

Lab 5 : Pandas Concatenate, Merge and Join

DINESH KUMAR K

225229108

```
In [72]: import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

```
In [73]: north_america = pd.read_csv('north_america_2000_2010.csv',index_col=0)
south_america = pd.read_csv('south_america_2000_2010.csv',index_col=0)
```

```
In [74]: north_america
```

```
Out[74]:
```

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Country											
Canada	1779.0	1771.0	1754.0	1740.0	1760.0	1747	1745.0	1741.0	1735	1701.0	1703.0
Mexico	2311.2	2285.2	2271.2	2276.5	2270.6	2281	2280.6	2261.4	2258	2250.2	2242.4
USA	1836.0	1814.0	1810.0	1800.0	1802.0	1799	1800.0	1798.0	1792	1767.0	1778.0

```
In [75]: south_america
```

```
Out[75]:
```

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Country											
Chile	2263	2242	2250	2235	2232	2157	2165	2128	2095	2074	2069.6

```
In [76]: !type north_america_2000_2010.csv
```

```
Country,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010
Canada,1779,1771,1754,1740,1760,1747,1745,1741,1735,1701,1703
Mexico,2311.2,2285.2,2271.2,2276.5,2270.6,2281,2280.6,2261.4,2258,2250.2,2242.4
USA,1836,1814,1810,1800,1802,1799,1800,1798,1792,1767,1778
```

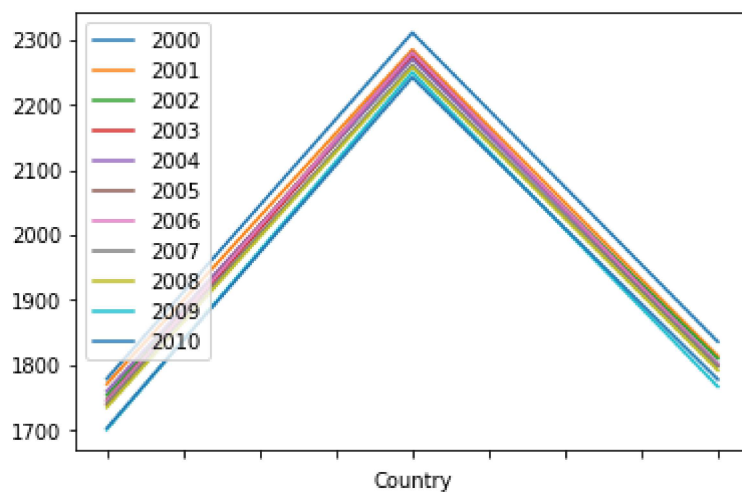
```
In [77]: !type south_america_2000_2010.csv
```

```
Country,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010
Chile,2263,2242,2250,2235,2232,2157,2165,2128,2095,2074,2069.6
```

Create line graphs for our yearly labor trends in north_america

In [78]: `north_america.plot()`

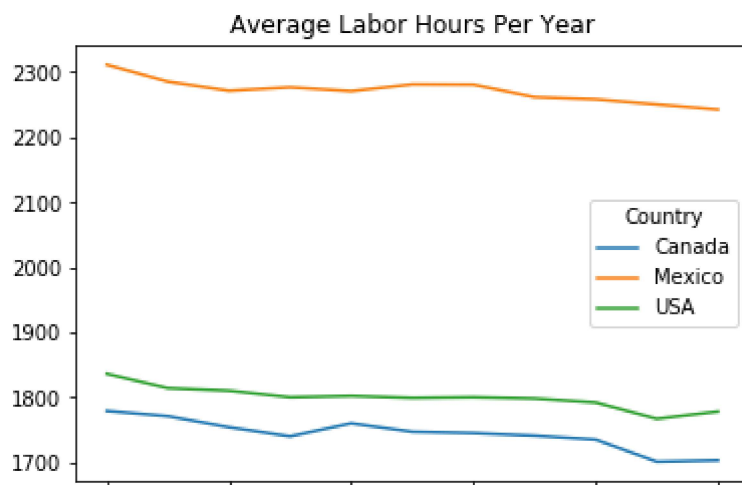
Out[78]: `<matplotlib.axes._subplots.AxesSubplot at 0x22ce67e3550>`



Plot transposed line graph of north_america dataframe, with title "Average Labor Hours Per Year"

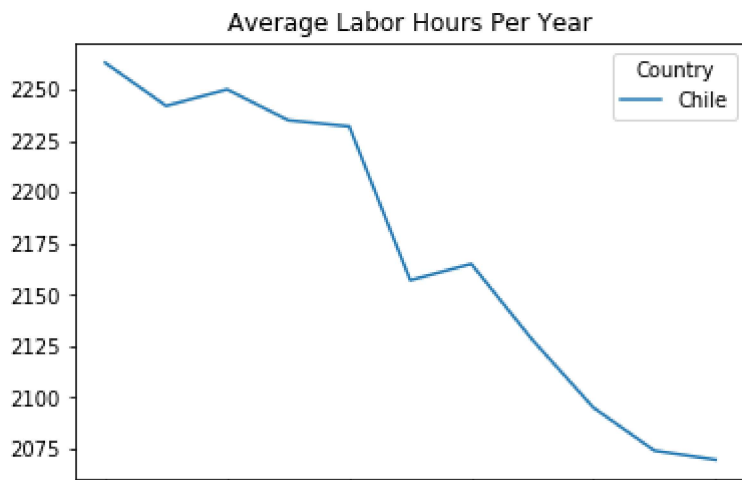
In [79]: `north_america.transpose().plot(title='Average Labor Hours Per Year')`

Out[79]: `<matplotlib.axes._subplots.AxesSubplot at 0x22ce7bbfa58>`



Plot transposed line graph of north_america dataframe, with title "Average Labor Hours Per Year"

```
In [80]: south_america.transpose().plot(title='Average Labor Hours Per Year')
plt.show()
```



Concatenate America Data

Concatenate north_america and south_america dataframes and store result in a dataframe, americas

```
In [81]: americas=pd.concat([north_america,south_america])
americas
```

```
Out[81]:
```

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Country											
Canada	1779.0	1771.0	1754.0	1740.0	1760.0	1747	1745.0	1741.0	1735	1701.0	1703.0
Mexico	2311.2	2285.2	2271.2	2276.5	2270.6	2281	2280.6	2261.4	2258	2250.2	2242.4
USA	1836.0	1814.0	1810.0	1800.0	1802.0	1799	1800.0	1798.0	1792	1767.0	1778.0
Chile	2263.0	2242.0	2250.0	2235.0	2232.0	2157	2165.0	2128.0	2095	2074.0	2069.6

Load the additional files

```
In [82]: americas_dfs = [americas]
for year in range(2011, 2016):
    filename = "./americas_{}.csv".format(year)
    df = pd.read_csv(filename, index_col=0)
    americas_dfs.append(df)
```

In [83]: `americas_dfs[1]`

Out[83]:

	2011
Country	
Canada	1700.0
Chile	2047.4
Mexico	2250.2
USA	1786.0

Concatenate `americas` and `americas_dfs` dataframes and store result in `americas`

In [84]: `americas=pd.concat(americas_dfs,axis=1)`
`americas.index.names=['country']`

In [85]: `americas`

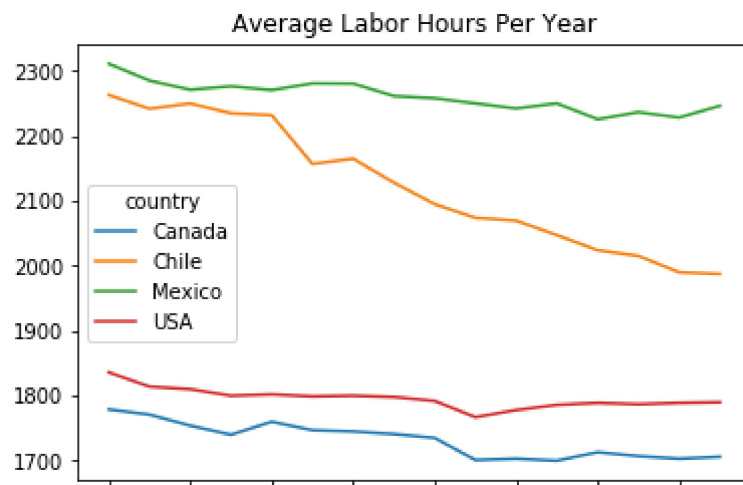
Out[85]:

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
country												
Canada	1779.0	1771.0	1754.0	1740.0	1760.0	1747	1745.0	1741.0	1735	1701.0	1703.0	1700.0
Chile	2263.0	2242.0	2250.0	2235.0	2232.0	2157	2165.0	2128.0	2095	2074.0	2069.6	2047.4
Mexico	2311.2	2285.2	2271.2	2276.5	2270.6	2281	2280.6	2261.4	2258	2250.2	2242.4	2250.2
USA	1836.0	1814.0	1810.0	1800.0	1802.0	1799	1800.0	1798.0	1792	1767.0	1778.0	1786.0

Now, plot transposed `americas` dataframe

```
In [86]: americas.transpose().plot(title='Average Labor Hours Per Year')
```

```
Out[86]: <matplotlib.axes._subplots.AxesSubplot at 0x22ce7c536d8>
```



Appending data from other Continents

The data collection team has provided CSV files for Asia, Europe, and the South Pacific for 2000 through 2015. Let's load these files in and have a preview

```
In [87]: asia=pd.read_csv('asia_2000_2015.csv',index_col=0)
asia
```

```
Out[87]:
```

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Country															
Israel	2017	1979	1993	1974	1942	1931	1919	1931	1929	1927	1918	1920	1910	1867	1867
Japan	1821	1809	1798	1799	1787	1775	1784	1785	1771	1714	1733	1728	1745	1734	1734
Korea	2512	2499	2464	2424	2392	2351	2346	2306	2246	2232	2187	2090	2163	2079	2079
Russia	1982	1980	1982	1993	1993	1989	1998	1999	1997	1974	1976	1979	1982	1980	1980

```
In [88]: europe=pd.read_csv('europe_2000_2015.csv',index_col=0)
europe.head()
```

Out[88]:

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Country											
Austria	1807.4	1794.6	1792.2	1783.8	1786.8	1764.0	1746.2	1736.0	1728.5	1673.0	1668.6
Belgium	1595.0	1588.0	1583.0	1578.0	1573.0	1565.0	1572.0	1577.0	1570.0	1548.0	1546.0
Switzerland	1673.6	1635.0	1614.0	1626.8	1656.5	1651.7	1643.2	1632.7	1623.1	1614.9	1612.4
Czech Republic	1896.0	1818.0	1816.0	1806.0	1817.0	1817.0	1799.0	1784.0	1790.0	1779.0	1800.0
Germany	1452.0	1441.9	1430.9	1424.8	1422.2	1411.3	1424.7	1424.4	1418.4	1372.7	1389.9

```
In [89]: south_pacific=pd.read_csv('south_pacific_2000_2015.csv',index_col=0)
south_pacific
```

Out[89]:

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Country												
Australia	1778.7	1736.7	1731.7	1735.8	1734.5	1729.2	1720.5	1712.5	1717.2	1690	1691.5	1691.5
New Zealand	1836.0	1825.0	1826.0	1823.0	1830.0	1815.0	1795.0	1774.0	1761.0	1740	1755.0	1740

Append asia, europe and south_pacific to americas dataframe and assign to new dataframe world

```
In [90]: world=americas.append([asia,europe,south_pacific])
```

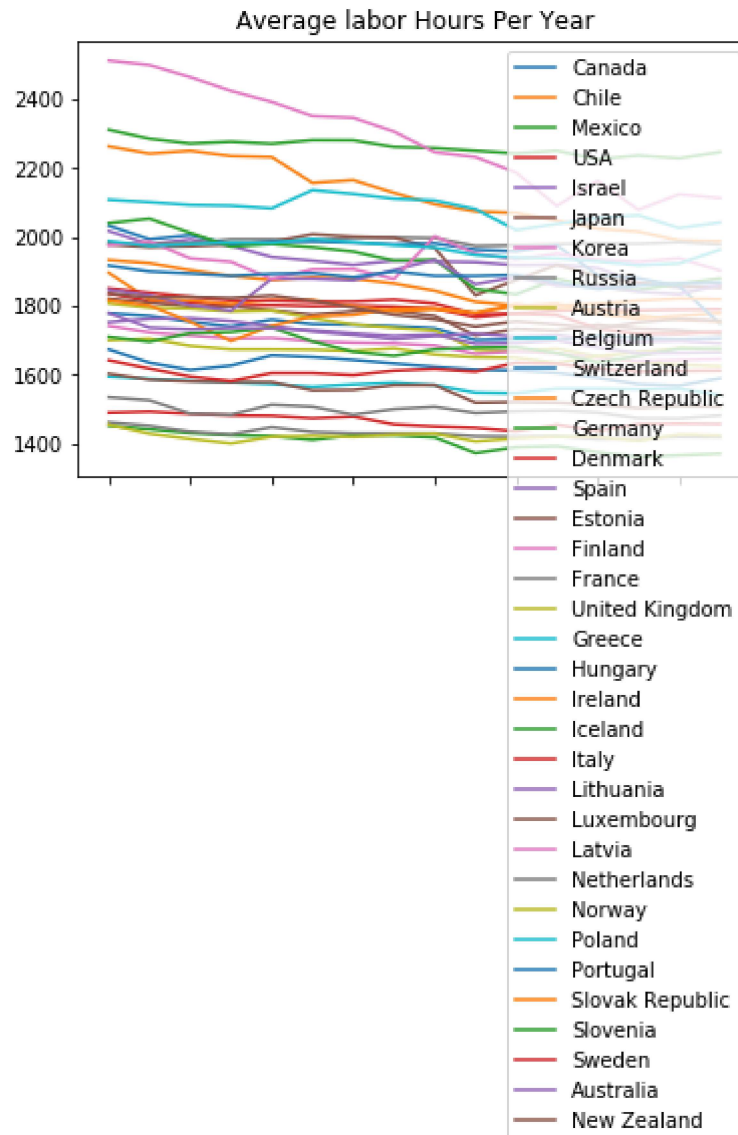
```
In [91]: world.index
```

```
Out[91]: Index(['Canada', 'Chile', 'Mexico', 'USA', 'Israel', 'Japan', 'Korea',
               'Russia', 'Austria', 'Belgium', 'Switzerland', 'Czech Republic',
               'Germany', 'Denmark', 'Spain', 'Estonia', 'Finland', 'France',
               'United Kingdom', 'Greece', 'Hungary', 'Ireland', 'Iceland', 'Italy',
               'Lithuania', 'Luxembourg', 'Latvia', 'Netherlands', 'Norway', 'Poland',
               'Portugal', 'Slovak Republic', 'Slovenia', 'Sweden', 'Australia',
               'New Zealand'],
              dtype='object')
```

Plot, transposed world dataframe

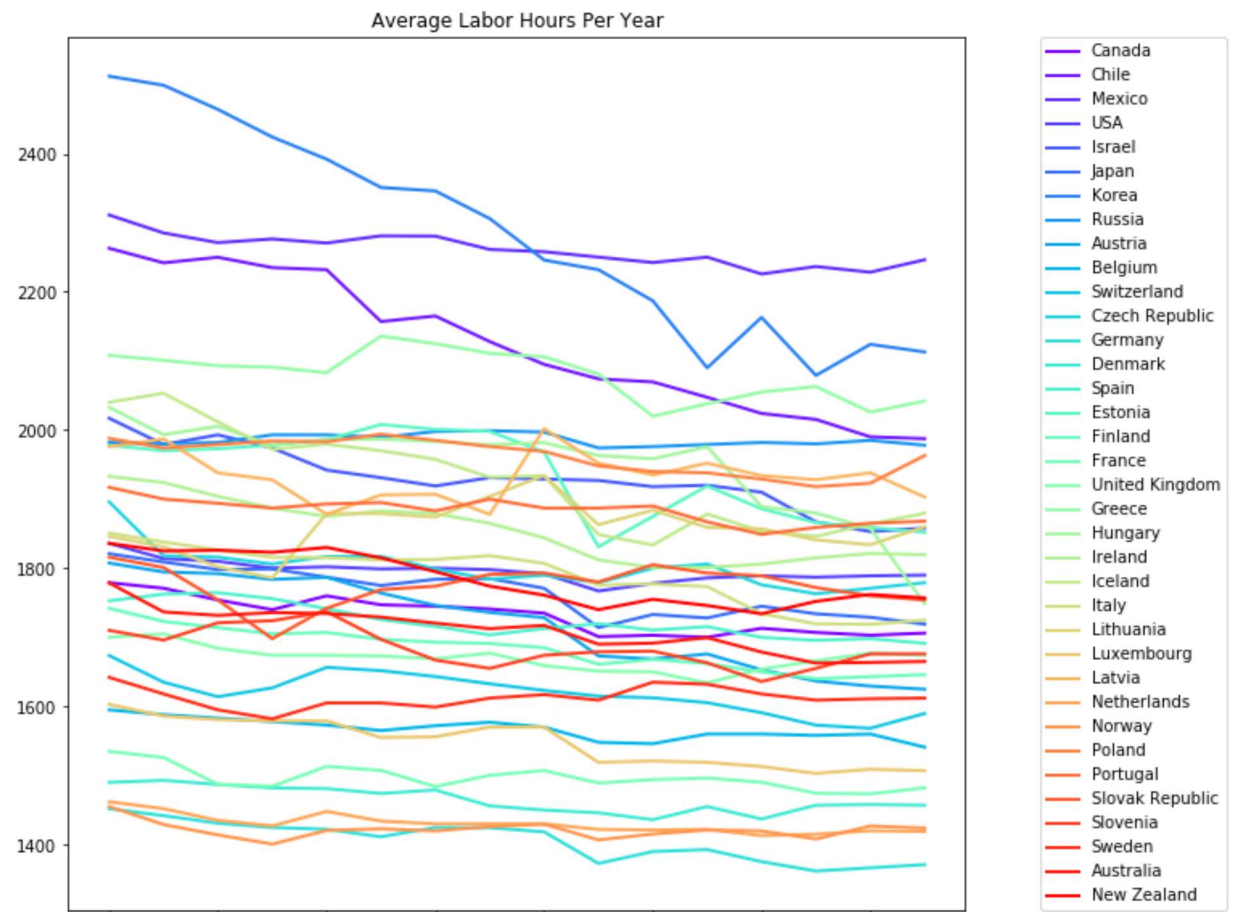
```
In [92]: world.transpose().plot(title='Average labor Hours Per Year')
```

```
Out[92]: <matplotlib.axes._subplots.AxesSubplot at 0x22ce801a860>
```



let us customize this plot, so that country names appear outside the chart

```
In [93]: world.transpose().plot(figsize=(10,10),colormap='rainbow',linewidth=2,title='Average Labor Hours Per Year')
plt.legend(loc='right',bbox_to_anchor=(1.3,0.5))
plt.show()
```



Merging Historical Labor Data


```
In [94]: historical = pd.read_csv('historical.csv', index_col=0)
historical.head()
```

Out[94]:

	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	...	1990	1991	
Country														
Australia	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1779.5	1774.90	177
Austria	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
Belgium	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1662.9	1625.79	166
Canada	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1789.5	1767.50	176
Switzerland	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	1673.10	166

5 rows × 50 columns



```
In [95]: print("World rows & columns: ", world.shape)
print("Historical rows & columns: ", historical.shape)
```

```
World rows & columns: (36, 16)
Historical rows & columns: (39, 50)
```

Merge historical dataframe with world dataframe and store in a new variable, world_historical

```
In [96]: world_historical = pd.merge(historical, world, left_index=True, right_index=True,
```

Print size of world_historical dataframe

```
In [97]: print(world_historical.shape)
```

```
(36, 66)
```

Print top-5 of world_historical dataframe

In [98]: `world_historical.head()`

Out[98]:

	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	...	2006
Canada	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1745.0
Chile	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2165.0
Mexico	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2280.6
USA	1960.0	1975.5	1978.0	1980.0	1970.5	1992.5	1990.0	1962.0	1936.5	1947.0	...	1800.0
Israel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1919.0

5 rows × 66 columns



Joining Historical Data

Use join method to join historical dataframe and world dataframe and store result in world_historical dataframe

In [99]: `world_historical = historical.join(world, how='right')`
`world_historical.head()`

Out[99]:

	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	...	2006
Canada	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1745.0
Chile	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2165.0
Mexico	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2280.6
USA	1960.0	1975.5	1978.0	1980.0	1970.5	1992.5	1990.0	1962.0	1936.5	1947.0	...	1800.0
Israel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1919.0

5 rows × 66 columns

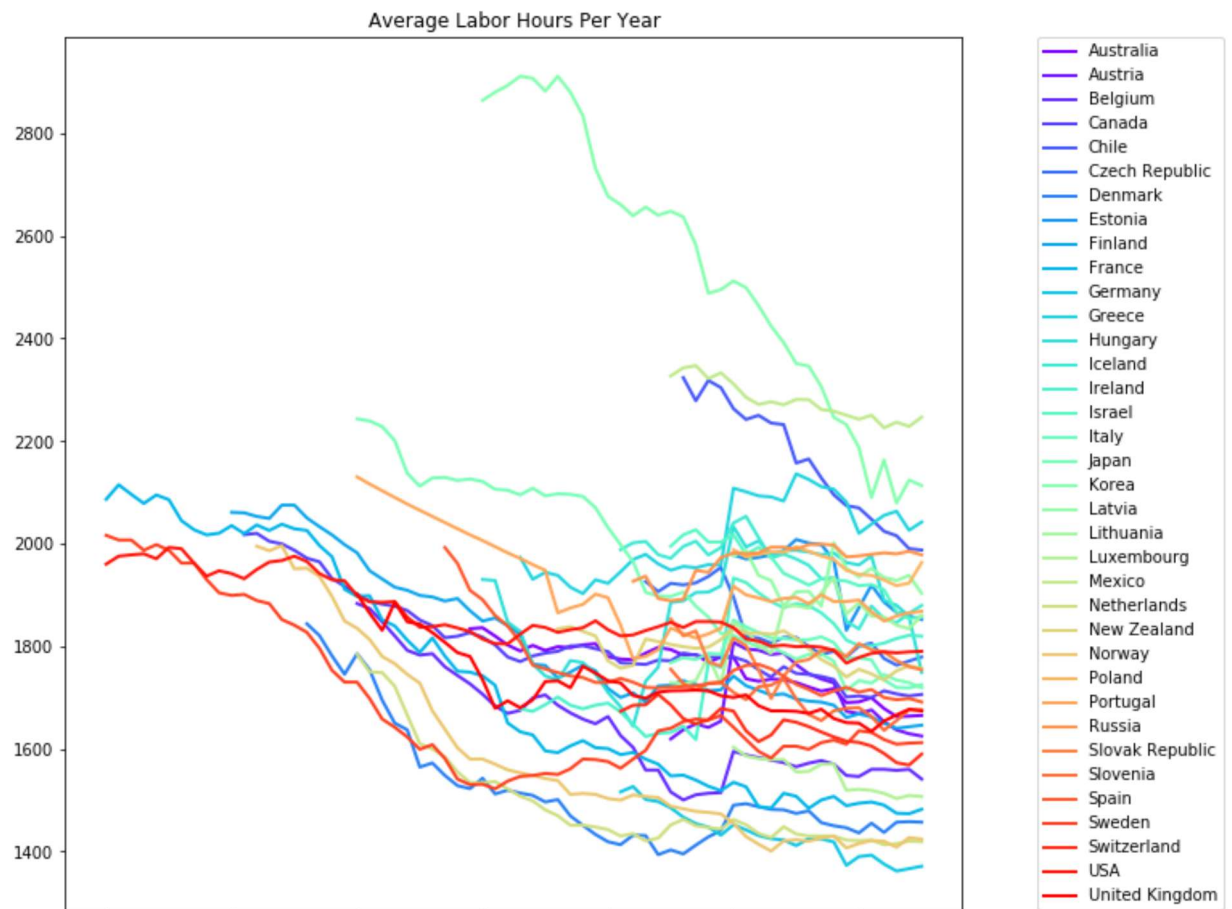


Plot our world labor data

In [100]: `world_historical.sort_index(inplace=True)`

Plot, transposed world_historical dataframe

```
In [101]: world_historical.transpose().plot(figsize=(10,10),colormap='rainbow',linewidth=2,
plt.legend(loc='right',bbox_to_anchor=(1.3,0.5))
plt.show()
```



Which country worked longer hours per year?

```
In [102]: world_historical.index.name='country'
```

```
In [103]: world_historical.groupby('country').max()
```

Out[103]:

	1950	1951	1952	1953	1954	1955	1956
country							
Australia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Austria	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Belgium	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Canada	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Chile	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Czech Republic	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Denmark	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Estonia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Finland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
France	2086.380005	2114.61499	2096.035034	2078.25	2094.825012	2085.534973	2044.694
Germany	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Greece	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Hungary	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Iceland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ireland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Israel	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Italy	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Japan	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Latvia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Lithuania	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Luxembourg	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mexico	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Netherlands	NaN	NaN	NaN	NaN	NaN	NaN	NaN
New Zealand	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Norway	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Poland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Portugal	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Russia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Slovak Republic	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Slovenia	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	1950	1951	1952	1953	1954	1955	1956
country							
Spain	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sweden	2016.000000	2007.00000	2007.000000	1987.00	1998.000000	1987.000000	1962.000000
Switzerland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
USA	1960.000000	1975.50000	1978.000000	1980.00	1970.500000	1992.500000	1990.000000
United Kingdom	NaN	NaN	NaN	NaN	NaN	NaN	NaN

36 rows × 66 columns

```
In [104]: world_historical.groupby('country').min()
```

```
Out[104]:
```

	1950	1951	1952	1953	1954	1955	1956
country							
Australia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Austria	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Belgium	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Canada	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Chile	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Czech Republic	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Denmark	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Estonia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Finland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
France	2086.380005	2114.61499	2096.035034	2078.25	2094.825012	2085.534973	2044.694
Germany	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Greece	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Hungary	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Iceland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ireland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Israel	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Italy	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Japan	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Latvia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Lithuania	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Luxembourg	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mexico	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Netherlands	NaN	NaN	NaN	NaN	NaN	NaN	NaN
New Zealand	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Norway	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Poland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Portugal	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Russia	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Slovak Republic	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Slovenia	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	1950	1951	1952	1953	1954	1955	1956
country							
Spain	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sweden	2016.000000	2007.00000	2007.000000	1987.00	1998.000000	1987.000000	1962.000000
Switzerland	NaN	NaN	NaN	NaN	NaN	NaN	NaN
USA	1960.000000	1975.50000	1978.000000	1980.00	1970.500000	1992.500000	1990.000000
United Kingdom	NaN	NaN	NaN	NaN	NaN	NaN	NaN

36 rows × 66 columns

In []: