DV0101EN-2-2-1-Area-Plots-Histograms-and-Bar-Charts-py-v2.0

October 10, 2019

Area Plots, Histograms, and Bar Plots

0.1 Introduction

In this lab, we will continue exploring the Matplotlib library and will learn how to create additional plots, namely area plots, histograms, and bar charts.

0.2 Table of Contents

1 Exploring Datasets with *pandas* and Matplotlib

Toolkits: The course heavily relies on *pandas* and **Numpy** for data wrangling, analysis, and visualization. The primary plotting library that we are exploring in the course is Matplotlib.

Dataset: Immigration to Canada from 1980 to 2013 - International migration flows to and from selected countries - The 2015 revision from United Nation's website.

The dataset contains annual data on the flows of international migrants as recorded by the countries of destination. The data presents both inflows and outflows according to the place of birth, citizenship or place of previous / next residence both for foreigners and nationals. For this lesson, we will focus on the Canadian Immigration data.

2 Downloading and Prepping Data

Import Primary Modules. The first thing we'll do is import two key data analysis modules: *pandas* and **Numpy**.

[1]:
```
import numpy as np  # useful for many scientific computing in Python
import pandas as pd # primary data structure library
```

[2]:
```
pip install xlrd
```

```
Collecting xlrd
  Downloading https://files.pythonhosted.org/packages/b0/16/63576a1a001752
e34bf8ea62e367997530dc553b689356b9879339cf45a4/xlrd-1.2.0-py2.py3-none-any.whl
(103kB)
    |                | 112kB 2.2MB/s eta 0:00:01
Installing collected packages: xlrd
Successfully installed xlrd-1.2.0
Note: you may need to restart the kernel to use updated packages.
```

Let's download and import our primary Canadian Immigration dataset using *pandas* `read_excel()` method. Normally, before we can do that, we would need to download a module which *pandas* requires to read in excel files. This module is **xlrd**. For your convenience, we have pre-installed this module, so you would not have to worry about that. Otherwise, you would need to run the following line of code to install the **xlrd** module:

```
!conda install -c anaconda xlrd --yes
```

Download the dataset and read it into a *pandas* dataframe.

[3]:
```
df_can = 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
                      )

print('Data downloaded and read into a dataframe!')
```

```
Data downloaded and read into a dataframe!
```

Let's take a look at the first five items in our dataset.

[4]:
```
df_can.head()
```

[4]:

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

```
1   702   560   716   561   539   620   603
2  3623  4005  5393  4752  4325  3774  4331
3     0     0     0     0     0     0     0
4     1     0     0     0     0     1     1

[5 rows x 43 columns]
```

Let's find out how many entries there are in our dataset.

[5]:
```
# print the dimensions of the dataframe
print(df_can.shape)
```

```
(195, 43)
```

Clean up data. We will make some modifications to the original dataset to make it easier to create our visualizations. Refer to `Introduction to Matplotlib and Line Plots` lab for the rational and detailed description of the changes.

1. Clean up the dataset to remove columns that are not informative to us for visualization (eg. Type, AREA, REG).

[6]:
```
df_can.drop(['AREA', 'REG', 'DEV', 'Type', 'Coverage'], axis=1, inplace=True)

# let's view the first five elements and see how the dataframe was changed
df_can.head()
```

[6]:
```
           OdName AreaName          RegName             DevName  1980  1981  \
0     Afghanistan     Asia    Southern Asia  Developing regions    16    39
1         Albania   Europe  Southern Europe   Developed regions     1     0
2         Algeria   Africa  Northern Africa  Developing regions    80    67
3  American Samoa  Oceania        Polynesia  Developing regions     0     1
4         Andorra   Europe  Southern Europe   Developed regions     0     0

   1982  1983  1984  1985  …  2004  2005  2006  2007  2008  2009  2010  \
0    39    47    71   340  …  2978  3436  3009  2652  2111  1746  1758
1     0     0     0     0  …  1450  1223   856   702   560   716   561
2    71    69    63    44  …  3616  3626  4807  3623  4005  5393  4752
3     0     0     0     0  …     0     0     1     0     0     0     0
4     0     0     0     0  …     0     0     1     1     0     0     0

   2011  2012  2013
0  2203  2635  2004
1   539   620   603
2  4325  3774  4331
3     0     0     0
4     0     1     1

[5 rows x 38 columns]
```

Notice how the columns Type, Coverage, AREA, REG, and DEV got removed from the dataframe.

2. Rename some of the columns so that they make sense.

[7]:
```
df_can.rename(columns={'OdName':'Country', 'AreaName':'Continent','RegName':
 ↪'Region'}, inplace=True)

# let's view the first five elements and see how the dataframe was changed
df_can.head()
```

[7]:

	Country	Continent	Region	DevName	1980	1981	1982	1983	1984	1985	…	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
0	Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	…	2978	3436	3009	2652	2111	1746	1758	2203	2635	2004
1	Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	…	1450	1223	856	702	560	716	561	539	620	603
2	Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	…	3616	3626	4807	3623	4005	5393	4752	4325	3774	4331
3	American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	…	0	0	1	0	0	0	0	0	0	0
4	Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	…	0	0	1	1	0	0	0	0	1	1

[5 rows x 38 columns]

Notice how the column names now make much more sense, even to an outsider.

3. For consistency, ensure that all column labels of type string.

[8]:
```
# let's examine the types of the column labels
all(isinstance(column, str) for column in df_can.columns)
```

[8]:
```
False
```

Notice how the above line of code returned *False* when we tested if all the column labels are of type **string**. So let's change them all to **string** type.

[9]:
```
df_can.columns = list(map(str, df_can.columns))

# let's check the column labels types now
all(isinstance(column, str) for column in df_can.columns)
```

[9]:
```
True
```

4. Set the country name as index - useful for quickly looking up countries using .loc method.

[10]:
```
df_can.set_index('Country', inplace=True)

# let's view the first five elements and see how the dataframe was changed
df_can.head()
```

[10]:

Country	Continent	Region	DevName	1980	1981	1982	1983	1984	1985	1986	…	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496	…	2978	3436	3009	2652	2111	1746	1758	2203	2635	2004
Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	1	…	1450	1223	856	702	560	716	561	539	620	603
Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69	…	3616	3626	4807	3623	4005	5393	4752	4325	3774	4331
American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	0	…	0	0	1	0	0	0	0	0	0	0
Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	2	…	0	0	1	1	0	0	0	0	1	1

[5 rows x 37 columns]

[11]:
```
df_can.head()
```

[11]:

Country	Continent	Region	DevName	1980	1981	1982	1983	1984	1985	1986	…	2004	2005	2006	2007
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496	…	2978	3436	3009	2652
Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	1	…	1450	1223	856	702
Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69	…	3616	3626	4807	3623
American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	0	…	0	0	1	0
Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	2	…	0	0	1	1

```
Afghanistan       39    47    71   340   496  …  2978  3436  3009  2652
Albania            0     0     0     0     1  …  1450  1223   856   702
Algeria           71    69    63    44    69  …  3616  3626  4807  3623
American Samoa     0     0     0     0     0  …     0     0     1     0
Andorra            0     0     0     0     2  …     0     0     1     1

                2008  2009  2010  2011  2012  2013
Country
Afghanistan     2111  1746  1758  2203  2635  2004
Albania          560   716   561   539   620   603
Algeria         4005  5393  4752  4325  3774  4331
American Samoa     0     0     0     0     0     0
Andorra            0     0     0     0     1     1

[5 rows x 37 columns]
```

Notice how the country names now serve as indices.

5. Add total column.

[12]:
```
df_can['Total'] = df_can.sum(axis=1)

# let's view the first five elements and see how the dataframe was changed
df_can.head()
```

[12]:
```
               Continent           Region             DevName  1980  1981  \
Country
Afghanistan         Asia    Southern Asia  Developing regions    16    39
Albania           Europe  Southern Europe   Developed regions     1     0
Algeria           Africa  Northern Africa  Developing regions    80    67
American Samoa   Oceania        Polynesia  Developing regions     0     1
Andorra           Europe  Southern Europe   Developed regions     0     0

                1982  1983  1984  1985  1986  …  2005  2006  2007  2008  \
Country                                       …
Afghanistan       39    47    71   340   496  …  3436  3009  2652  2111
Albania            0     0     0     0     1  …  1223   856   702   560
Algeria           71    69    63    44    69  …  3626  4807  3623  4005
American Samoa     0     0     0     0     0  …     0     1     0     0
Andorra            0     0     0     0     2  …     0     1     1     0

                2009  2010  2011  2012  2013  Total
Country
Afghanistan     1746  1758  2203  2635  2004  58639
Albania          716   561   539   620   603  15699
Algeria         5393  4752  4325  3774  4331  69439
American Samoa     0     0     0     0     0      6
Andorra            0     0     0     1     1     15
```

```
[5 rows x 38 columns]
```

Now the dataframe has an extra column that presents the total number of immigrants from each country in the dataset from 1980 - 2013. So if we print the dimension of the data, we get:

[13]:
```
print ('data dimensions:', df_can.shape)
```

```
data dimensions: (195, 38)
```

So now our dataframe has 38 columns instead of 37 columns that we had before.

[14]:
```
# finally, let's create a list of years from 1980 - 2013
# this will come in handy when we start plotting the data
years = list(map(str, range(1980, 2014)))

years
```

[14]:
```
['1980',
 '1981',
 '1982',
 '1983',
 '1984',
 '1985',
 '1986',
 '1987',
 '1988',
 '1989',
 '1990',
 '1991',
 '1992',
 '1993',
 '1994',
 '1995',
 '1996',
 '1997',
 '1998',
 '1999',
 '2000',
 '2001',
 '2002',
 '2003',
 '2004',
 '2005',
 '2006',
 '2007',
 '2008',
 '2009',
```

```
 '2010',
 '2011',
 '2012',
 '2013']
```

3 Visualizing Data using Matplotlib

Import `Matplotlib` and **Numpy**.

[15]:
```
# use the inline backend to generate the plots within the browser
%matplotlib inline

import matplotlib as mpl
import matplotlib.pyplot as plt

mpl.style.use('ggplot') # optional: for ggplot-like style

# check for latest version of Matplotlib
print ('Matplotlib version: ', mpl.__version__) # >= 2.0.0
```

```
Matplotlib version:  3.1.1
```

4 Area Plots

In the last module, we created a line plot that visualized the top 5 countries that contribued the most immigrants to Canada from 1980 to 2013. With a little modification to the code, we can visualize this plot as a cumulative plot, also knows as a **Stacked Line Plot** or **Area plot**.

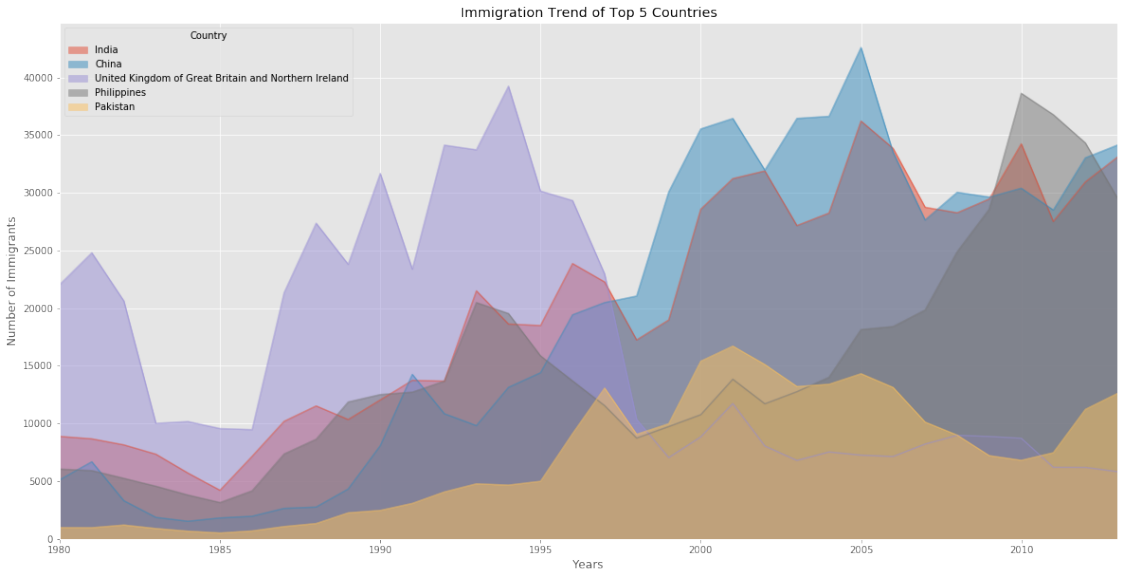
[16]:
```
df_can.sort_values(['Total'], ascending=False, axis=0, inplace=True)

# get the top 5 entries
df_top5 = df_can.head()

# transpose the dataframe
df_top5 = df_top5[years].transpose()

df_top5.head()
```

[16]:
```
Country  India  China  United Kingdom of Great Britain and Northern Ireland  \
1980      8880   5123                                                22045
1981      8670   6682                                                24796
1982      8147   3308                                                20620
1983      7338   1863                                                10015
1984      5704   1527                                                10170

Country  Philippines  Pakistan
1980            6051       978
```

1981	5921	972
1982	5249	1201
1983	4562	900
1984	3801	668

Area plots are stacked by default. And to produce a stacked area plot, each column must be either all positive or all negative values (any NaN values will defaulted to 0). To produce an unstacked plot, pass `stacked=False`.

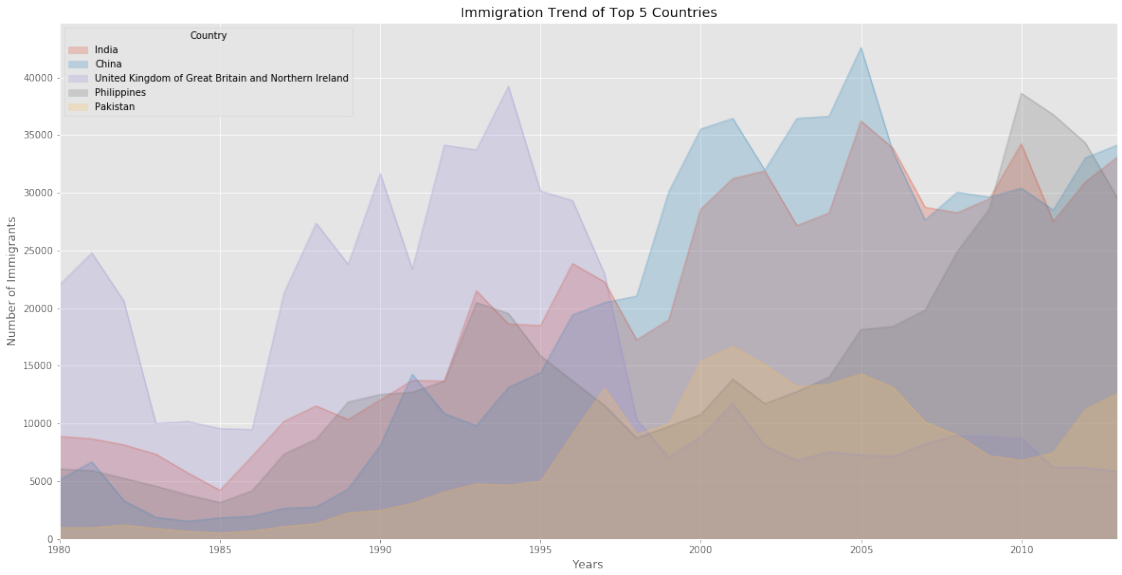
[17]:
```
df_top5.index = df_top5.index.map(int) # let's change the index values of␣
↪df_top5 to type integer for plotting
df_top5.plot(kind='area',
             stacked=False,
             figsize=(20, 10), # pass a tuple (x, y) size
             )

plt.title('Immigration Trend of Top 5 Countries')
plt.ylabel('Number of Immigrants')
plt.xlabel('Years')

plt.show()
```



The unstacked plot has a default transparency (alpha value) at 0.5. We can modify this value by passing in the `alpha` parameter.

[18]:
```
df_top5.plot(kind='area',
             alpha=0.25, # 0-1, default value a= 0.5
             stacked=False,
```

```
             figsize=(20, 10),
            )

plt.title('Immigration Trend of Top 5 Countries')
plt.ylabel('Number of Immigrants')
plt.xlabel('Years')

plt.show()
```



4.0.1 Two types of plotting

As we discussed in the video lectures, there are two styles/options of ploting with `matplotlib`. Plotting using the Artist layer and plotting using the scripting layer.

Option 1: Scripting layer (procedural method) - using matplotlib.pyplot as 'plt'

You can use `plt` i.e. `matplotlib.pyplot` and add more elements by calling different methods procedurally; for example, `plt.title(...)` to add title or `plt.xlabel(...)` to add label to the x-axis.

```
# Option 1: This is what we have been using so far
df_top5.plot(kind='area', alpha=0.35, figsize=(20, 10))
plt.title('Immigration trend of top 5 countries')
plt.ylabel('Number of immigrants')
plt.xlabel('Years')
```

Option 2: Artist layer (Object oriented method) - using an `Axes` instance from Matplotlib (preferred)

You can use an `Axes` instance of your current plot and store it in a variable (eg. `ax`). You can add more elements by calling methods with a little change in syntax (by adding "*set_*" to the

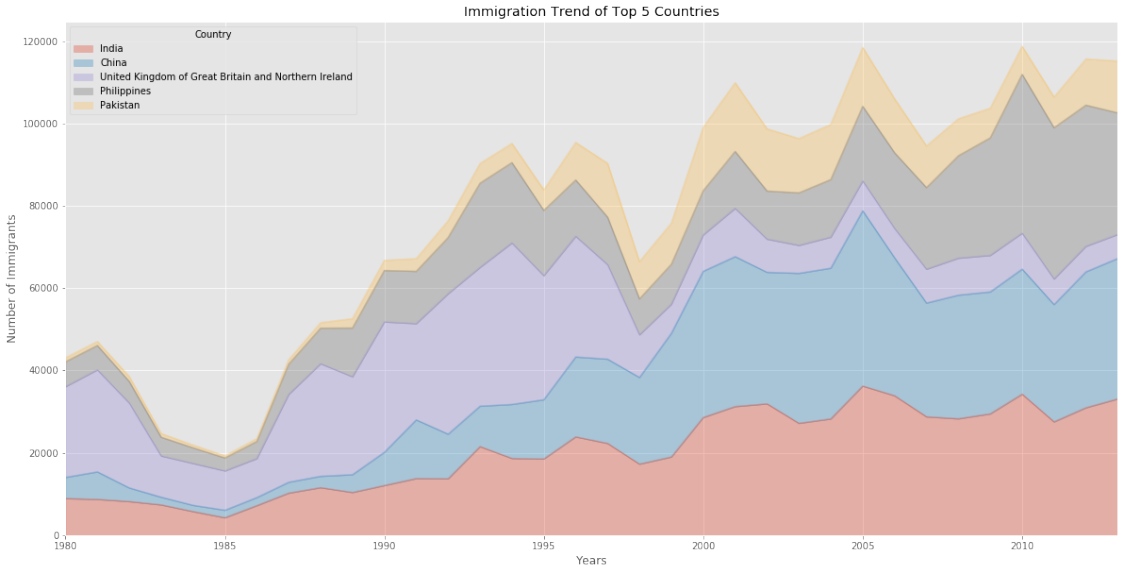
previous methods). For example, use `ax.set_title()` instead of `plt.title()` to add title, or `ax.set_xlabel()` instead of `plt.xlabel()` to add label to the x-axis.

This option sometimes is more transparent and flexible to use for advanced plots (in particular when having multiple plots, as you will see later).

In this course, we will stick to the **scripting layer**, except for some advanced visualizations where we will need to use the **artist layer** to manipulate advanced aspects of the plots.

[19]:
```
# option 2: preferred option with more flexibility
ax = df_top5.plot(kind='area', alpha=0.35, figsize=(20, 10))

ax.set_title('Immigration Trend of Top 5 Countries')
ax.set_ylabel('Number of Immigrants')
ax.set_xlabel('Years')
```

[19]:
```
Text(0.5, 0, 'Years')
```

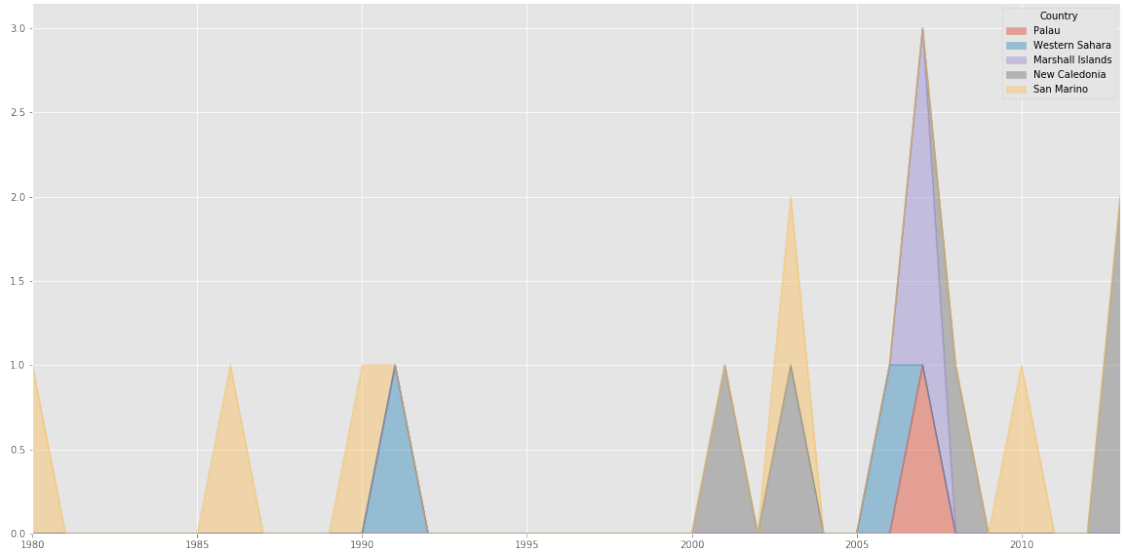


Question: Use the scripting layer to create a stacked area plot of the 5 countries that contributed the least to immigration to Canada **from** 1980 to 2013. Use a transparency value of 0.45.

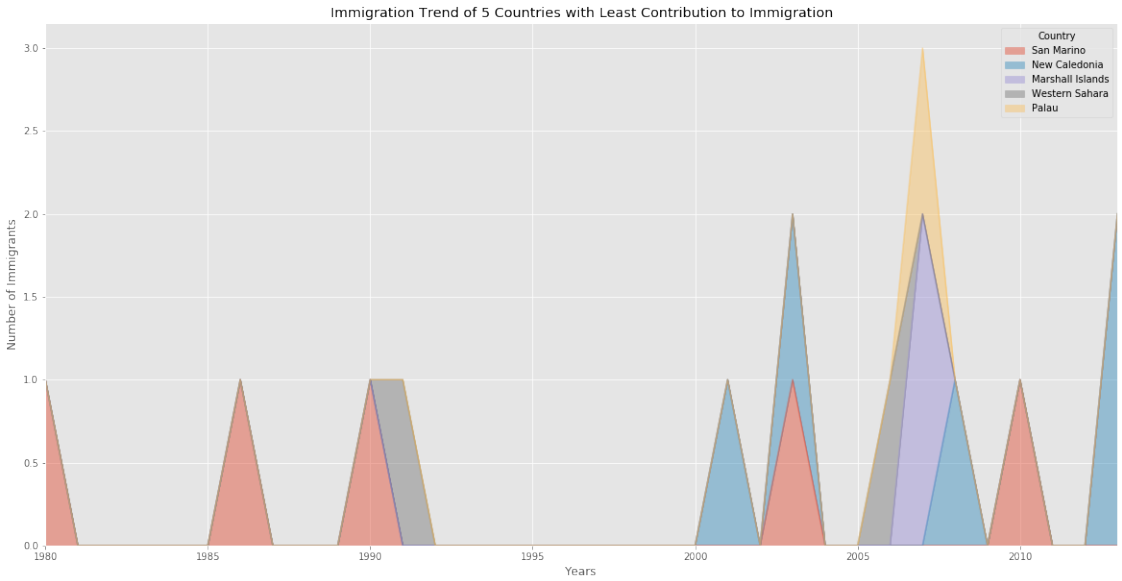
[32]:
```
### type your answer here
df_can=df_can.sort_values(by='Total',ascending=True, axis=0)
df_bot5=df_can.head(5)

df_bot5=df_bot5[years].transpose()

df_bot5.plot(kind='area',
             alpha=0.45,
             figsize=(20,10))
```

```
plt.ylabel='number of immigrants'
plt.xlabel='year'
plt.show
```

[32]: <function matplotlib.pyplot.show(*args, **kw)>



[20]:
```
df_least5 = df_can.tail(5)
 # transpose the dataframe
df_least5 = df_least5[years].transpose()
df_least5.head()

df_least5.index = df_least5.index.map(int) # let's change the index values of␣
 ↪df_least5 to type integer for plotting
df_least5.plot(kind='area', alpha=0.45, figsize=(20, 10))

plt.title('Immigration Trend of 5 Countries with Least Contribution to␣
 ↪Immigration')
plt.ylabel('Number of Immigrants')
plt.xlabel('Years')

plt.show()
```



[30]:
```
df_least5
```

[30]:

Country	Pakistan	Philippines	\
1980	978	6051	
1981	972	5921	
1982	1201	5249	
1983	900	4562	
1984	668	3801	
1985	514	3150	
1986	691	4166	
1987	1072	7360	
1988	1334	8639	
1989	2261	11865	
1990	2470	12509	
1991	3079	12718	
1992	4071	13670	
1993	4777	20479	
1994	4666	19532	
1995	4994	15864	
1996	9125	13692	
1997	13073	11549	
1998	9068	8735	
1999	9979	9734	
2000	15400	10763	
2001	16708	13836	
2002	15110	11707	
2003	13205	12758	
2004	13399	14004	

2005	14314	18139
2006	13127	18400
2007	10124	19837
2008	8994	24887
2009	7217	28573
2010	6811	38617
2011	7468	36765
2012	11227	34315
2013	12603	29544

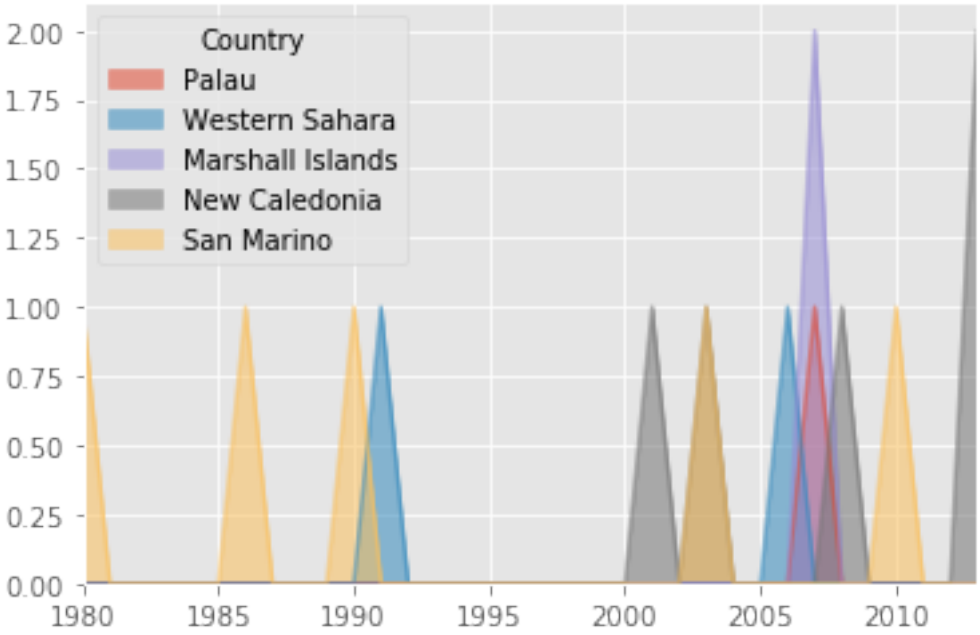
Country	United Kingdom of Great Britain and Northern Ireland	China	India
1980	22045	5123	8880
1981	24796	6682	8670
1982	20620	3308	8147
1983	10015	1863	7338
1984	10170	1527	5704
1985	9564	1816	4211
1986	9470	1960	7150
1987	21337	2643	10189
1988	27359	2758	11522
1989	23795	4323	10343
1990	31668	8076	12041
1991	23380	14255	13734
1992	34123	10846	13673
1993	33720	9817	21496
1994	39231	13128	18620
1995	30145	14398	18489
1996	29322	19415	23859
1997	22965	20475	22268
1998	10367	21049	17241
1999	7045	30069	18974
2000	8840	35529	28572
2001	11728	36434	31223
2002	8046	31961	31889
2003	6797	36439	27155
2004	7533	36619	28235
2005	7258	42584	36210
2006	7140	33518	33848
2007	8216	27642	28742
2008	8979	30037	28261
2009	8876	29622	29456
2010	8724	30391	34235
2011	6204	28502	27509
2012	6195	33024	30933
2013	5827	34129	33087

Double-click **here** for the solution.

Question: Use the artist layer to create an unstacked area plot of the 5 countries that contributed the least to immigration to Canada **from** 1980 to 2013. Use a transparency value of 0.55.

[36]:
```
### type your answer here
art=df_bot5.plot(kind='area',
                 stacked=False,
                 alpha=0.55)
art.set_ylabel='immigrants'
art.set_xlabel='years'
plt.show
```

[36]: <function matplotlib.pyplot.show(*args, **kw)>



Double-click **here** for the solution.

5 Histograms

A histogram is a way of representing the *frequency* distribution of numeric dataset. The way it works is it partitions the x-axis into *bins*, assigns each data point in our dataset to a bin, and then counts the number of data points that have been assigned to each bin. So the y-axis is the frequency or the number of data points in each bin. Note that we can change the bin size and usually one needs to tweak it so that the distribution is displayed nicely.

Question: What is the frequency distribution of the number (population) of new immigrants from the various countries to Canada in 2013?

Before we proceed with creating the histogram plot, let's first examine the data split into intervals.

To do this, we will us **Numpy**'s `histrogram` method to get the bin ranges and frequency counts as follows:

[21]:
```
# let's quickly view the 2013 data
df_can['2013'].head()
```

[21]:
```
Country
India                                                33087
China                                                34129
United Kingdom of Great Britain and Northern Ireland     5827
Philippines                                          29544
Pakistan                                             12603
Name: 2013, dtype: int64
```

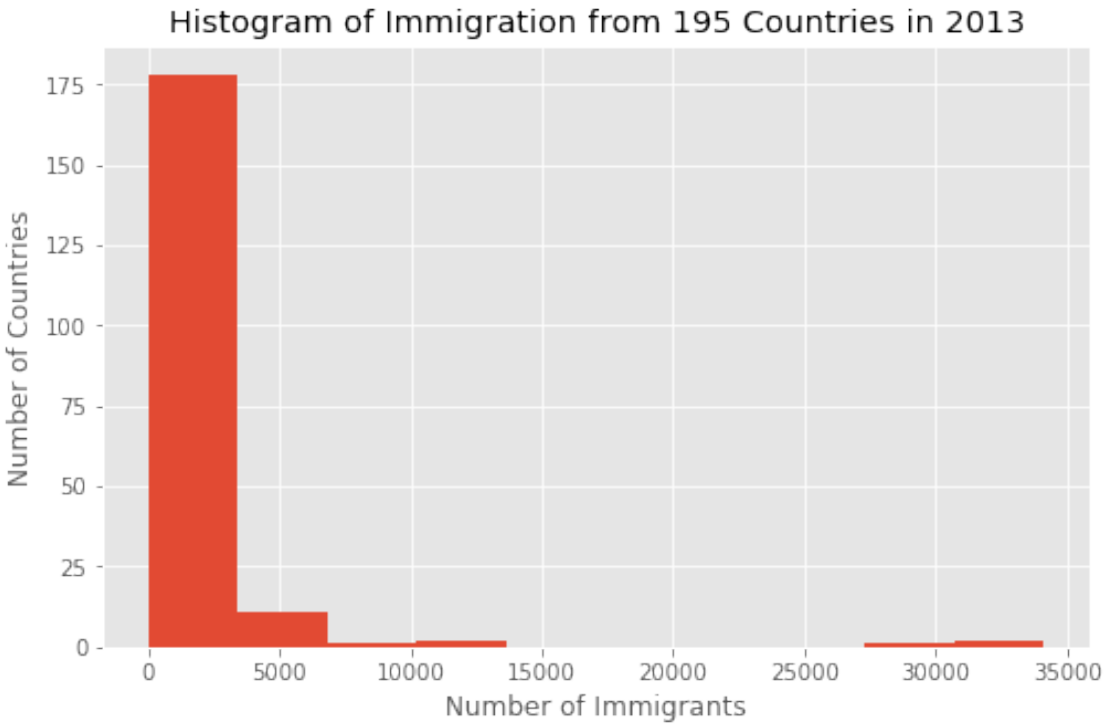
[22]:
```
# np.histogram returns 2 values
count, bin_edges = np.histogram(df_can['2013'])

print(count) # frequency count
print(bin_edges) # bin ranges, default = 10 bins
```

```
[178  11   1   2   0   0   0   0   1   2]
[    0.   3412.9  6825.8 10238.7 13651.6 17064.5 20477.4 23890.3 27303.2
 30716.1 34129. ]
```

By default, the `histrogram` method breaks up the dataset into 10 bins. The figure below summarizes the bin ranges and the frequency distribution of immigration in 2013. We can see that in 2013: * 178 countries contributed between 0 to 3412.9 immigrants * 11 countries contributed between 3412.9 to 6825.8 immigrants * 1 country contributed between 6285.8 to 10238.7 immigrants, and so on..

We can easily graph this distribution by passing `kind=hist` to `plot()`.

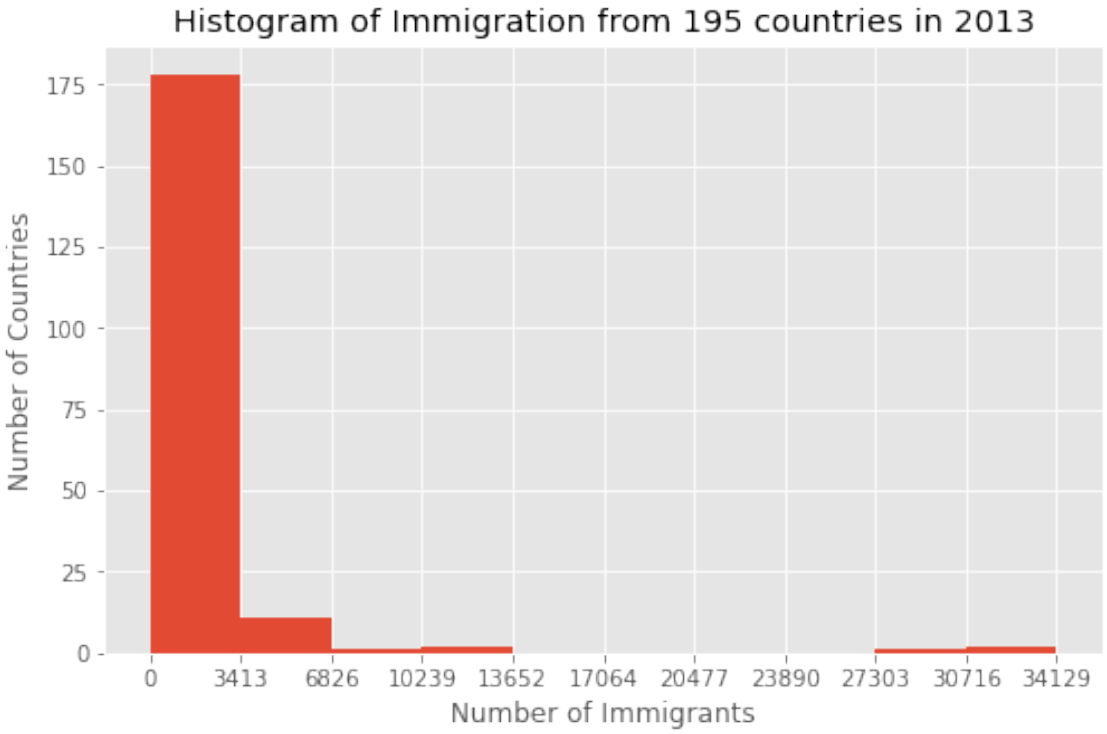
[23]:
```
df_can['2013'].plot(kind='hist', figsize=(8, 5))

plt.title('Histogram of Immigration from 195 Countries in 2013') # add a title␣
 ↪to the histogram
plt.ylabel('Number of Countries') # add y-label
plt.xlabel('Number of Immigrants') # add x-label

plt.show()
```



In the above plot, the x-axis represents the population range of immigrants in intervals of 3412.9. The y-axis represents the number of countries that contributed to the aforementioned population.

Notice that the x-axis labels do not match with the bin size. This can be fixed by passing in a `xticks` keyword that contains the list of the bin sizes, as follows:

[24]:
```
# 'bin_edges' is a list of bin intervals
count, bin_edges = np.histogram(df_can['2013'])

df_can['2013'].plot(kind='hist', figsize=(8, 5), xticks=bin_edges)

plt.title('Histogram of Immigration from 195 countries in 2013') # add a title␣
 ↪to the histogram
plt.ylabel('Number of Countries') # add y-label
plt.xlabel('Number of Immigrants') # add x-label

plt.show()
```



Side Note: We could use `df_can['2013'].plot.hist()`, instead. In fact, throughout this lesson, using `some_data.plot(kind='type_plot', ...)` is equivalent to `some_data.plot.type_plot(...)`. That is, passing the type of the plot as argument or method behaves the same.

See the *pandas* documentation for more info http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.plot.html.

We can also plot multiple histograms on the same plot. For example, let's try to answer the following questions using a histogram.

Question: What is the immigration distribution for Denmark, Norway, and Sweden for years 1980 - 2013?

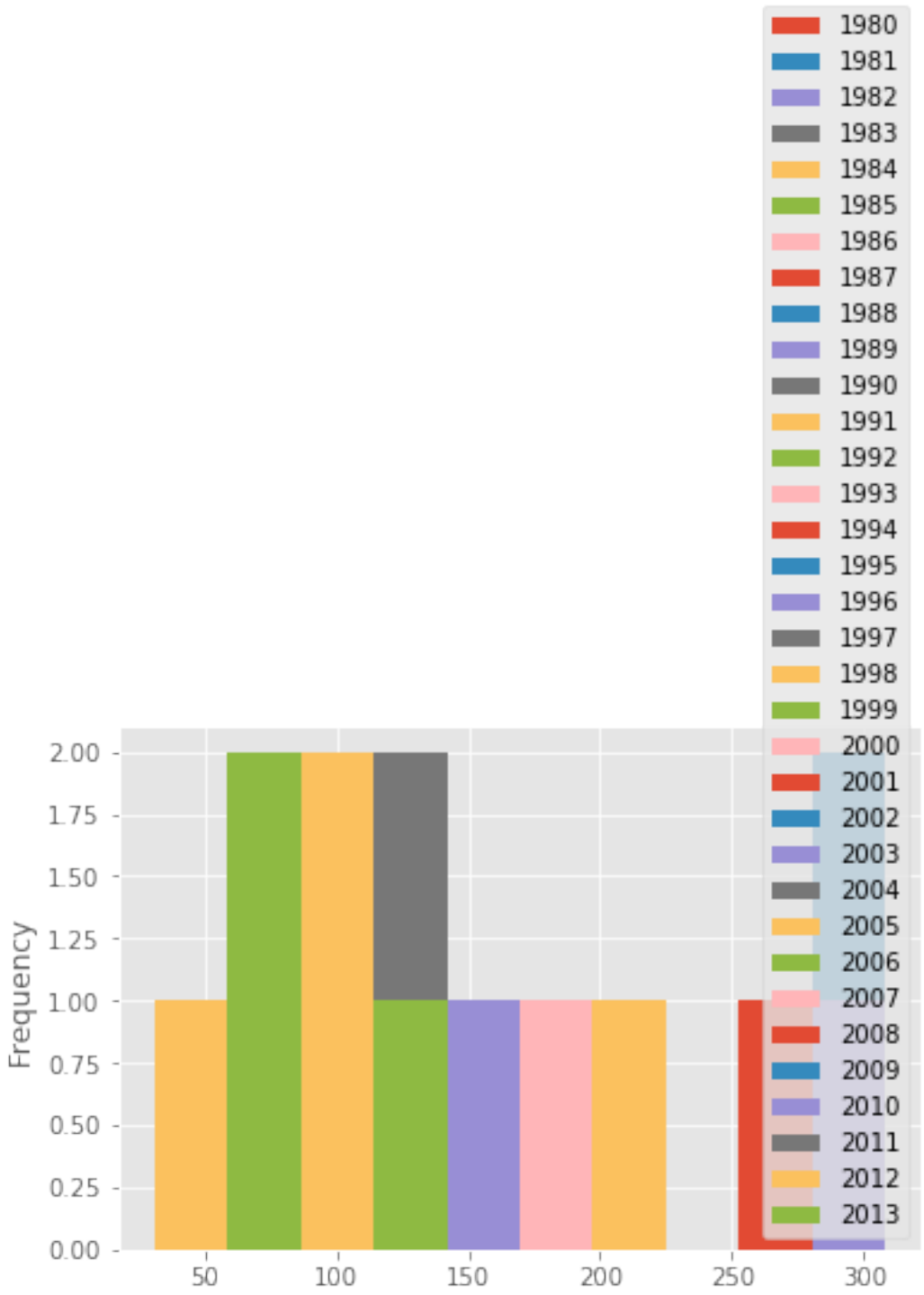
[25]:
```
# let's quickly view the dataset
df_can.loc[['Denmark', 'Norway', 'Sweden'], years]
```

[25]:
```
         1980  1981  1982  1983  1984  1985  1986  1987  1988  1989  … \
Country                                                              …
Denmark   272   293   299   106    93    73    93   109   129   129  …
Norway    116    77   106    51    31    54    56    80    73    76  …
Sweden    281   308   222   176   128   158   187   198   171   182  …

         2004  2005  2006  2007  2008  2009  2010  2011  2012  2013
Country
```

```
Denmark     89     62    101     97    108     81     92     93     94     81
Norway      73     57     53     73     66     75     46     49     53     59
Sweden     129    205    139    193    165    167    159    134    140    140

[3 rows x 34 columns]
```

[26]:
```
# generate histogram
df_can.loc[['Denmark', 'Norway', 'Sweden'], years].plot.hist()
```

[26]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3b7b68c550>
```



That does not look right!

Don't worry, you'll often come across situations like this when creating plots. The solution often lies in how the underlying dataset is structured.

Instead of plotting the population frequency distribution of the population for the 3 countries, *pandas* instead plotted the population frequency distribution for the `years`.

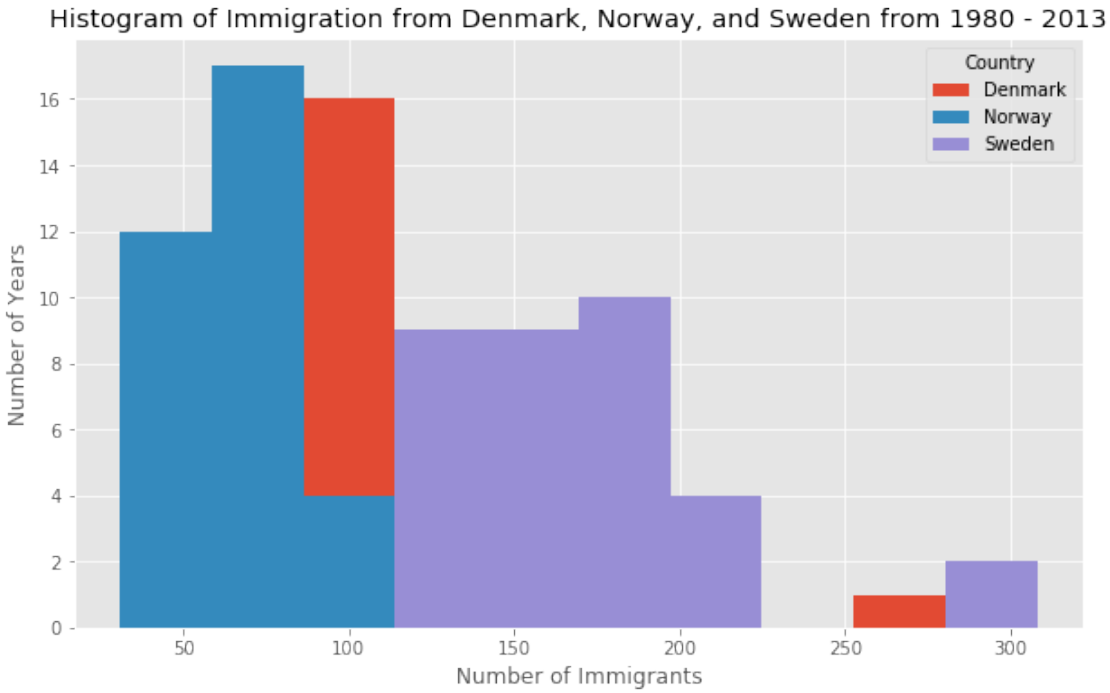
This can be easily fixed by first transposing the dataset, and then plotting as shown below.

[27]:
```
# transpose dataframe
df_t = df_can.loc[['Denmark', 'Norway', 'Sweden'], years].transpose()
df_t.head()
```

[27]:

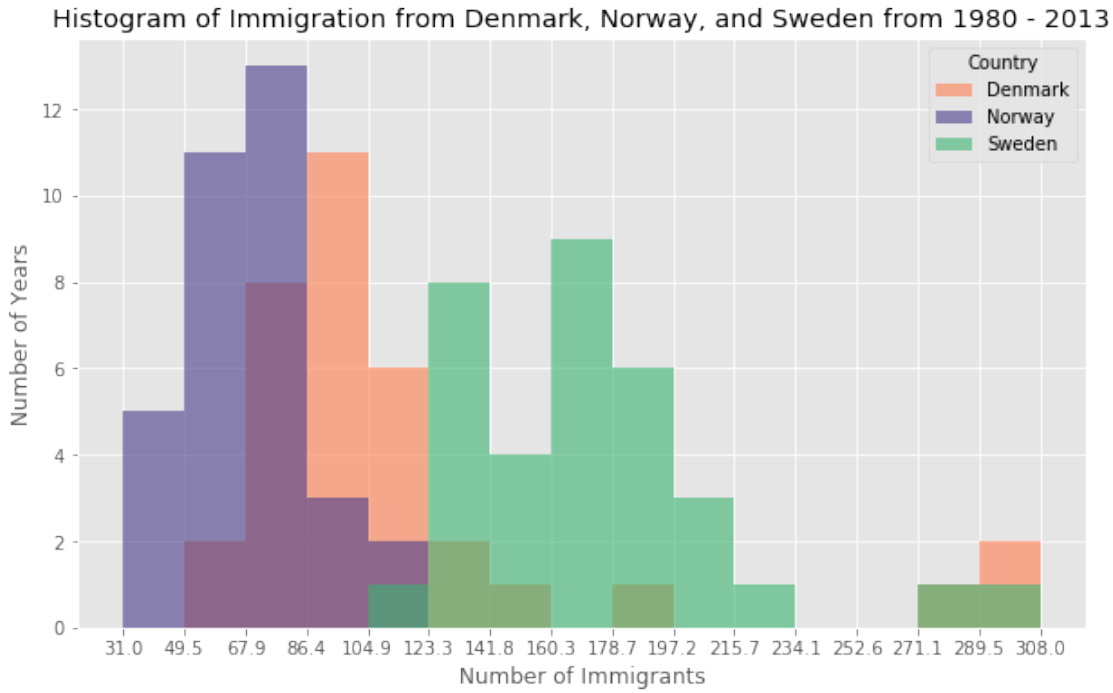
Country	Denmark	Norway	Sweden
1980	272	116	281
1981	293	77	308
1982	299	106	222
1983	106	51	176
1984	93	31	128

[28]:
```
# generate histogram
df_t.plot(kind='hist', figsize=(10, 6))

plt.title('Histogram of Immigration from Denmark, Norway, and Sweden from 1980␣
 ↪- 2013')
plt.ylabel('Number of Years')
plt.xlabel('Number of Immigrants')

plt.show()
```



Let's make a few modifications to improve the impact and aesthetics of the previous plot: * increase the bin size to 15 by passing in `bins` parameter * set transparency to 60% by passing in `alpha` paramemter * label the x-axis by passing in `x-label` paramater * change the colors of the plots by passing in `color` parameter

[29]:
```
# let's get the x-tick values
count, bin_edges = np.histogram(df_t, 15)

# un-stacked histogram
df_t.plot(kind ='hist',
          figsize=(10, 6),
          bins=15,
          alpha=0.6,
          xticks=bin_edges,
          color=['coral', 'darkslateblue', 'mediumseagreen']
         )

plt.title('Histogram of Immigration from Denmark, Norway, and Sweden from 1980␣
 ↪- 2013')
plt.ylabel('Number of Years')
plt.xlabel('Number of Immigrants')

plt.show()
```

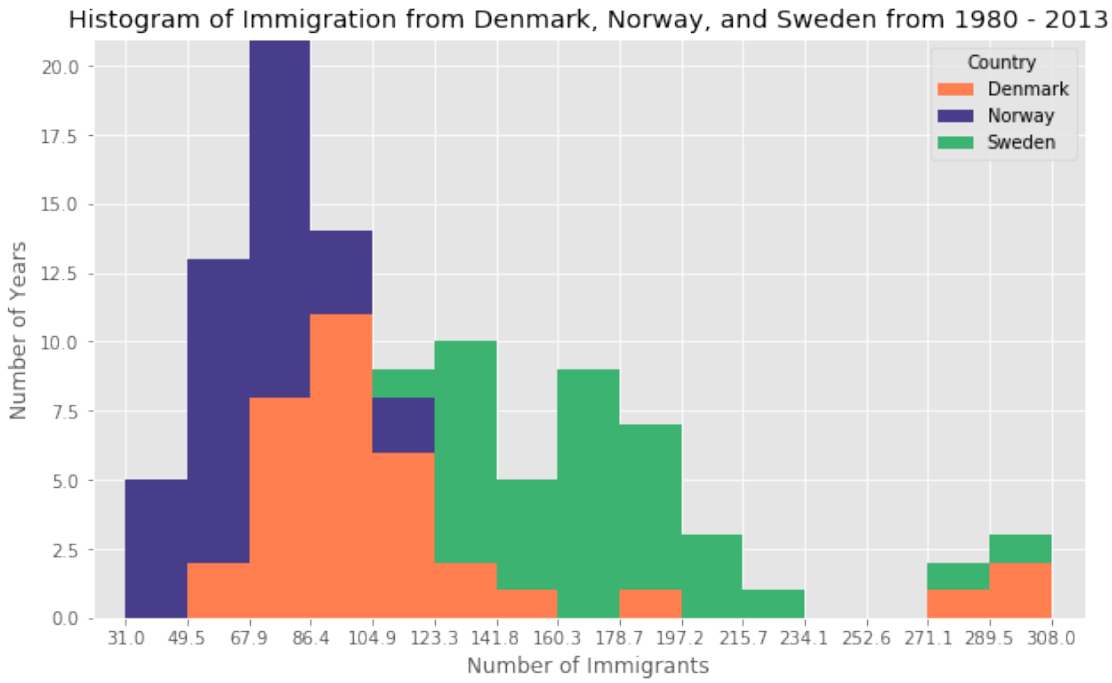


Tip: For a full listing of colors available in Matplotlib, run the following code in your python shell:

```
import matplotlib
for name, hex in matplotlib.colors.cnames.items():
    print(name, hex)
```

If we do no want the plots to overlap each other, we can stack them using the `stacked` paramemter. Let's also adjust the min and max x-axis labels to remove the extra gap on the edges of the plot. We can pass a tuple (min,max) using the `xlim` paramater, as show below.

[30]:
```
count, bin_edges = np.histogram(df_t, 15)
xmin = bin_edges[0] - 10   #  first bin value is 31.0, adding buffer of 10 for␣
 ↪aesthetic purposes
xmax = bin_edges[-1] + 10  #  last bin value is 308.0, adding buffer of 10 for␣
 ↪aesthetic purposes

# stacked Histogram
df_t.plot(kind='hist',
          figsize=(10, 6),
          bins=15,
          xticks=bin_edges,
          color=['coral', 'darkslateblue', 'mediumseagreen'],
          stacked=True,
          xlim=(xmin, xmax)
         )

plt.title('Histogram of Immigration from Denmark, Norway, and Sweden from 1980␣
 ↪- 2013')
plt.ylabel('Number of Years')
plt.xlabel('Number of Immigrants')

plt.show()
```

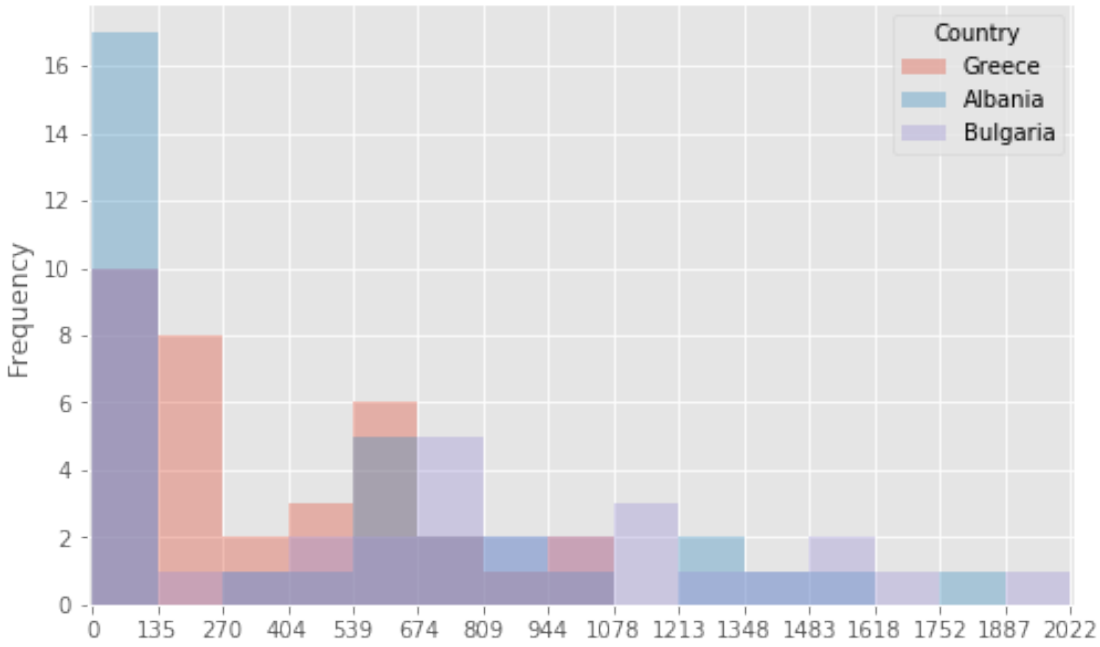


Question: Use the scripting layer to display the immigration distribution for Greece, Albania, and Bulgaria for years 1980 - 2013? Use an overlapping plot with 15 bins and a transparency value of 0.35.

[32]:
```
### type your answer here

df_greece=df_can.loc[['Greece','Albania','Bulgaria'],years].transpose()
# df_greece
```

[34]:
```
count, bin_edges=np.histogram(df_greece,15)
xmin=bin_edges[0]-10
xmax=bin_edges[-1]+10

df_greece.plot(kind='hist', bins=15, alpha=0.35, stacked=False,␣
 ↪xticks=bin_edges, xlim=(xmin, xmax), figsize=(8,5))
```

[34]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3b7b2cf4e0>
```



Double-click **here** for the solution.

6 Bar Charts (Dataframe)

A bar plot is a way of representing data where the *length* of the bars represents the magnitude/size of the feature/variable. Bar graphs usually represent numerical and categorical variables grouped in intervals.

To create a bar plot, we can pass one of two arguments via `kind` parameter in `plot()`:

- `kind=bar` creates a *vertical* bar plot
- `kind=barh` creates a *horizontal* bar plot

Vertical bar plot

In vertical bar graphs, the x-axis is used for labelling, and the length of bars on the y-axis corresponds to the magnitude of the variable being measured. Vertical bar graphs are particuarly useful in analyzing time series data. One disadvantage is that they lack space for text labelling at the foot of each bar.

Let's start off by analyzing the effect of Iceland's Financial Crisis:

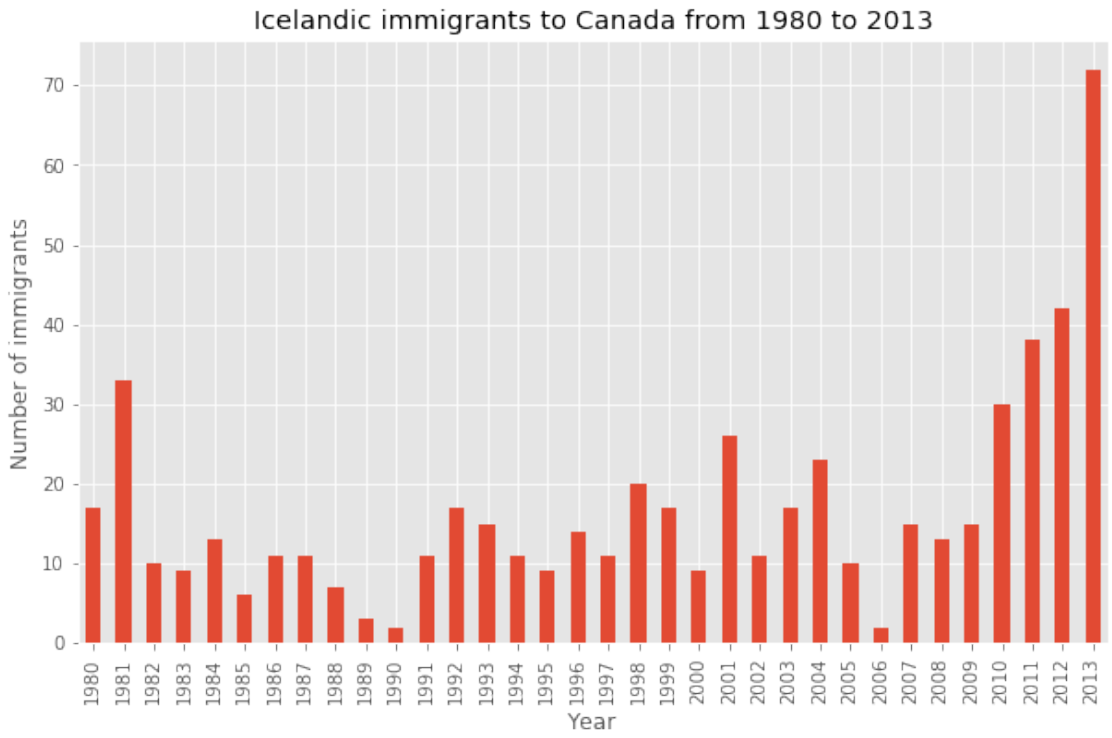
The 2008 - 2011 Icelandic Financial Crisis was a major economic and political event in Iceland. Relative to the size of its economy, Iceland's systemic banking collapse was the largest experienced by any country in economic history. The crisis led to a severe economic depression in 2008 - 2011 and significant political unrest.

Question: Let's compare the number of Icelandic immigrants (country = 'Iceland') to Canada from year 1980 to 2013.

[35]:
```
# step 1: get the data
df_iceland = df_can.loc['Iceland', years]
df_iceland.head()
```

[35]:
```
1980    17
1981    33
1982    10
1983     9
1984    13
Name: Iceland, dtype: object
```

[36]:
```
# step 2: plot data
df_iceland.plot(kind='bar', figsize=(10, 6))

plt.xlabel('Year') # add to x-label to the plot
plt.ylabel('Number of immigrants') # add y-label to the plot
plt.title('Icelandic immigrants to Canada from 1980 to 2013') # add title to␣
 ↪the plot

plt.show()
```



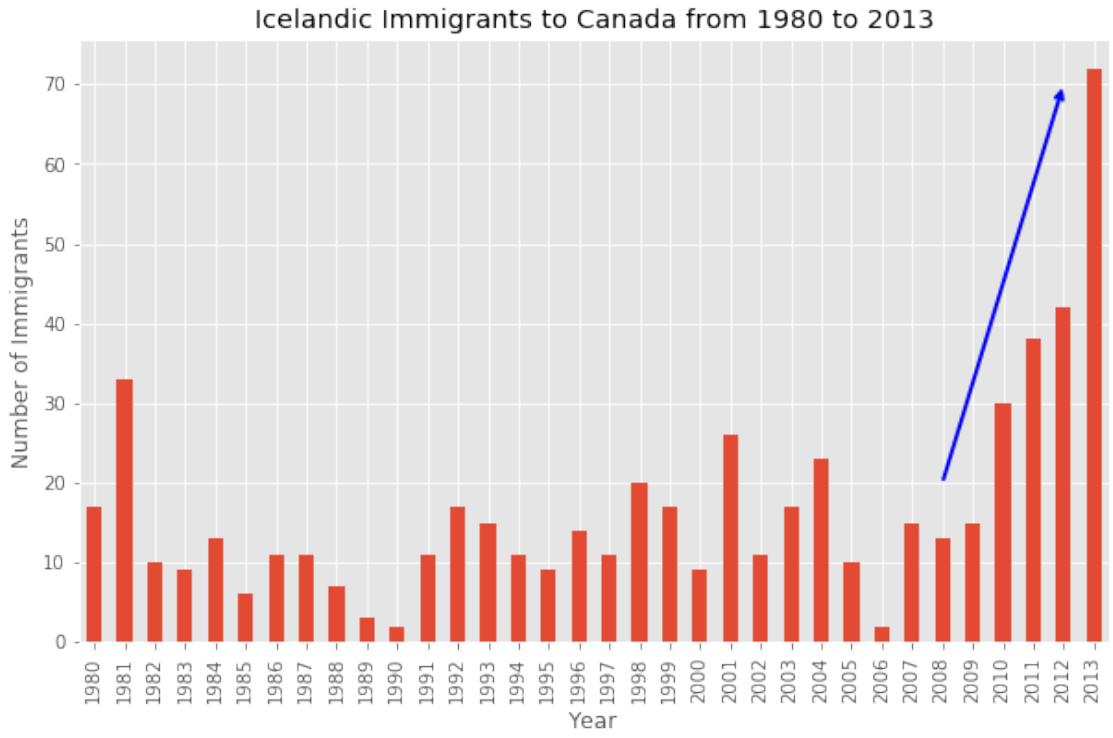
The bar plot above shows the total number of immigrants broken down by each year. We can clearly see the impact of the financial crisis; the number of immigrants to Canada started increasing rapidly

after 2008.

Let's annotate this on the plot using the `annotate` method of the **scripting layer** or the **pyplot interface**. We will pass in the following parameters: - `s`: str, the text of annotation. - `xy`: Tuple specifying the (x,y) point to annotate (in this case, end point of arrow). - `xytext`: Tuple specifying the (x,y) point to place the text (in this case, start point of arrow). - `xycoords`: The coordinate system that xy is given in - 'data' uses the coordinate system of the object being annotated (default). - `arrowprops`: Takes a dictionary of properties to draw the arrow: - `arrowstyle`: Specifies the arrow style, `'->'` is standard arrow. - `connectionstyle`: Specifies the connection type. `arc3` is a straight line. - `color`: Specifes color of arror. - `lw`: Specifies the line width.

I encourage you to read the Matplotlib documentation for more details on annotations: http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.annotate.

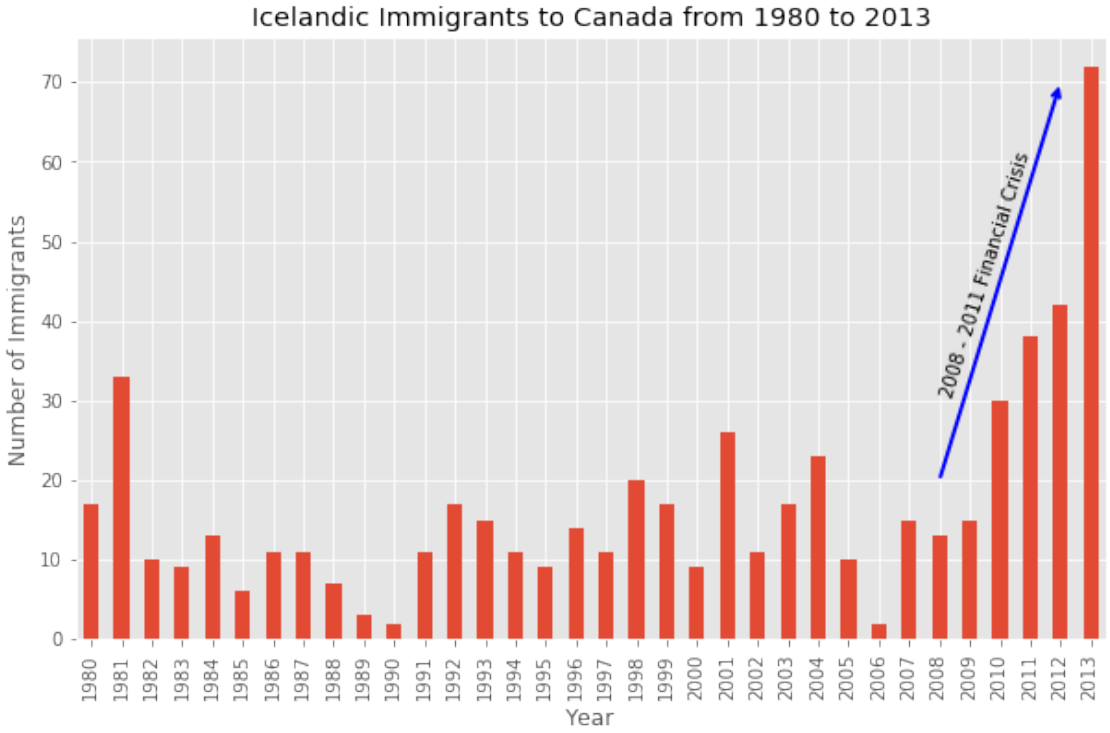
[37]:
```
df_iceland.plot(kind='bar', figsize=(10, 6), rot=90) # rotate the bars by 90␣
↪degrees

plt.xlabel('Year')
plt.ylabel('Number of Immigrants')
plt.title('Icelandic Immigrants to Canada from 1980 to 2013')

# Annotate arrow
plt.annotate('',                      # s: str. Will leave it blank for no text
             xy=(32, 70),             # place head of the arrow at point (year␣
↪2012 , pop 70)
             xytext=(28, 20),         # place base of the arrow at point (year␣
↪2008 , pop 20)
             xycoords='data',         # will use the coordinate system of the␣
↪object being annotated
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3',␣
↪color='blue', lw=2)
            )

plt.show()
```



Let's also annotate a text to go over the arrow. We will pass in the following additional parameters: - `rotation`: rotation angle of text in degrees (counter clockwise) - `va`: vertical alignment of text ['center' | 'top' | 'bottom' | 'baseline'] - `ha`: horizontal alignment of text ['center' | 'right' | 'left']

[38]:
```
df_iceland.plot(kind='bar', figsize=(10, 6), rot=90)

plt.xlabel('Year')
plt.ylabel('Number of Immigrants')
plt.title('Icelandic Immigrants to Canada from 1980 to 2013')

# Annotate arrow
plt.annotate('',                      # s: str. will leave it blank for no text
             xy=(32, 70),             # place head of the arrow at point (year␣
 ↪2012 , pop 70)
             xytext=(28, 20),         # place base of the arrow at point (year␣
 ↪2008 , pop 20)
             xycoords='data',         # will use the coordinate system of the␣
 ↪object being annotated
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3',␣
 ↪color='blue', lw=2)
            )

# Annotate Text
```

```
plt.annotate('2008 - 2011 Financial Crisis', # text to display
             xy=(28, 30),                    # start the text at at point (year␣
 ↪2008 , pop 30)
             rotation=72.5,                  # based on trial and error to␣
 ↪match the arrow
             va='bottom',                    # want the text to be vertically␣
 ↪'bottom' aligned
             ha='left',                      # want the text to be horizontally␣
 ↪'left' algned.
             )

plt.show()
```



Horizontal Bar Plot

Sometimes it is more practical to represent the data horizontally, especially if you need more room for labelling the bars. In horizontal bar graphs, the y-axis is used for labelling, and the length of bars on the x-axis corresponds to the magnitude of the variable being measured. As you will see, there is more room on the y-axis to label categetorical variables.

Question: Using the scripting layter and the `df_can` dataset, create a *horizontal* bar plot showing the *total* number of immigrants to Canada from the top 15 countries, for the period 1980 - 2013. Label each country with the total immigrant count.

Step 1: Get the data pertaining to the top 15 countries.

[45]:
```
### type your answer here
df_can.sort_values(by='Total', ascending=False, axis=0, inplace=True)
df_can_15=df_can.head(15)
df_can_15=df_can_15['Total']
df_can_15
```

[45]:
```
Country
India                                              691904
China                                              659962
United Kingdom of Great Britain and Northern Ireland    551500
Philippines                                        511391
Pakistan                                           241600
United States of America                           241122
Iran (Islamic Republic of)                         175923
Sri Lanka                                          148358
Republic of Korea                                  142581
Poland                                             139241
Lebanon                                            115359
France                                             109091
Jamaica                                            106431
Viet Nam                                            97146
Romania                                             93585
Name: Total, dtype: int64
```

Double-click **here** for the solution.

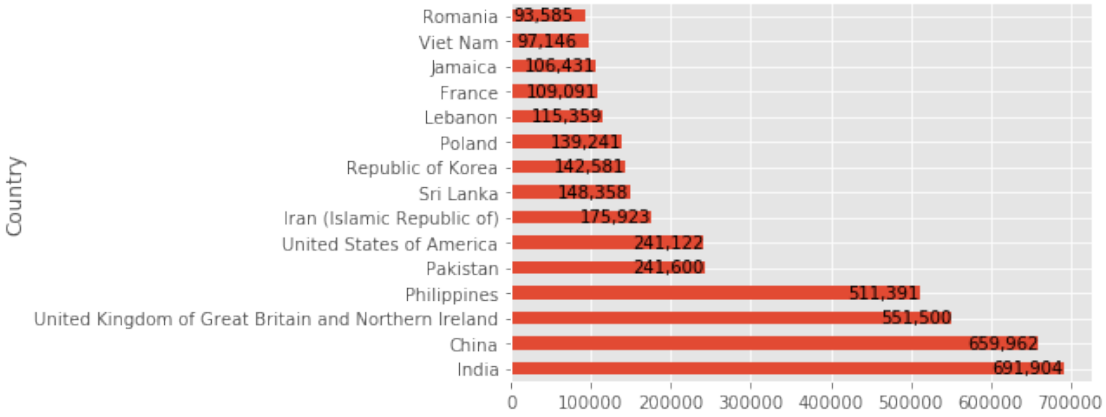
[51]:
```
df_can_15
```

[51]:
```
Country
India                                              691904
China                                              659962
United Kingdom of Great Britain and Northern Ireland    551500
Philippines                                        511391
Pakistan                                           241600
United States of America                           241122
Iran (Islamic Republic of)                         175923
Sri Lanka                                          148358
Republic of Korea                                  142581
Poland                                             139241
Lebanon                                            115359
France                                             109091
Jamaica                                            106431
Viet Nam                                            97146
Romania                                             93585
Name: Total, dtype: int64
```

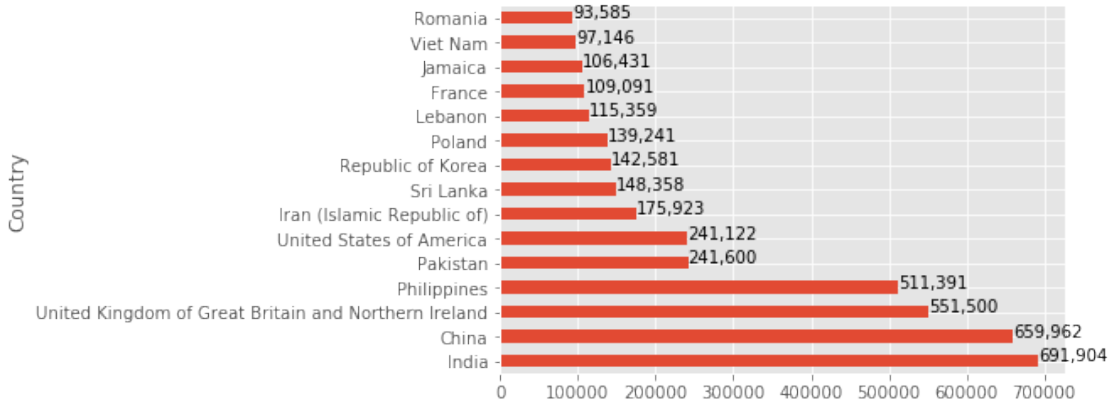
Step 2: Plot data: 1. Use `kind='barh'` to generate a bar chart with horizontal bars. 2. Make sure to choose a good size for the plot and to label your axes and to give the plot a title. 3. Loop

through the countries and annotate the immigrant population using the anotate function of the scripting interface.

[65]:
```
### type your answer here
df_can_15.plot(kind='barh')

for i, v in enumerate(df_can_15):
    label = format(int(v), ',')
    plt.annotate(label,
                 xy=(v-90000,i-0.2),
                 color='black')

plt.show()
```



[58]:
```
### type your answer here
df_can_15.plot(kind='barh')

for i, v in enumerate(df_can_15):
    label = format(int(v), ',')
    plt.annotate(label,
                 xy=(v,i),
                 color='black')

plt.show()
```



Double-click **here** for the solution.

6.0.1 Thank you for completing this lab!

This notebook was originally created by Jay Rajasekharan with contributions from Ehsan M. Kermani, and Slobodan Markovic.

This notebook was recently revamped by Alex Aklson. I hope you found this lab session interesting. Feel free to contact me if you have any questions!

This notebook is part of a course on **Coursera** called *Data Visualization with Python*. If you accessed this notebook outside the course, you can take this course online by clicking here.

Copyright © 2019 Cognitive Class. This notebook and its source code are released under the terms of the MIT License.