Data Visualization

Agenda

1. Line Plot
2. Area Plot
3. Histogram
4. Bar Chart

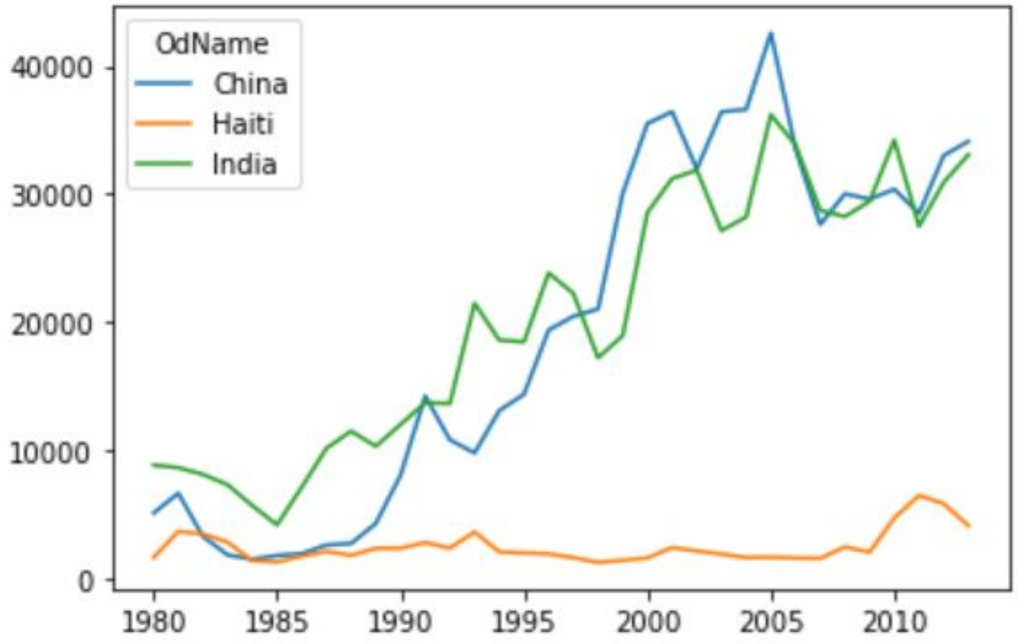
Line Plot

Line Plot - Complete Example

In [1]:
```
import pandas as pd

df = pd.read_csv('canada-mig-dataset.csv')
df1 = df.loc[ df['OdName'].isin(["China", "India", "Haiti"]) ]
df2 = df1.set_index('OdName')
df3 = df2.iloc[:, 8:42]
df4 = df3.transpose()
df4.plot(kind='line')
```

Out[1]: <AxesSubplot:>



Area Plot

Area Plot (Area Chart or Area Graph)

- is an **extension of** (based on) the **line plot**
- is commonly used when trying to **compare** two or more **quantities**
- depicts **accumulated totals** using numbers/percentages **over time**

Area Plot Example - Cell 1

In [1]:
```
import pandas as pd

df = pd.read_csv('canada-mig-dataset.csv')

df.head()
```

Out[1]:

	Type	Coverage	OdName	AREA	AreaName	REG	RegName	DEV	DevName	1980	...	2004	2005	2006	2007	2008	2009	2010	2011	2012
0	Immigrants	Foreigners	Afghanistan	935	Asia	5501	Southern Asia	902	Developing regions	16	...	2978	3436	3009	2652	2111	1746	1758	2203	2635
1	Immigrants	Foreigners	Albania	908	Europe	925	Southern Europe	901	Developed regions	1	...	1450	1223	856	702	560	716	561	539	620
2	Immigrants	Foreigners	Algeria	903	Africa	912	Northern Africa	902	Developing regions	80	...	3616	3626	4807	3623	4005	5393	4752	4325	3774
3	Immigrants	Foreigners	American Samoa	909	Oceania	957	Polynesia	902	Developing regions	0	...	0	0	1	0	0	0	0	0	0
4	Immigrants	Foreigners	Andorra	908	Europe	925	Southern Europe	901	Developed regions	0	...	0	0	1	1	0	0	0	0	1

5 rows × 43 columns

Area Plot Example - Cell 2

In [2]:
```
df1 = df.set_index('OdName')
df1.head()
```

Out[2]:

OdName	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2004	2005	2006	2007	2008	2009	2010	2011	2
Afghanistan	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	16	39	...	2978	3436	3009	2652	2111	1746	1758	2203	2
Albania	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	1	0	...	1450	1223	856	702	560	716	561	539	
Algeria	Immigrants	Foreigners	903	Africa	912	Northern Africa	902	Developing regions	80	67	...	3616	3626	4807	3623	4005	5393	4752	4325	3
American Samoa	Immigrants	Foreigners	909	Oceania	957	Polynesia	902	Developing regions	0	1	...	0	0	1	0	0	0	0	0	
Andorra	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	0	0	...	0	0	1	1	0	0	0	0	

5 rows × 42 columns

Area Plot Example - Cell 3

In [3]:
```
df1['Total'] = df1.iloc[:, 8:42].sum(axis=1)
df1.head()
```

Out[3]:

e	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2005	2006	2007	2008	2009	2010	2011	2012	2013	Total
n	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	16	39	...	3436	3009	2652	2111	1746	1758	2203	2635	2004	58639
a	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	1	0	...	1223	856	702	560	716	561	539	620	603	15699
a	Immigrants	Foreigners	903	Africa	912	Northern Africa	902	Developing regions	80	67	...	3626	4807	3623	4005	5393	4752	4325	3774	4331	69439
n a	Immigrants	Foreigners	909	Oceania	957	Polynesia	902	Developing regions	0	1	...	0	1	0	0	0	0	0	0	0	6
a	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	0	0	...	0	1	1	0	0	0	0	1	1	15

3 columns

Area Plot Example - Cell 4

In [4]:
```
df1.sort_values(by=['Total'], ascending = False)
```

Out[4]:

OdName	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2005	2006	2007	2008	2009	2010	2011
India	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	8880	8670	...	36210	33848	28742	28261	29456	34235	27509
China	Immigrants	Foreigners	935	Asia	906	Eastern Asia	902	Developing regions	5123	6682	...	42584	33518	27642	30037	29622	30391	28502
United Kingdom of Great Britain and Northern Ireland	Immigrants	Foreigners	908	Europe	924	Northern Europe	901	Developed regions	22045	24796	...	7258	7140	8216	8979	8876	8724	6204
Unknown	Immigrants	Foreigners	999	World	999	World	999	World	44000	18078	...	4785	4583	4348	4197	3402	3731	2554
Philippines	Immigrants	Foreigners	935	Asia	920	South-Eastern Asia	902	Developing regions	6051	5921	...	18139	18400	19837	24887	28573	38617	36765
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Area Plot Example - Cell 5

In [5]:
```
df1.sort_values(by=['Total'], ascending = False,  inplace = True)
df1.head()
```

Out[5]:

OdName	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2005	2006	2007	2008	2009	2010	2011
India	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	8880	8670	...	36210	33848	28742	28261	29456	34235	27509
China	Immigrants	Foreigners	935	Asia	906	Eastern Asia	902	Developing regions	5123	6682	...	42584	33518	27642	30037	29622	30391	28502
United Kingdom of Great Britain and Northern Ireland	Immigrants	Foreigners	908	Europe	924	Northern Europe	901	Developed regions	22045	24796	...	7258	7140	8216	8979	8876	8724	6204
Unknown	Immigrants	Foreigners	999	World	999	World	999	World	44000	18078	...	4785	4583	4348	4197	3402	3731	2554
Philippines	Immigrants	Foreigners	935	Asia	920	South-Eastern Asia	902	Developing regions	6051	5921	...	18139	18400	19837	24887	28573	38617	36765

5 rows × 43 columns

Area Plot Example - Cell 6

In [6]:
```
df2 = df1.head()
df2[list(map(str, range(1980,2014)))]
```

Out[6]:

OdName	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	...	2004	2005	2006	2007	2008	2009	2010	2011	2012	
India	8880	8670	8147	7338	5704	4211	7150	10189	11522	10343	...	28235	36210	33848	28742	28261	29456	34235	27509	30933	3
China	5123	6682	3308	1863	1527	1816	1960	2643	2758	4323	...	36619	42584	33518	27642	30037	29622	30391	28502	33024	3
United Kingdom of Great Britain and Northern Ireland	22045	24796	20620	10015	10170	9564	9470	21337	27359	23795	...	7533	7258	7140	8216	8979	8876	8724	6204	6195	
Unknown	44000	18078	16904	13635	14855	14368	13303	17304	22279	27118	...	3739	4785	4583	4348	4197	3402	3731	2554	1681	
Philippines	6051	5921	5249	4562	3801	3150	4166	7360	8639	11865	...	14004	18139	18400	19837	24887	28573	38617	36765	34315	2

5 rows × 34 columns

Area Plot Example - Cell 7

In [7]:
```
df3 = df2[list(map(str, range(1980,2014)))].transpose()
df3.head()
```

Out[7]:

OdName	India	China	United Kingdom of Great Britain and Northern Ireland	Unknown	Philippines
1980	8880	5123	22045	44000	6051
1981	8670	6682	24796	18078	5921
1982	8147	3308	20620	16904	5249
1983	7338	1863	10015	13635	4562
1984	5704	1527	10170	14855	3801

Area Plot Example - Cell 8

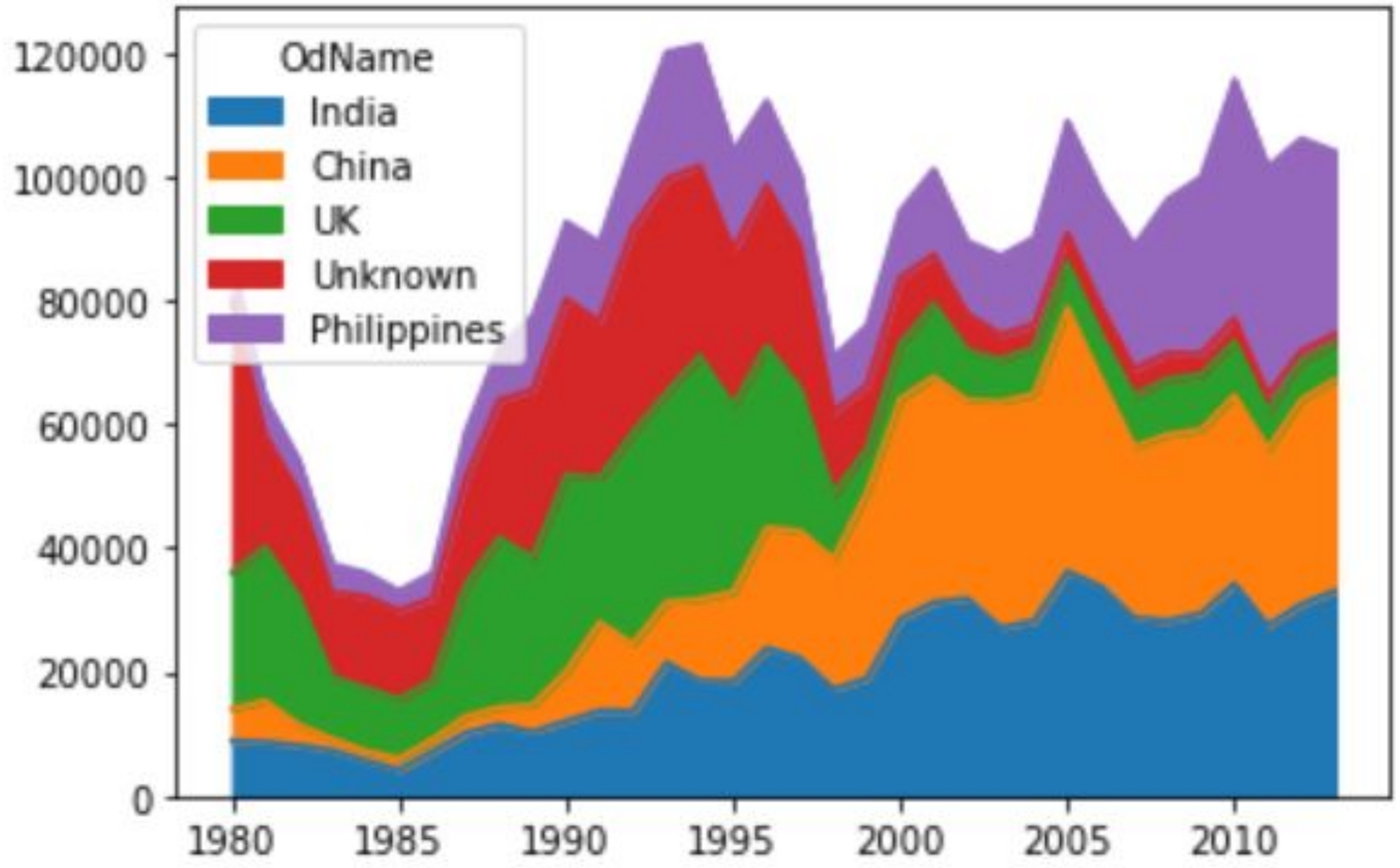
In [8]:
```
df4 = df3.rename(columns = {"United Kingdom of Great Britain and Northern Ireland":"UK"})
df4.head()
```

Out[8]:

OdName	India	China	UK	Unknown	Philippines
1980	8880	5123	22045	44000	6051
1981	8670	6682	24796	18078	5921
1982	8147	3308	20620	16904	5249
1983	7338	1863	10015	13635	4562
1984	5704	1527	10170	14855	3801

Area Plot Example - Cell 9

```
In [9]: import matplotlib.pyplot as plt
        df4.plot(kind='area')
        plt.show()
```

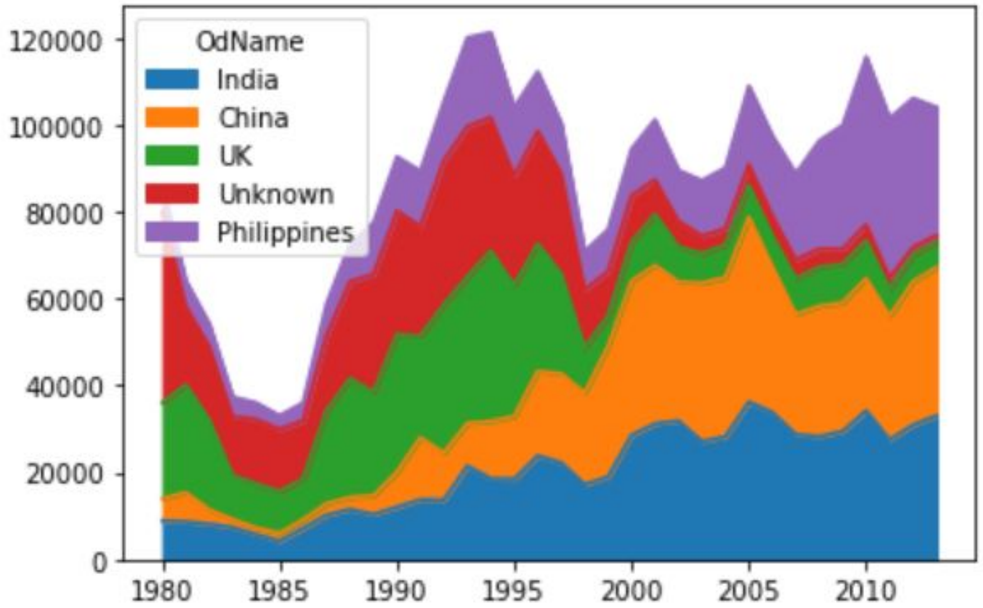


Area Plot - Complete Example

```
import pandas as pd

df0 = pd.read_csv('canada-mig-dataset.csv')
df1 = df0.set_index('OdName')
df1['Total'] = df1.iloc[:, 8:42].sum(axis=1)
df1.sort_values(by=['Total'], ascending = False,  inplace = True)
df2 = df1.head()
df3 = df2[list(map(str, range(1980,2014)))].transpose()
df4 = df3.rename(columns = {"United Kingdom of Great Britain and Northern Irelanc
df4.plot(kind='area')
```

```
<AxesSubplot:>
```



Histogram

Histogram

- represents the **frequency distribution** of a **numeric dataset**

- The way it works is:
  - partitions the spread of the numeric data into bins
  - assigns each datapoint in the dataset to a bin
  - counts the number of datapoints that have been assigned to each bin

- **Vertical axis**
  - is actually the frequency or the number of datapoints in each bin

Histogram Example - Cell 1

In [1]:
```
import pandas as pd

df = pd.read_csv('canada-mig-dataset.csv')

df.head()
```

Out[1]:

	Type	Coverage	OdName	AREA	AreaName	REG	RegName	DEV	DevName	1980	...	2004	2005	2006	2007	2008	2009	2010	2011	2012
0	Immigrants	Foreigners	Afghanistan	935	Asia	5501	Southern Asia	902	Developing regions	16	...	2978	3436	3009	2652	2111	1746	1758	2203	2635
1	Immigrants	Foreigners	Albania	908	Europe	925	Southern Europe	901	Developed regions	1	...	1450	1223	856	702	560	716	561	539	620
2	Immigrants	Foreigners	Algeria	903	Africa	912	Northern Africa	902	Developing regions	80	...	3616	3626	4807	3623	4005	5393	4752	4325	3774
3	Immigrants	Foreigners	American Samoa	909	Oceania	957	Polynesia	902	Developing regions	0	...	0	0	1	0	0	0	0	0	0
4	Immigrants	Foreigners	Andorra	908	Europe	925	Southern Europe	901	Developed regions	0	...	0	0	1	1	0	0	0	0	1

5 rows × 43 columns

Histogram Example - Cell 2

In [2]:
```
df1 = df.set_index('OdName')
df1.head()
```

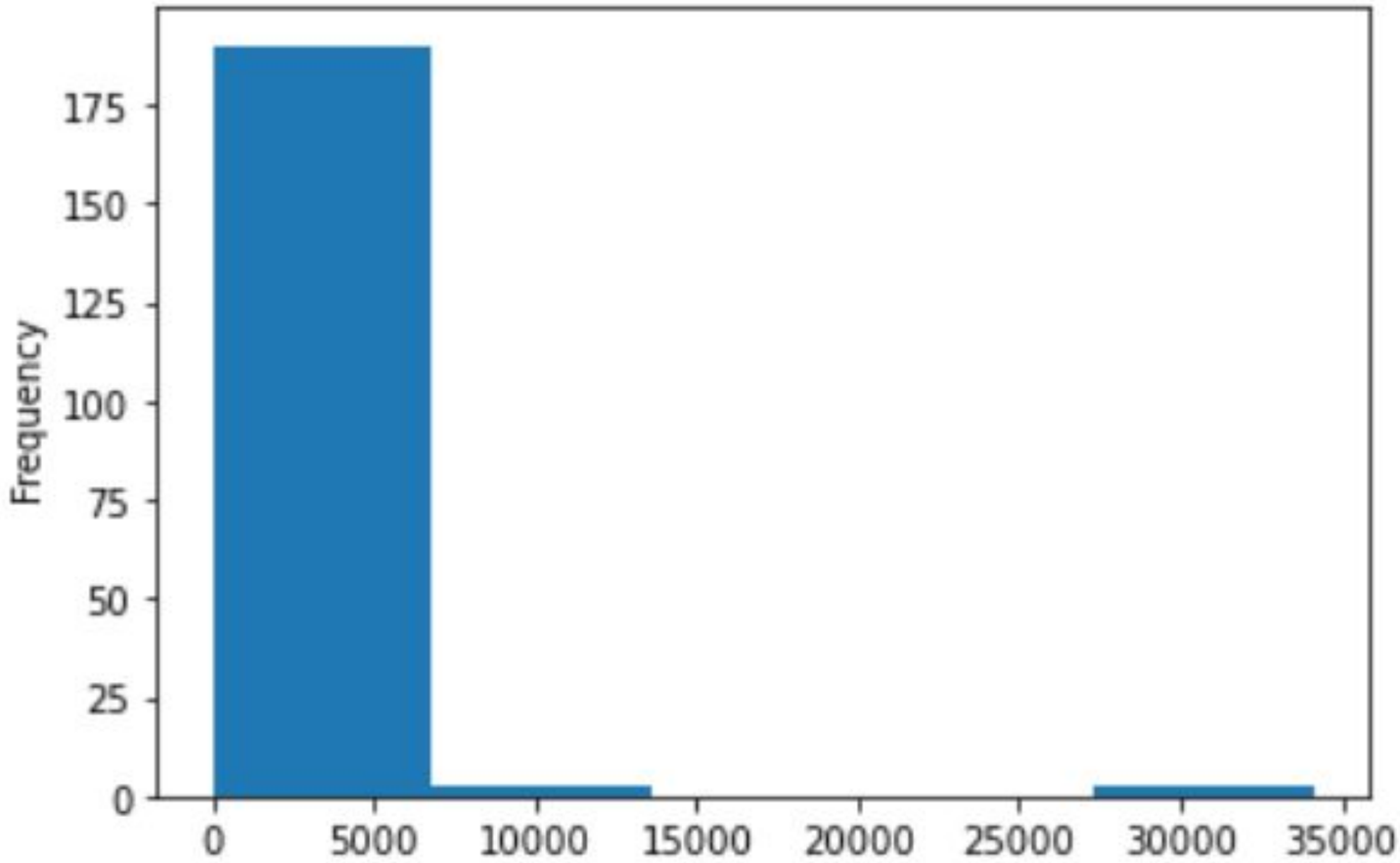
Out[2]:

OdName	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2004	2005	2006	2007	2008	2009	2010	2011	2
Afghanistan	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	16	39	...	2978	3436	3009	2652	2111	1746	1758	2203	2
Albania	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	1	0	...	1450	1223	856	702	560	716	561	539	
Algeria	Immigrants	Foreigners	903	Africa	912	Northern Africa	902	Developing regions	80	67	...	3616	3626	4807	3623	4005	5393	4752	4325	3
American Samoa	Immigrants	Foreigners	909	Oceania	957	Polynesia	902	Developing regions	0	1	...	0	0	1	0	0	0	0	0	
Andorra	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	0	0	...	0	0	1	1	0	0	0	0	

5 rows × 42 columns

Histogram Example - Cell 3

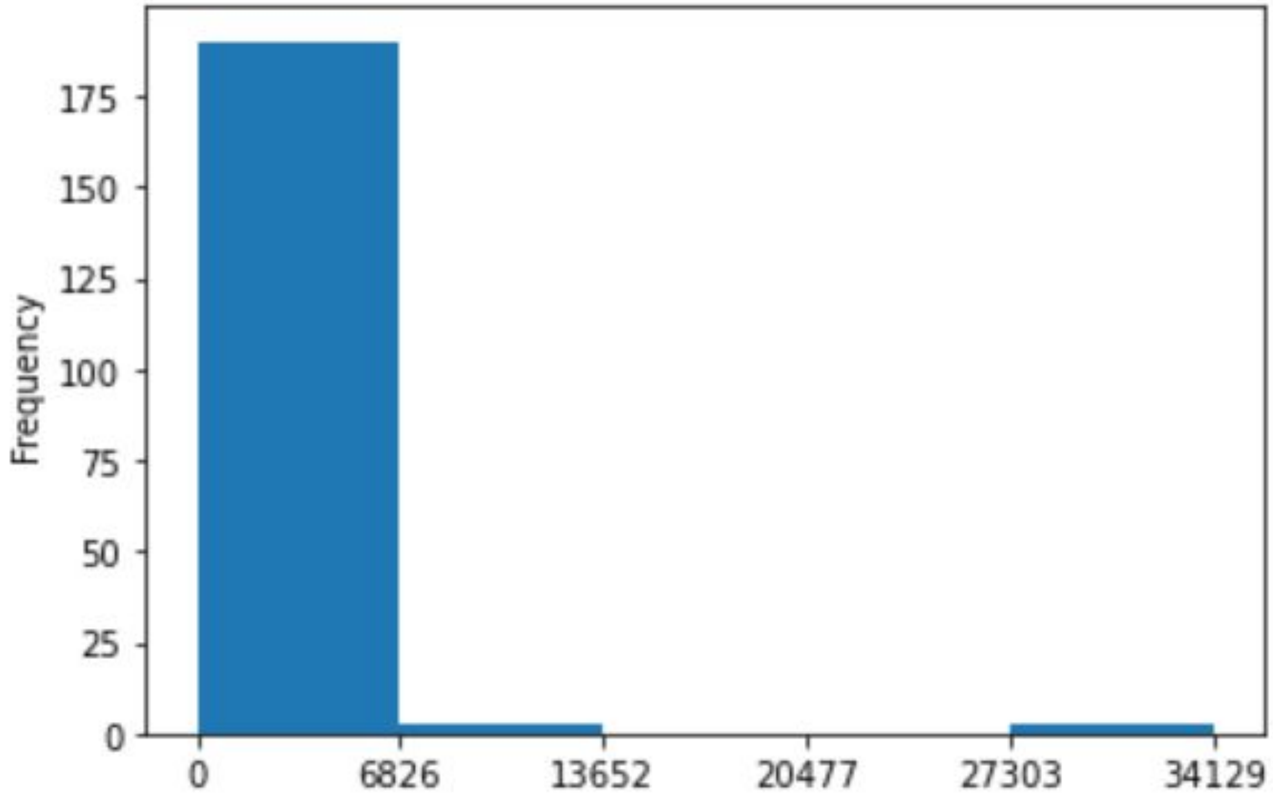
```
In [3]: df1['2013'].plot(kind='hist', bins = 5)
Out[3]: <AxesSubplot:ylabel='Frequency'>
```



Histogram Example - Cell 4

In [4]:
```
import numpy as np
count, bin_edges = np.histogram(df1['2013'], bins=5)
df1['2013'].plot(kind='hist', bins = 5, xticks = bin_edges)
```

Out[4]: <AxesSubplot:ylabel='Frequency'>

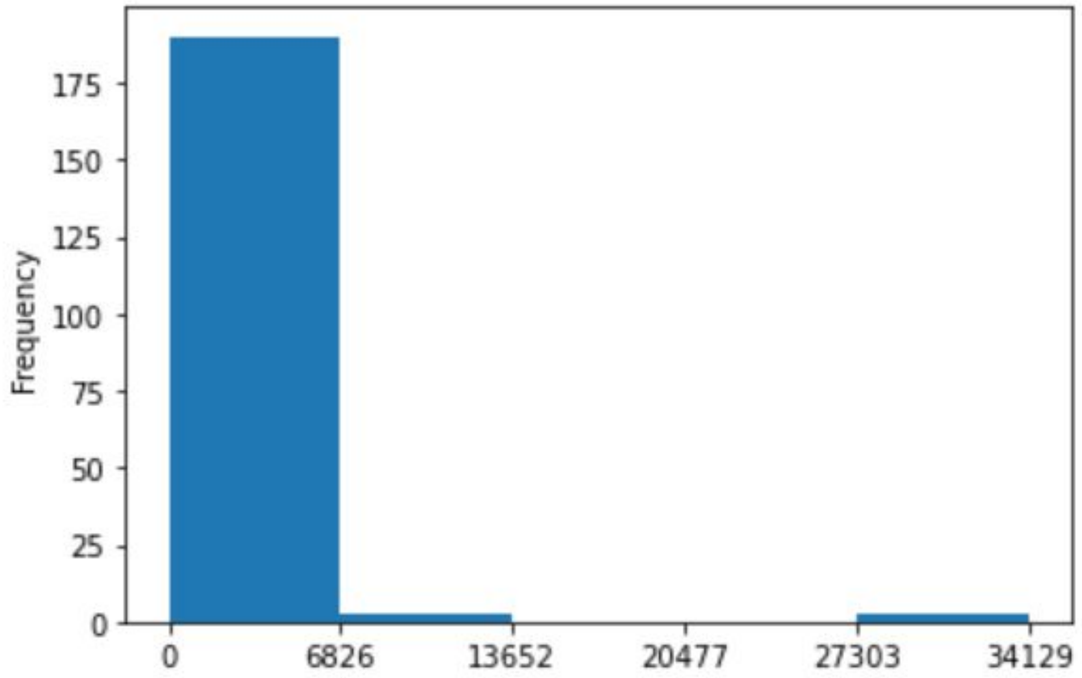


Histogram - Complete Example

In [1]:
```
import pandas as pd
import numpy as np

df0 = pd.read_csv('canada-mig-dataset.csv')
df1 = df0.set_index('OdName')
count, bin_edges = np.histogram(df1['2013'], bins=5)
df1['2013'].plot(kind='hist', bins = 5, xticks = bin_edges)
```

Out[1]: <AxesSubplot:ylabel='Frequency'>



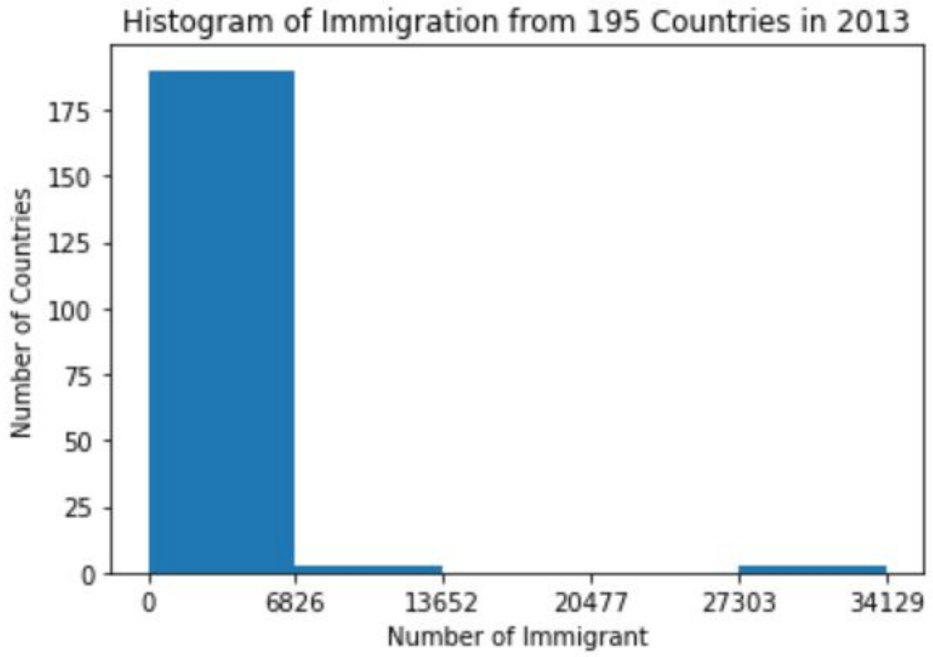
Histogram - Complete Example

- Add Title
- Add X Label
- Add Y Label

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df0 = pd.read_csv('canada-mig-dataset.csv')
df1 = df0.set_index('OdName')
count, bin_edges = np.histogram(df1['2013'], bins=5)
df1['2013'].plot(kind='hist', bins = 5, xticks = bin_edges)

plt.title('Histogram of Immigration from 195 Countries in 2013')
plt.xlabel('Number of Immigrant')
plt.ylabel('Number of Countries')
plt.show()
```



Bar Chart

Bar Chart (Bar Graph)

- `is a very popular visualization tool`
  - the length of each bar is proportional to the value of the item that it represents

- `commonly used:`
  - to compare the values of a variable at a given point in time

- `Example:`
  - visualizing in discrete fashion how immigration from Iceland to Canada evolved
  - bar height represents total immigration from Iceland to Canada in a specific year

Bar Chart Example - Cell 1

In [1]:
```
import pandas as pd

df = pd.read_csv('canada-mig-dataset.csv')

df.head()
```

Out[1]:

	Type	Coverage	OdName	AREA	AreaName	REG	RegName	DEV	DevName	1980	...	2004	2005	2006	2007	2008	2009	2010	2011	2012
0	Immigrants	Foreigners	Afghanistan	935	Asia	5501	Southern Asia	902	Developing regions	16	...	2978	3436	3009	2652	2111	1746	1758	2203	2635
1	Immigrants	Foreigners	Albania	908	Europe	925	Southern Europe	901	Developed regions	1	...	1450	1223	856	702	560	716	561	539	620
2	Immigrants	Foreigners	Algeria	903	Africa	912	Northern Africa	902	Developing regions	80	...	3616	3626	4807	3623	4005	5393	4752	4325	3774
3	Immigrants	Foreigners	American Samoa	909	Oceania	957	Polynesia	902	Developing regions	0	...	0	0	1	0	0	0	0	0	0
4	Immigrants	Foreigners	Andorra	908	Europe	925	Southern Europe	901	Developed regions	0	...	0	0	1	1	0	0	0	0	1

5 rows × 43 columns

Bar Chart Example - Cell 2

In [2]:
```
df1 = df.set_index('OdName')
df1.head()
```

Out[2]:

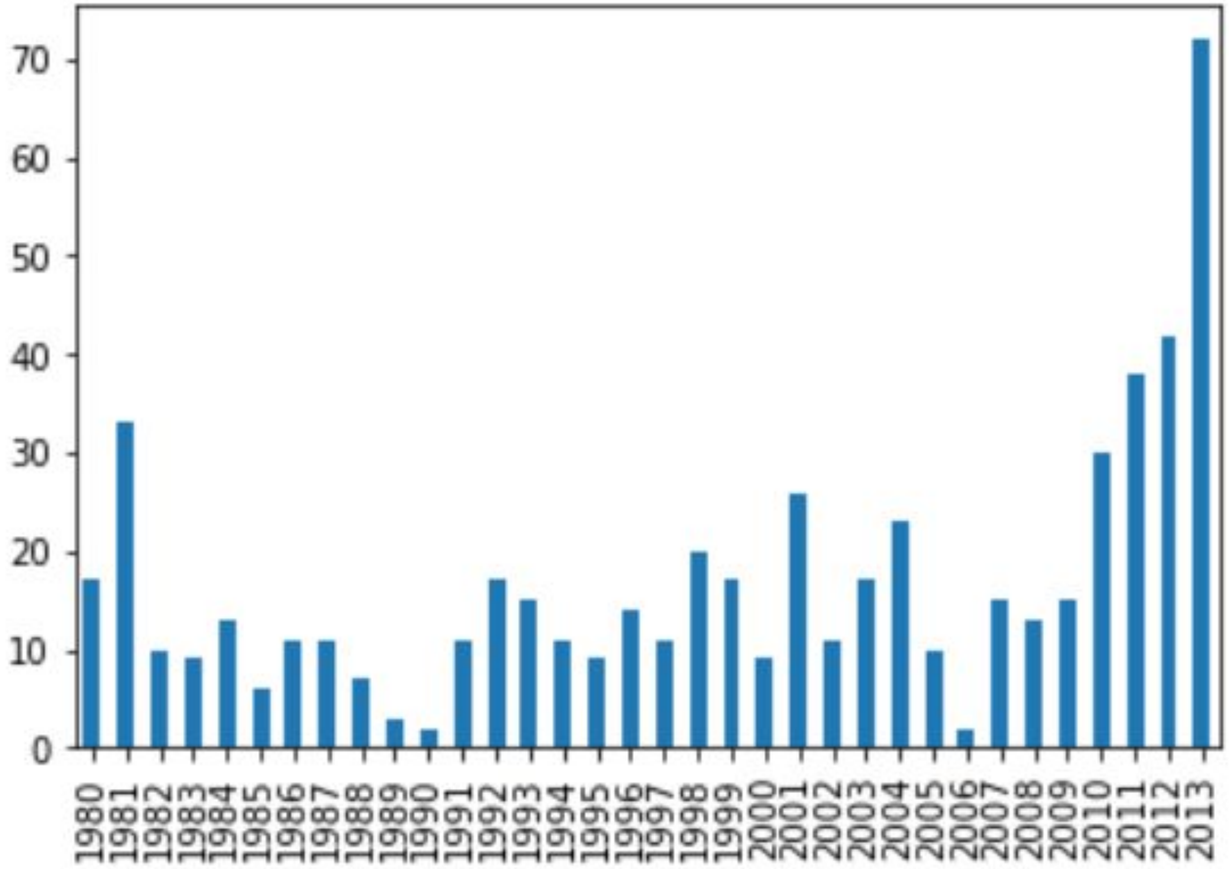
OdName	Type	Coverage	AREA	AreaName	REG	RegName	DEV	DevName	1980	1981	...	2004	2005	2006	2007	2008	2009	2010	2011	2
Afghanistan	Immigrants	Foreigners	935	Asia	5501	Southern Asia	902	Developing regions	16	39	...	2978	3436	3009	2652	2111	1746	1758	2203	2
Albania	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	1	0	...	1450	1223	856	702	560	716	561	539	
Algeria	Immigrants	Foreigners	903	Africa	912	Northern Africa	902	Developing regions	80	67	...	3616	3626	4807	3623	4005	5393	4752	4325	3
American Samoa	Immigrants	Foreigners	909	Oceania	957	Polynesia	902	Developing regions	0	1	...	0	0	1	0	0	0	0	0	
Andorra	Immigrants	Foreigners	908	Europe	925	Southern Europe	901	Developed regions	0	0	...	0	0	1	1	0	0	0	0	

5 rows × 42 columns

Bar Chart Example - Cell 3

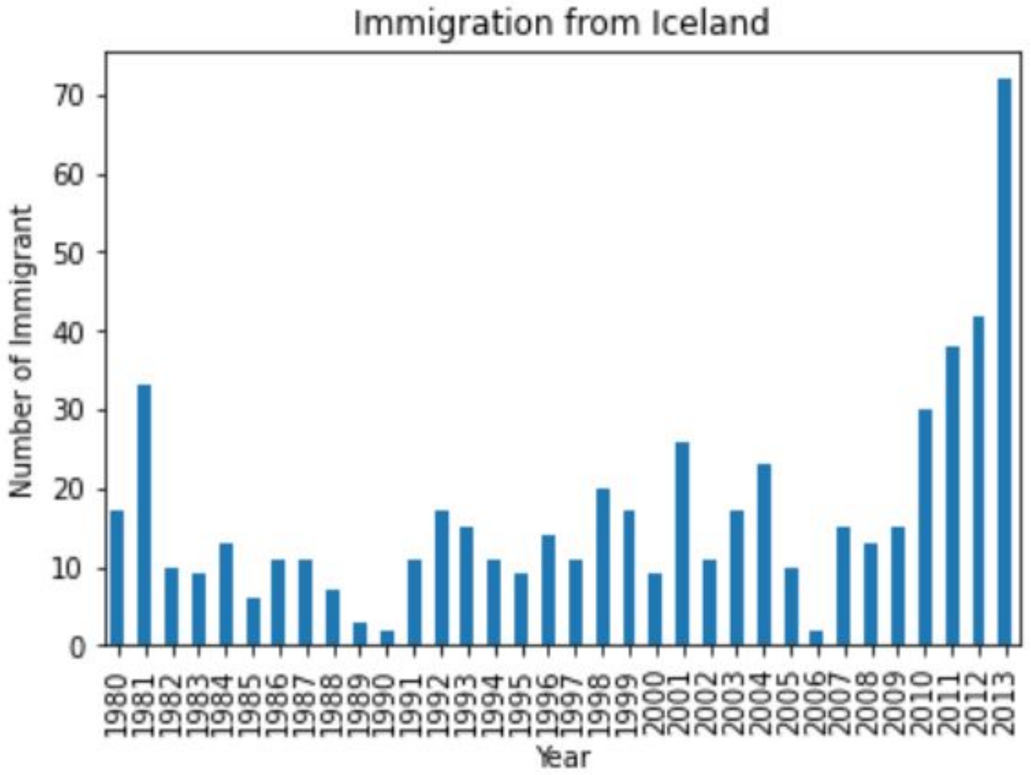
```
In [3]: df2 = df1.loc[ "Iceland", list(map(str, range(1980,2014))) ]
        df2.plot(kind='bar')
```

```
Out[3]: <AxesSubplot:>
```



Bar Chart Example - Cell 4

In [4]:
```
import numpy as np
import matplotlib.pyplot as plt
df2.plot(kind='bar')
plt.title("Immigration from Iceland")
plt.ylabel("Number of Immigrant")
plt.xlabel("Year")
plt.show()
```

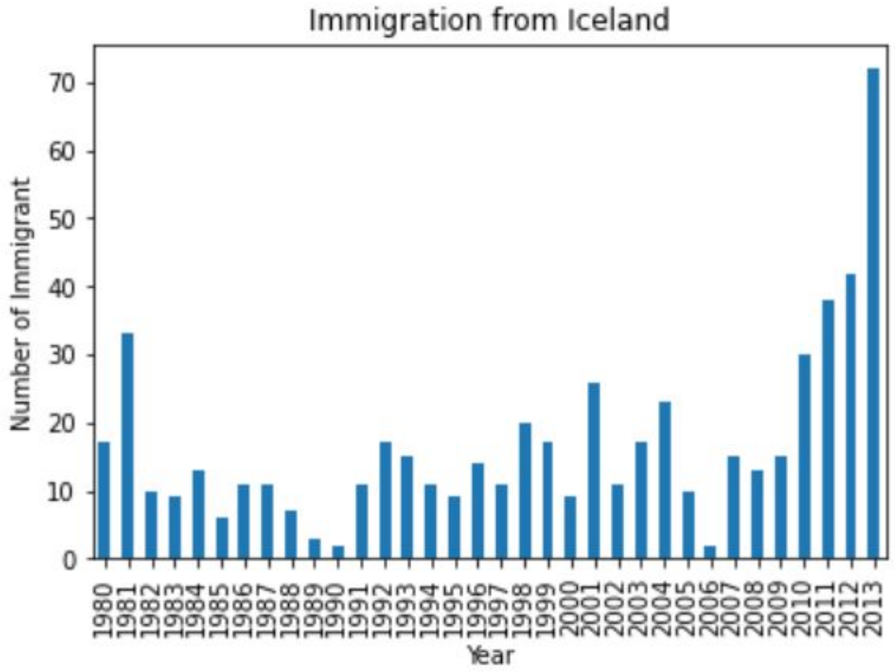


Bar Chart - Complete Example

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt

df0 = pd.read_csv('canada-mig-dataset.csv')
df1 = df0.set_index('OdName')
df2 = df1.loc[ "Iceland", list(map(str, range(1980,2014))) ]
df2.plot(kind='bar')

plt.title("Immigration from Iceland")
plt.ylabel("Number of Immigrant")
plt.xlabel("Year")
plt.show()
```



Questions

Links

https://github.com/fcai-b/dv

References

1. https://www.coursera.org/learn/python-for-data-visualization