Import Libraries

In [1]:

```
import numpy as np
import pandas as pd
from pandas import DataFrame as df
import matplotlib.pyplot as plt
import seaborn as sns
```

Load Dataset

In [10]:

```
canada = pd.read_excel('https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DV0101EN/labs/Data_Files/Canada.xlsx',
 sheet_name='Canada by Citizenship',
 skiprows=range(20),
 skipfooter=2)
```

In [12]:

```
canada.head()
#spare df
canada1 = canada
```

Rename columns

In [13]:

```
canada.rename(columns={'OdName':'Country', 'AreaName':'Continent', 'RegName':'Continent-Region'}, inplace=True)
canada.drop(['AREA','REG','DEV','Type','Coverage'], inplace=True, axis='columns')
canada.isnull().sum().sum()
```

Out[13]:

0

In [14]:

```
canada.head(5)
```

Out[14]:

	Country	Continent	Continent-Region	DevName	1980	1981	1982	1983	1984	1985	...
0	Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	...
1	Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	...
2	Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	...
3	American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	...
4	Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	...

5 rows × 38 columns

In [15]:

```
canada.index.values
```

Out[15]:

```
array([  0,   1,   2,   3,   4,   5,   6,   7,   8,   9,  10,  11,
12,
        13,  14,  15,  16,  17,  18,  19,  20,  21,  22,  23,  24,
25,
        26,  27,  28,  29,  30,  31,  32,  33,  34,  35,  36,  37,
38,
        39,  40,  41,  42,  43,  44,  45,  46,  47,  48,  49,  50,
51,
        52,  53,  54,  55,  56,  57,  58,  59,  60,  61,  62,  63,
64,
        65,  66,  67,  68,  69,  70,  71,  72,  73,  74,  75,  76,
77,
        78,  79,  80,  81,  82,  83,  84,  85,  86,  87,  88,  89,
90,
        91,  92,  93,  94,  95,  96,  97,  98,  99, 100, 101, 102,
103,
       104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115,
116,
       117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128,
129,
       130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141,
142,
       143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154,
155,
       156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167,
168,
       169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180,
181,
       182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193,
194])
```

Reassign Index

In [16]:

```
canada.set_index('Country', inplace=True)
```

In [18]:

```
canada.head()
```

Out[18]:

Country	Continent	Continent-Region	DevName	1980	1981	1982	1983	1984	1985	1986
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496
Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	1
Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69
American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	0
Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	2

5 rows × 37 columns

Convert columns to str

In [154]:

```
canada.dtypes
canada.columns = list(map(str,canada.columns));
```

Subsetting

In [23]:

```
canada[canada['Continent']=='Asia'].head(4)
```

Out[23]:

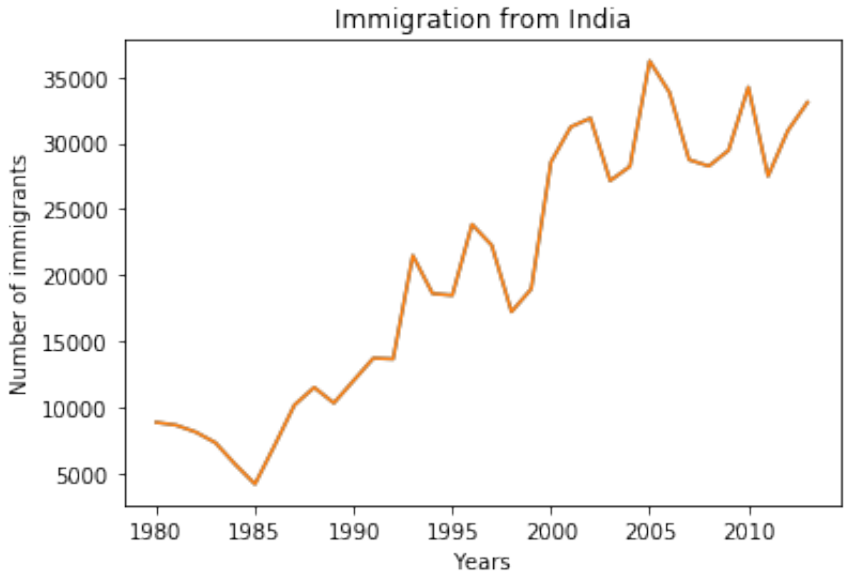
Country	Continent	Continent-Region	DevName	1980	1981	1982	1983	1984	1985	1986
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496
Armenia	Asia	Western Asia	Developing regions	0	0	0	0	0	0	0
Azerbaijan	Asia	Western Asia	Developing regions	0	0	0	0	0	0	0
Bahrain	Asia	Western Asia	Developing regions	0	2	1	1	1	3	0

4 rows × 37 columns

Viewing Line chart

In [30]:

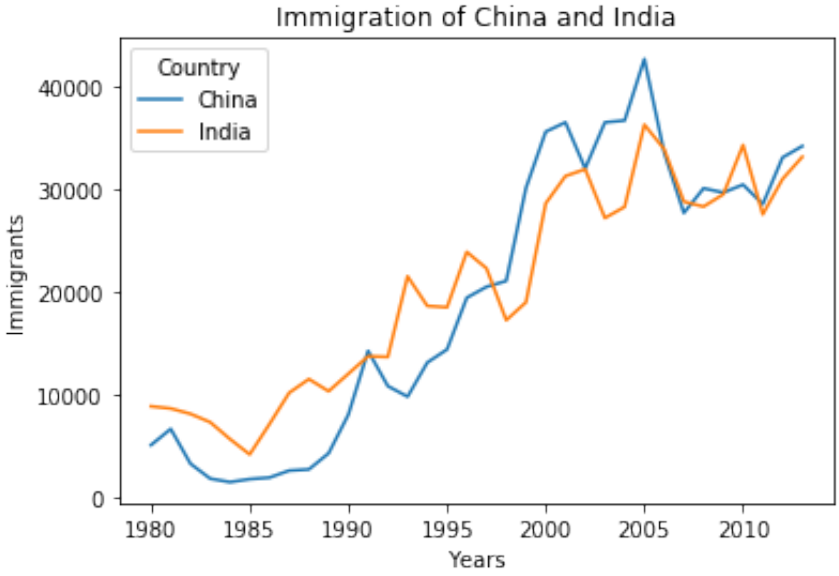
```
years = list(map(str,range(1980,2014)))
india_imgt = canada.loc['India', years]
india_imgt.plot();
india_imgt.plot(kind='line');
plt.title('Immigration from India');
plt.ylabel('Number of immigrants');
plt.xlabel('Years');
```



Plotting two countries and comparing

In [34]:

```
canada.loc[['China','India'], years].transpose().plot();
plt.title('Immigration of China and India');
plt.xlabel('Years');
plt.ylabel('Immigrants');
```



Assignment

Which two countries have similar immigration trends over the years 1980-2013?

In [118]:

```
canada['total_immigration'] = canada.sum(axis='columns')
canada.sort_values(by='total_immigration', ascending=False, axis='index', inpl
ace=True)
canada.head(10)
```

Out[118]:

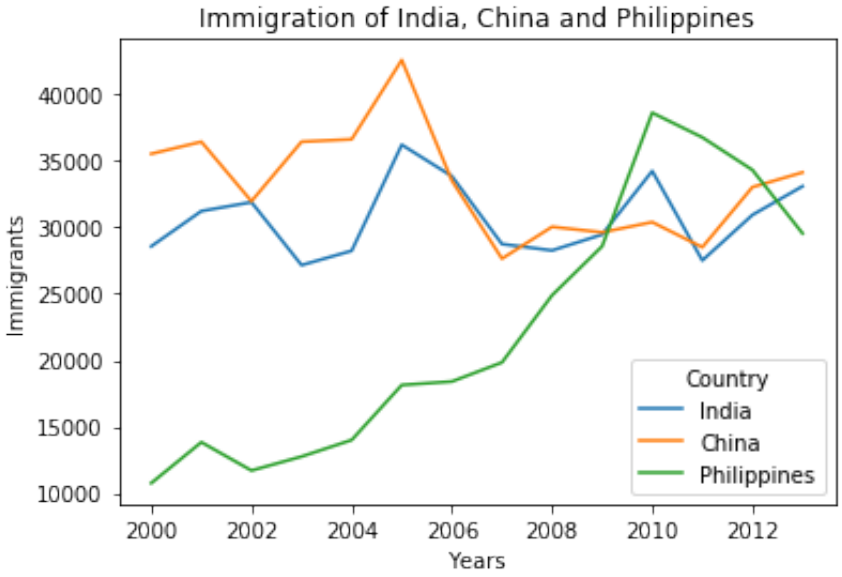
Country	Continent	Continent-Region	DevName	1980	1981	1982	1983	1984	1985	1
India	Asia	Southern Asia	Developing regions	8880	8670	8147	7338	5704	4211	7
China	Asia	Eastern Asia	Developing regions	5123	6682	3308	1863	1527	1816	1
United Kingdom of Great Britain and Northern Ireland	Europe	Northern Europe	Developed regions	22045	24796	20620	10015	10170	9564	9
Philippines	Asia	South-Eastern Asia	Developing regions	6051	5921	5249	4562	3801	3150	4
Pakistan	Asia	Southern Asia	Developing regions	978	972	1201	900	668	514	
United States of America	Northern America	Northern America	Developed regions	9378	10030	9074	7100	6661	6543	7
Iran (Islamic Republic of)	Asia	Southern Asia	Developing regions	1172	1429	1822	1592	1977	1648	1
Sri Lanka	Asia	Southern Asia	Developing regions	185	371	290	197	1086	845	1
Republic of Korea	Asia	Eastern Asia	Developing regions	1011	1456	1572	1081	847	962	1
Poland	Europe	Eastern Europe	Developed regions	863	2930	5881	4546	3588	2819	4

10 rows × 38 columns

Plotting three similar countries after year 2009 in the top 10 list.

In [124]:

```
canada.loc[['India','China','Philippines'], years].transpose().plot();
plt.title('Immigration of India, China and Philippines');
plt.xlabel('Years');
plt.ylabel('Immigrants');
```



Since 2009, India China and Philippines have shown similar trends.

As it is impossible to compare 195 countries and their charts, Last 3 year's average is taken and compared in the following steps:

Taking avg. of last 3 years and cleaning the Series

In [203]:

```
canada['avglast3'] = (canada['2011']+canada['2012']+canada['2013'])/3
l3 = canada['avglast3']
l3 = l3.sort_values(ascending = False )
l3 = l3.round()
l3 = l3[l3>2]
```

In [243]:

```
l3.head(10)
```

Out[243]:

```
Country
Philippines                                                  33541.0
China                                                        31885.0
India                                                        30510.0
Pakistan                                                     10433.0
Iran (Islamic Republic of)                                    8768.0
United States of America                                      8023.0
United Kingdom of Great Britain and Northern Ireland          6075.0
Haiti                                                         5508.0
France                                                        5328.0
Iraq                                                          5052.0
Name: avglast3, dtype: float64
```

Removing outliers

In [283]:

```
l5 = l3[l3<11000];
```

Calculating Standard Deviation for bins

In [282]:

```
np.std(l5)
```

Out[282]:

```
1629.5934270734047
```

Creating bins

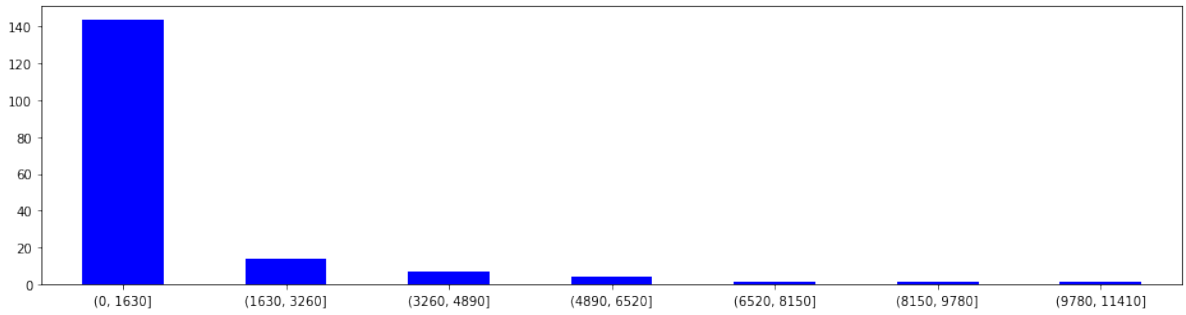
In [237]:

```
type(l5)
dl5 = df(l5)
dl5
ahg = pd.Series(range(0,12000,1630))
bins = ahg
dl5['binned'] = pd.cut(dl5['avglast3'], bins)
```

Plotting the bar graph

In [238]:

```
out = dl5['binned']
ax = out.value_counts(sort=False).plot.bar(rot=0, color="b", figsize=(16,4))
plt.show()
```



Bin values

In [271]:

```
dl5['binned'].value_counts(sort=True)
```

Out[271]:

```
(0, 1630]        144
(1630, 3260]      14
(3260, 4890]       7
(4890, 6520]       4
(9780, 11410]      1
(8150, 9780]       1
(6520, 8150]       1
Name: binned, dtype: int64
```

Countries showing similar trends (2011-2013), grouped

In [280]:

```
print('(9780, 11410]')
print(' {ho} {jo}'.format(ho = l5.index[0], jo = l5[0]))
print('')
print('')
print('(8150, 9780]')
print(' {ho} {jo}'.format(ho = l5.index[1], jo = l5[1]))
print('')
print('')
print('(6520, 8150]')
print(' {ho} {jo}'.format(ho = l5.index[2], jo = l5[2]))
print('')
print('')
print('(4890, 6520]')
print(' {jo}'.format(jo = l5[3:7]))
print('')
print('')
print('(3260, 4890]')
print(' {jo}'.format(jo = l5[7:14]))
print('')
print('')
print('(1630, 3260]')
print(' {jo}'.format(jo = l5[14:28]))
print('')
print('')
print('(0, 1630]')
print(' {jo}'.format(jo = l5[28:172]))
```

```
(9780, 11410]
 Pakistan 10433.0


(8150, 9780]
 Iran (Islamic Republic of) 8768.0


(6520, 8150]
 United States of America 8023.0


(4890, 6520]
 Country
United Kingdom of Great Britain and Northern Ireland    6075.0
Haiti                                                    5508.0
France                                                   5328.0
Iraq                                                     5052.0
Name: avglast3, dtype: float64


(3260, 4890]
 Country
Republic of Korea    4804.0
```

```
Egypt                4794.0
Algeria              4143.0
Mexico               4057.0
Colombia             3913.0
Morocco              3846.0
Nigeria              3573.0
Name: avglast3, dtype: float64


(1630, 3260]
 Country
Bangladesh                          3041.0
Sri Lanka                           3014.0
Ukraine                             2422.0
Lebanon                             2286.0
Afghanistan                         2281.0
Jamaica                             2240.0
Cameroon                            2195.0
Russian Federation                  2169.0
Israel                              2016.0
Ethiopia                            1878.0
Viet Nam                            1855.0
Somalia                             1715.0
Democratic Republic of the Congo    1663.0
Germany                             1657.0
Name: avglast3, dtype: float64


(0, 1630]
 Country
Romania              1625.0
Brazil               1621.0
Tunisia              1524.0
Eritrea              1412.0
Jordan               1365.0
                      ...
Equatorial Guinea       6.0
Mozambique              6.0
Brunei Darussalam       5.0
Cabo Verde              4.0
Tonga                   3.0
Name: avglast3, Length: 144, dtype: float64
```

In [ ]:

In [ ]: