```
In [1]: import matplotlib as mpl
        %matplotlib inline
        import matplotlib.pyplot as plt

In [2]: import numpy as np
        import pandas as pd

In [3]: df_can = pd.read_excel('canada.xlsx', sheet_name = 'Canada by Citizenship',
                               skiprows = range(0,20),
                               skiprows = range(21,38),
                               usecols = 66)

        df_can.head()
```

	Type	Coverage	Software	AREA	Western NIS	Regional	NIS	Dorfman	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
0	Immigrants	Foreigners	Algeria	903	Europe	522	Southern Europe	908	Greece	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	Immigrants	Foreigners	Albania	908	Europe	522	Southern Europe	914	Developed regions	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	Immigrants	Foreigners	Angola	904	Africa	910	Sub-Saharan Africa	911	Developing regions	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	Immigrants	Foreigners	Argentina	909	Americas	905	South America	931	Developed regions	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	Immigrants	Foreigners	Armenia	908	Europe	522	Eastern Europe	925	Developing regions	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

```
In [4]: df_can.shape
```

```
(195, 43)
```

```
In [5]: # clean up the dataset to remove unnecessary columns (eg. beta)
        df_can.drop(['AREA', 'REG', 'DEV', 'Type', 'Coverage'], axis=1, inplace=True)

        # let's rename the columns so that they make sense
        df_can.rename(columns={'OdeName':'Country', 'AreaName':'Continent', 'RegName':'Region'}, inplace=True)

        # for sake of simplicity, let's use df_can.columns to replace df_can.columns
        all_columns = list(df_can.columns)
        # let's convert the version of the dataset to a list
        new_columns = list(map(str, all_columns))
        df_can.columns = new_columns

        # let's perform a quick summary of the dataset
        print('data dimensions: {}'.format(df_can.shape))
```

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

```
In [6]: df_can.head()
```

	Country	Continent	Region	Dorfman	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
0	Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496	741	828	1076	1405	1799	2021	2335	2571	2854	3064	3214	3443	3548	3743	4083	4259	4274	4293	4132	4167	4265	4244	4158	4035	3837	3622
1	Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69	132	251	329	304	327	548	548	843	1057	1031	1004	1148	1223	1267	1171	1224	1626	1854	2426	2626	2629	2099	2943	3358	4879	3294
3	American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5 rows × 38 columns

```
In [7]: # set the country name as index - useful for quickly looking for countries
        df_can.set_index('Country', inplace=True)

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

Country	Continent	Region	Dorfman	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
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496	741	828	1076	1405	1799	2021	2335	2571	2854	3064	3214	3443	3548	3743	4083	4259	4274	4293	4132	4167	4265	4244	4158	4035	3837	3622
Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69	132	251	329	304	327	548	548	843	1057	1031	1004	1148	1223	1267	1171	1224	1626	1854	2426	2626	2629	2099	2943	3358	4879	3294
American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5 rows × 38 columns

```
In [8]: # notice now the country names now serve as indices
        df_can.loc['Japan', :]
```

Country	Continent	Region	Dorfman	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
Japan	Asia	Eastern Asia	Developed regions	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

1 rows × 38 columns

```
In [9]: # notice now the country names now serve as indices
        df_can.loc['Japan', :].values
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       
```