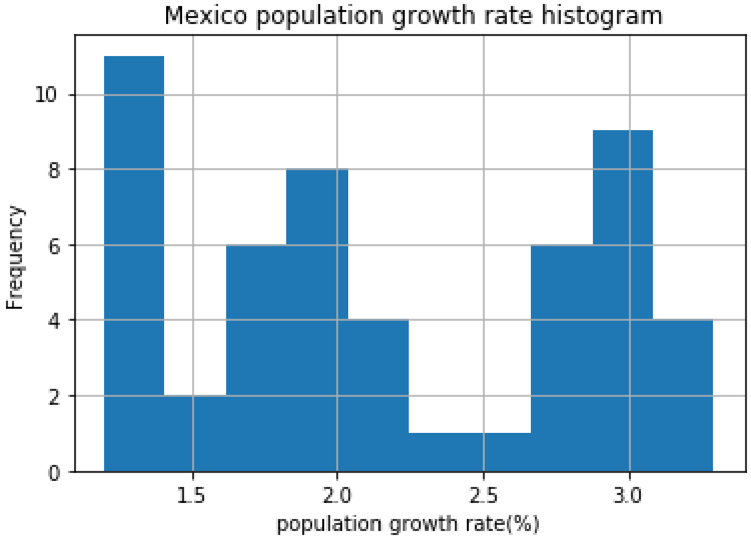
RESEARCH QUESTION 1: Observe and analyse the population growth rate for each country and compare the trends of all 3 countries

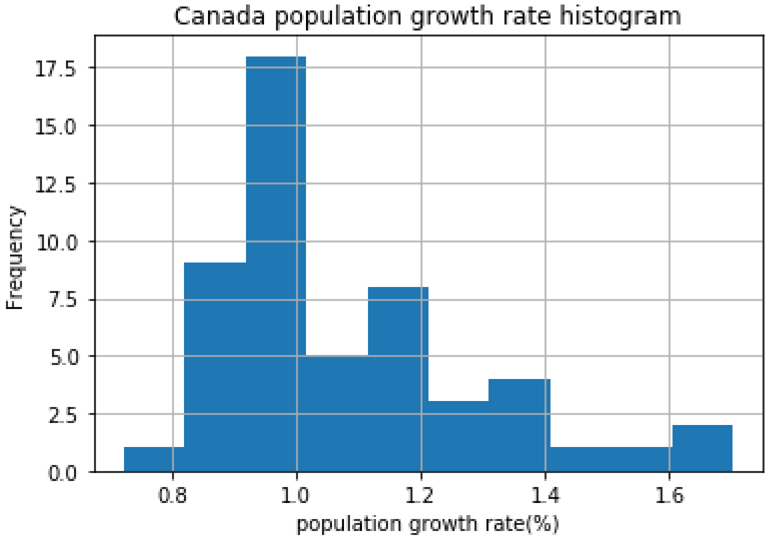
In [88]:

```
df_pop1.hist(column='population growth rate(%)')
plt.xlabel("population growth rate(%)")
plt.ylabel("Frequency")
plt.title('USA population growth rate histogram');
```

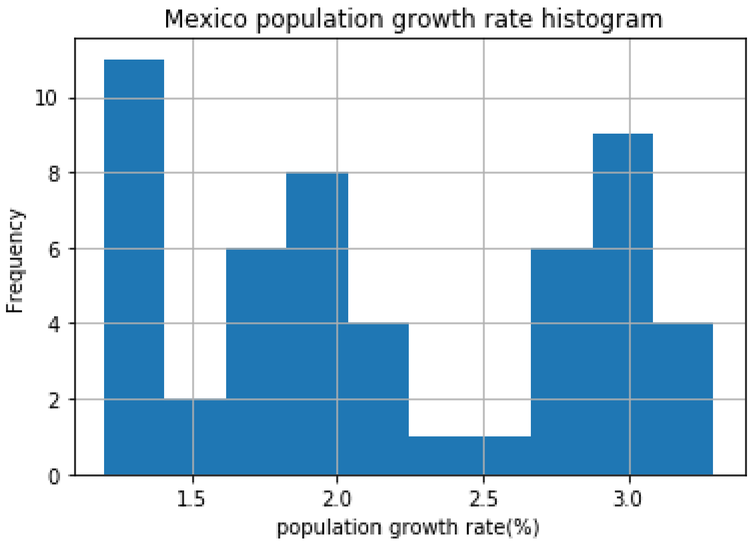
```
<matplotlib.figure.Figure at 0x2730fb85d30>
```


USA population growth rate histogram
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.8
1.0
1.2
1.4
1.6
population growth rate(%)

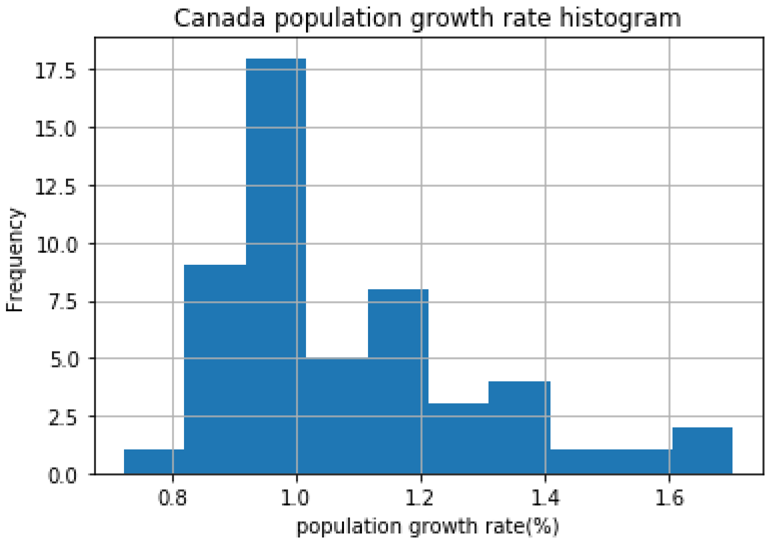

Mexico population growth rate histogram
Frequency
0
2
4
6
8
10
1.5
2.0
2.5
3.0
population growth rate(%)


Canada population growth rate histogram
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.8
1.0
1.2
1.4
1.6
population growth rate(%)

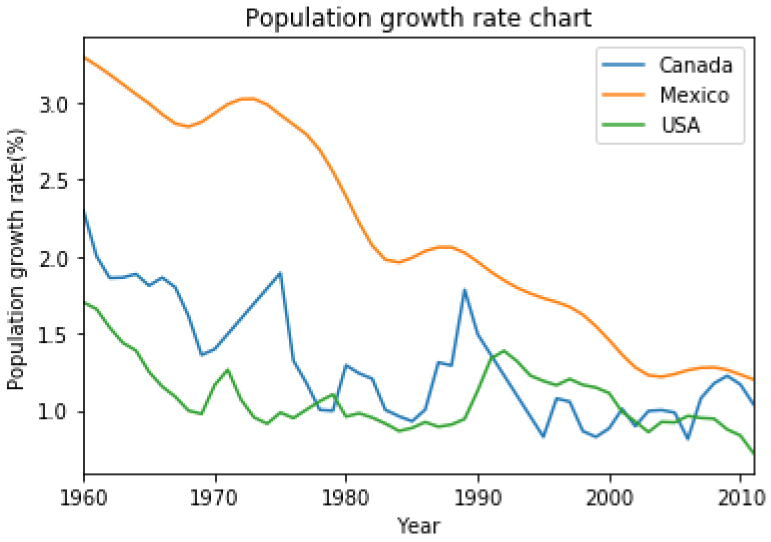
```
In [81]: df_pop2.hist(column='population growth rate(%)')
         plt.xlabel("population growth rate(%)")
         plt.ylabel("Frequency")
         plt.title(' Mexico population growth rate histogram');
```



```
In [83]: df_pop1.hist(column='population growth rate(%)')
         plt.xlabel("population growth rate(%)")
         plt.ylabel("Frequency")
         plt.title('Canada population growth rate histogram');
```



In [37]:
```
df_POP.reset_index(drop=True)
df_POP.groupby('country')['population growth rate(%)'].plot(x='index',y='population growth rate(%)',legend=True, title='Population growth rate chart')
plt.xlabel('Year')
plt.ylabel('Population growth rate(%)');
```

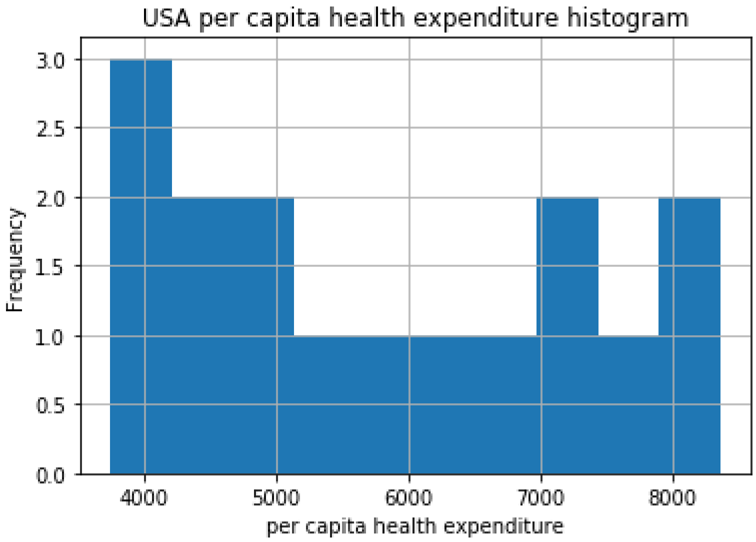


1.For USA, the popoulation growth rate was highest for the year 1960. The next highest peak was for the year 1992. The least growth rate was for the latest recorded year 2011, which is on a downward sloping curve and based on that it looks like, the growth rate might have declined further close to zero for the next few years too.

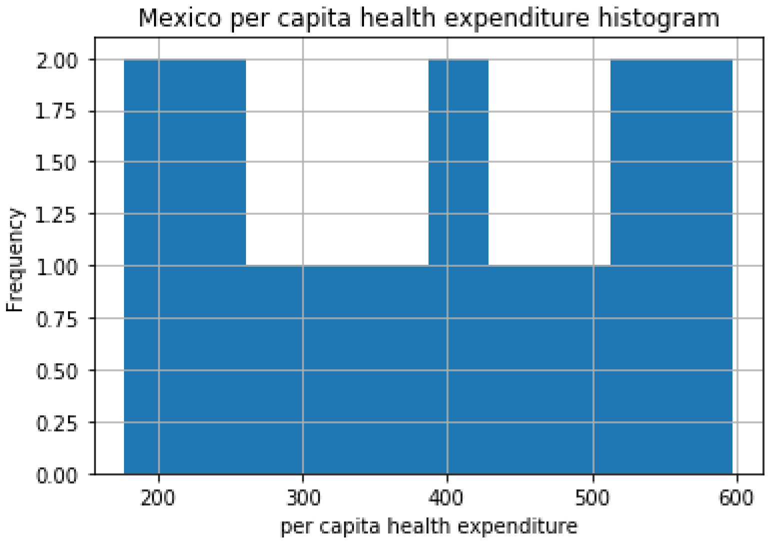
1. For Mexico, the growth rate started at highest value of all the 3 countries, and as expected, followed a decreasing trend while showing deviations at three intervals: 1969-73,85-88 and 2004-07.
2. In Canada,there is sharp rise and then immediate decline in population growth rates for the time intervals 1968-1978 & 1988-1995. Looking at the curve near 2010, there is a decreasing trend in growth rate around that area.
3. Comparing the curves of all 3 countries, irrespective of where the curve began on y-axis, all 3 countries seem to have their peaks in growth rates at around the same time, although the peaks are at different heights for each of them and for different time intervals. As with the shape of the curve for the latest five years, growth rate seems to be declining for all 3 countries.

RESEARCH QUESTION 2: Analyse and compare the per capita health expenditure data for all 3 countries

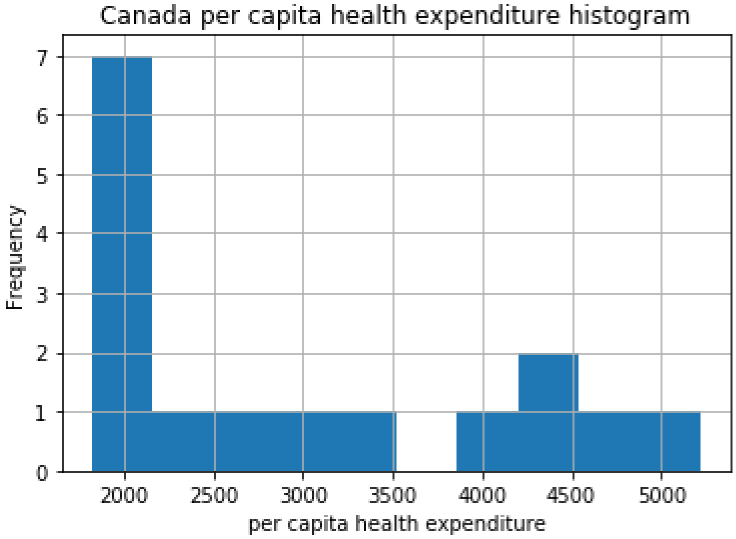
In [84]:
```
df_health1.hist(column='per capita health expenditure')
plt.xlabel("per capita health expenditure")
plt.ylabel("Frequency")
plt.title('USA per capita health expenditure histogram');
```



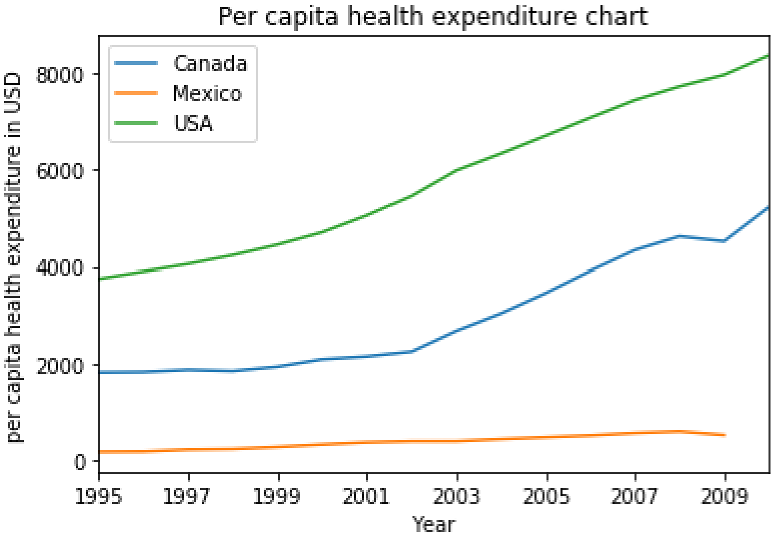
In [90]:
```
df_health2.hist(column='per capita health expenditure')
plt.xlabel("per capita health expenditure")
plt.ylabel("Frequency")
plt.title('Mexico per capita health expenditure histogram');
```



In [91]:
```
df_health3.hist(column='per capita health expenditure')
plt.xlabel("per capita health expenditure")
plt.ylabel("Frequency")
plt.title('Canada per capita health expenditure histogram');
```



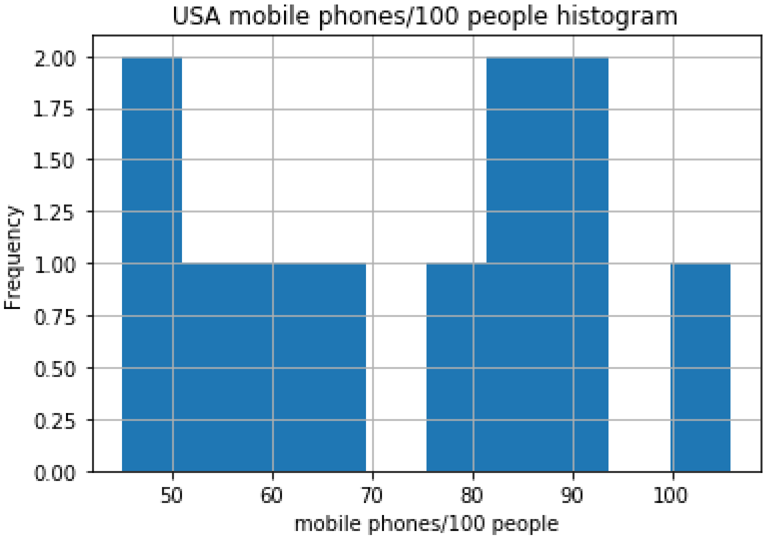
In [39]:
```
#df_HEA.set_index('index', inplace=True)
df_HEA.reset_index(drop=True)
df_HEA.groupby('country')['per capita health expenditure'].plot(x='index',y='p
er capita health expenditure',legend=True, title='Per capita health expenditur
e chart');
plt.xlabel('Year')
plt.ylabel('per capita health expenditure in USD');
```



The health expenditure per capita in USD for Canada and Mexico has remained constant from 1995 to 1998 and there after has increased very slightly for Mexico and considerably for Canada until 2002. We can see a sudden rise for Canada beginning from the year 2002 up till 2008. Then fromm 2009 onwards, it has increased again. For USA, the curve has been rising ever since 1995. But the slope seems to be decreasing, which means that the rate at which the cost has been going up from 2003 onwards is not as much as that for the period 2000-2003.

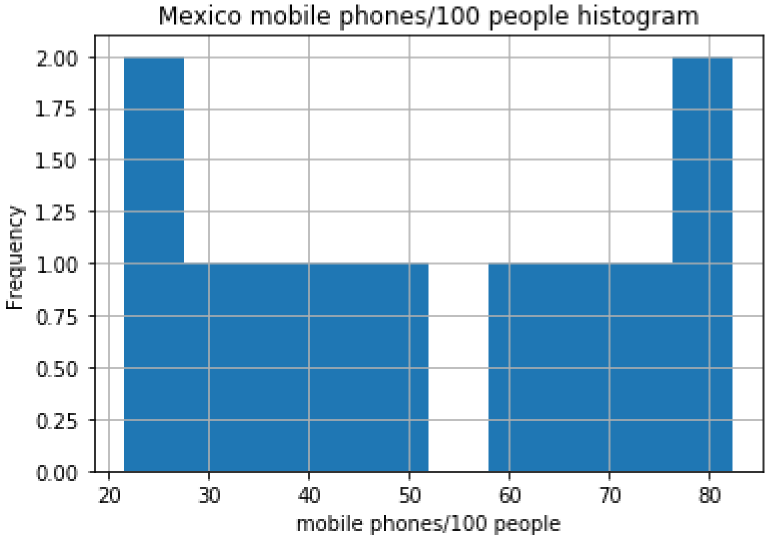
RESEARCH QUESTION 3: Observe and analyse the number of mobile phones per 100 people data for each country and compare the trends of all 3 countries.

In [92]:
```
df_phones1.hist(column='mobile phones/100 people')
plt.xlabel("mobile phones/100 people")
plt.ylabel("Frequency")
plt.title('USA mobile phones/100 people histogram');
```



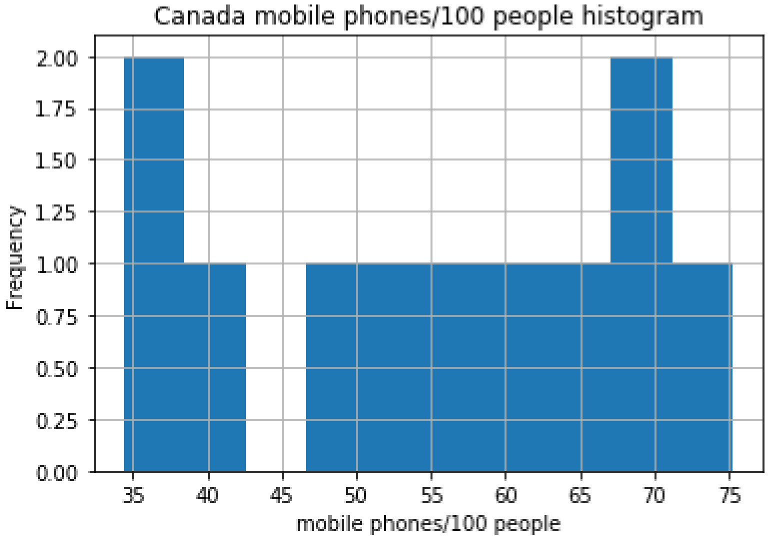
In [93]:

```
df_phones2.hist(column='mobile phones/100 people')
plt.xlabel("mobile phones/100 people")
plt.ylabel("Frequency")
plt.title('Mexico mobile phones/100 people histogram');
```

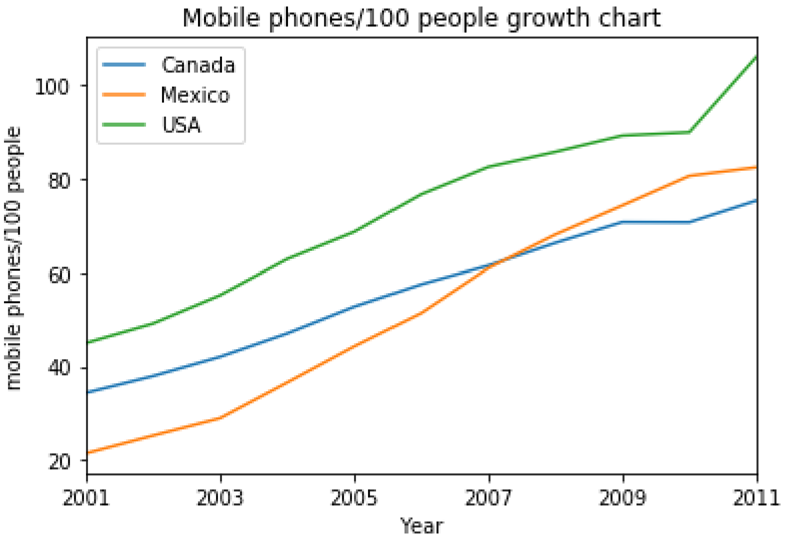


In [94]:

```
df_phones3.hist(column='mobile phones/100 people')
plt.xlabel("mobile phones/100 people")
plt.ylabel("Frequency")
plt.title('Canada mobile phones/100 people histogram');
```



In [41]:
```
#df_MOB.set_index('index', inplace=True)
df_MOB.reset_index(drop=True)
df_MOB.groupby('country')['mobile phones/100 people'].plot(x='index',y='mobile
 phones/100 people',legend=True,title='Mobile phones/100 people growth chart'
);
plt.xlabel('Year')
plt.ylabel('mobile phones/100 people');
```



1. The overall fastest growth in usage of mobile phones is for USA. Though the growth line for Mexico was below other countries, it soon touched the Canada growth line in 2007.From this point,the numbers for Mexico increased considerably faster than Canada.
2. At this point, it can be said that as mobile phones had become a necessity for each and every person and technlogy had spread widely, population of a country is a major deciding factor for this indicator. The number would certainly be higher for countries with larger populations. And before early 2000s, mobile technology in a country could have been its major deciding factor for this indicator.
3. For 2009-2010, the curve has been flat for USA and Canada. For 2010-2011, there is a marked difference in the slope of curves. The increase is highest for USA, then Canada and then the least growth can be seen for Mexico.

Conclusions

Limitations:

1. The data is available only upto 2011 for population growth and mobile phones data and upto 2010 for health expenditure data. This missing data is a limitation.
2. In future, when the data is available upto the current year or the data of few more years is added, the above visualisations can be enhanced and better analysis can be performed.
3. The accuracy of this analysis depends on the accuracy of data collected by Gapminder and applies only for the time duration for which the data has been collected.
4. No staistical testing has been performed on the data.
5. When comparing population growth rate with the other two indicators, it should be noted that population growth rate might only be one of the many factors that affect their variation over the period of time. For example, literacy rate and access to technology.

The variation of all the selected indicators vs the time period for which the data was recorded has been analysed and the significant points have been noted below.

1. Population growth rate has been following a very random pattern over the given time period, especially for USA and Canada. And there might be a multitude of factors responsible for variation of this indicator over the given time period.
2. Population growth might be one of the major deciding factors driving the growth in indicator health expenditure numbers and number of mobile phone users.(Though we have not analysed the actual population numbers here, it is clear from population growth rate chart that all 3 countries have had growth in their populations over the years for which we have the data for other two indicators)