Pyladies - 15.05.2017 - hands on data - odp

May 15, 2017

```
In [2]: import pandas as pd
        df = pd.read_excel('table030.xls', header=5, skiprows=[6,7,8,9,10,12], skip_footer=234,
        df.columns = df.columns.map(lambda x: x[0:4])
        df.index.name = None
        df.index = df.index.map(lambda x: x.replace('.', ''))
        df.dropna(axis=0, thresh=None, inplace=True)
```

```
In [3]: df
```

```
Out[3]:                      1950  1960  1970  1980  1981  1982  1983  1984  1985  \
        All ages, crude      11.4  10.6  11.6  11.9  12.0  12.2  12.1  12.4  12.4
        5-14 years            0.2   0.3   0.3   0.4   0.5   0.6   0.6   0.7   0.8
        15-24 years           4.5   5.2   8.8  12.3  12.2  12.1  11.8  12.4  12.8
          15-19 years         2.7   3.6   5.9   8.5   8.6   8.7   8.6   8.9   9.9
          20-24 years         6.2   7.1  12.2  16.1  15.7  15.2  14.6  15.5  15.4
        25-44 years          11.6  12.2  15.4  15.6  16.2  15.8  15.3  15.4  15.0
          25-34 years         9.1  10.0  14.1  16.0  16.3  16.0  15.8  15.6  15.3
          35-44 years        14.3  14.2  16.9  15.4  15.9  15.4  14.7  15.1  14.6
        45-64 years          23.5  22.0  20.6  15.9  16.2  16.8  16.4  16.8  16.3
          45-54 years        20.9  20.7  20.0  15.9  16.1  16.5  16.2  16.3  15.7
          55-64 years        26.8  23.7  21.4  15.9  16.4  17.0  16.6  17.4  16.8
        65 years and over    30.0  24.5  20.8  17.6  17.1  18.4  19.3  19.8  20.4
          65-74 years        29.6  23.0  20.8  16.9  16.2  17.4  17.8  18.9  18.7
          75-84 years        31.1  27.9  21.2  19.1  18.6  20.5  22.2  21.9  23.9
          85 years and over  28.8  26.0  19.0  19.2  17.8  17.6  19.4  18.6  19.4

                             1986  ...   2005  2006  2007  2008  2009  2010  2011  \
        All ages, crude      12.9  ...   11.0  11.2  11.5  11.8  12.0  12.4  12.7
        5-14 years            0.8  ...    0.7   0.5   0.5   0.5   0.6   0.7   0.7
        15-24 years          12.9  ...    9.9   9.8   9.6   9.9  10.0  10.5  11.0
          15-19 years        10.1  ...    7.5   7.1   6.7   7.2   7.5   7.5   8.3
          20-24 years        15.5  ...   12.4  12.5  12.6  12.7  12.6  13.6  13.6
        25-44 years          15.6  ...   13.9  14.0  14.5  14.6  14.6  15.0  15.4
          25-34 years        15.8  ...   12.7  12.7  13.3  13.2  13.1  14.0  14.6
          35-44 years        15.2  ...   15.1  15.2  15.7  15.9  16.1  16.0  16.2
        45-64 years          16.8  ...   15.3  16.0  16.7  17.5  17.9  18.6  18.6
          45-54 years        16.5  ...   16.5  17.2  17.7  18.6  19.2  19.6  19.8
```

```
  55-64 years        17.2  ...   13.7  14.4  15.3  16.0  16.4  17.5  17.1
65 years and over    21.6  ...   14.7  14.3  14.3  14.8  14.8  14.9  15.3
  65-74 years        19.9  ...   12.4  12.4  12.4  13.6  13.7  13.7  14.1
  75-84 years        25.0  ...   16.8  15.8  16.2  16.1  15.8  15.7  16.5
  85 years and over  21.1  ...   18.3  17.3  17.0  16.4  16.4  17.6  16.9

                     2012  2013  2014
All ages, crude      12.9  13.0  13.4
5-14 years            0.8   1.0   1.0
15-24 years          11.1  11.1  11.5
  15-19 years         8.3   8.3   8.7
  20-24 years        13.7  13.7  14.2
25-44 years          15.7  15.5  15.8
  25-34 years        14.7  14.8  15.1
  35-44 years        16.7  16.2  16.6
45-64 years          19.1  19.0  19.5
  45-54 years        20.0  19.7  20.2
  55-64 years        18.0  18.1  18.8
65 years and over    15.4  16.1  16.6
  65-74 years        14.0  15.0  15.6
  75-84 years        16.8  17.1  17.5
  85 years and over  17.8  18.6  19.3

[15 rows x 38 columns]
```

```
In [4]: df.head()
```

```
Out[4]:                  1950  1960  1970  1980  1981  1982  1983  1984  1985  1986  \
        All ages, crude  11.4  10.6  11.6  11.9  12.0  12.2  12.1  12.4  12.4  12.9
        5-14 years        0.2   0.3   0.3   0.4   0.5   0.6   0.6   0.7   0.8   0.8
        15-24 years       4.5   5.2   8.8  12.3  12.2  12.1  11.8  12.4  12.8  12.9
          15-19 years     2.7   3.6   5.9   8.5   8.6   8.7   8.6   8.9   9.9  10.1
          20-24 years     6.2   7.1  12.2  16.1  15.7  15.2  14.6  15.5  15.4  15.5

                         ...  2005  2006  2007  2008  2009  2010  2011  2012  2013  \
        All ages, crude  ...  11.0  11.2  11.5  11.8  12.0  12.4  12.7  12.9  13.0
        5-14 years       ...   0.7   0.5   0.5   0.5   0.6   0.7   0.7   0.8   1.0
        15-24 years      ...   9.9   9.8   9.6   9.9  10.0  10.5  11.0  11.1  11.1
          15-19 years    ...   7.5   7.1   6.7   7.2   7.5   7.5   8.3   8.3   8.3
          20-24 years    ...  12.4  12.5  12.6  12.7  12.6  13.6  13.6  13.7  13.7

                         2014
        All ages, crude  13.4
        5-14 years        1.0
        15-24 years      11.5
          15-19 years     8.7
          20-24 years    14.2

        [5 rows x 38 columns]
```

```
In [5]: df2 = pd.read_excel('Update_111_1.xlsx', index_col=0, header=2, skiprows=[3,4], skip_foo
```

```
In [6]: df2
```

```
Out[6]:       Wild Catch  Farmed Fish  Total Fish Production
        Year
        1950   17.157267     0.549871              17.707138
        1951   19.231613     0.681984              19.913597
        1952   21.132339     0.783993              21.916332
        1953   21.471768     0.923210              22.394978
        1954   23.159135     1.035641              24.194776
        1955   24.314154     1.164253              25.478407
        1956   25.902110     1.155524              27.057634
        1957   26.119223     1.507852              27.627075
        1958   26.608717     1.484374              28.093091
        1959   28.855170     1.608028              30.463198
        1960   30.901492     1.601541              32.503033
        1961   34.502465     1.463642              35.966107
        1962   37.482485     1.525561              39.008046
        1963   38.165946     1.705312              39.871258
        1964   42.266999     1.786919              44.053918
        1965   42.611290     1.960964              44.572254
        1966   46.217910     2.018305              48.236215
        1967   49.130492     2.072726              51.203218
        1968   52.121446     2.210066              54.331512
        1969   50.203587     2.292653              52.496240
        1970   55.350784     2.489182              57.839966
        1971   55.414307     2.658338              58.072645
        1972   50.633068     2.859333              53.492401
        1973   50.318887     2.976385              53.295272
        1974   53.124891     3.150059              56.274950
        1975   51.789776     3.484537              55.274313
        1976   55.102374     3.599727              58.702101
        1977   54.522449     3.985269              58.507718
        1978   56.941870     4.064889              61.006759
        1979   57.466001     4.183220              61.649221
        ...          ...          ...                    ...
        1983   61.409689     5.999410              67.409099
        1984   66.367655     6.677460              73.045115
        1985   67.939544     7.732004              75.671548
        1986   72.810218     8.843956              81.654174
        1987   73.443081    10.220098              83.663179
        1988   87.357491    11.681695              99.039186
        1989   87.924960    12.315219             100.240179
        1990   84.149669    13.074379              97.224048
        1991   83.247335    13.726148              96.973483
        1992   85.062590    15.409688             100.472278
        1993   86.406878    17.802261             104.209139
```

```
1994   91.969659    20.840020             112.809679
1995   92.052943    24.382690             116.435633
1996   93.633925    26.593276             120.227201
1997   92.926515    27.321941             120.248456
1998   85.543098    28.412950             113.956048
1999   91.259461    30.731507             121.990968
2000   93.306179    32.417738             125.723917
2001   90.536416    34.613626             125.150042
2002   90.647461    36.785687             127.433148
2003   87.934364    38.915093             126.849457
2004   92.304240    41.907649             134.211889
2005   92.145097    44.295996             136.441093
2006   89.878707    47.290220             137.168927
2007   90.170168    49.937426             140.107594
2008   89.579537    52.946447             142.525984
2009   89.461456    55.714357             145.175813
2010   88.544684    59.872600             148.417284
2011   91.800000    63.600000             155.400000
2012   90.100000    67.300000             157.400000

[63 rows x 3 columns]
```

```
In [7]: df2.describe()
```

```
Out[7]:        Wild Catch  Farmed Fish  Total Fish Production
        count   63.000000    63.000000              63.000000
        mean    62.809034    15.641514              78.450548
        std     25.221584    18.929883              41.591192
        min     17.157267     0.549871              17.707138
        25%     44.414600     1.989634              46.404235
        50%     59.931941     5.058457              64.990398
        75%     89.003070    26.957608             118.331417
        max     93.633925    67.300000             157.400000
```

```
In [8]: df3 = df.transpose().loc['1950':'2010', ['All ages, crude']]
        df.index.name = 'Year'
        df3.index
```

```
Out[8]: Index(['1950', '1960', '1970', '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'],
              dtype='object')
```

```
In [9]: df4 = df2.loc['1950':'2010', ['Wild Catch']]

        df4.index = df4.index.map(lambda x: str(x))
```

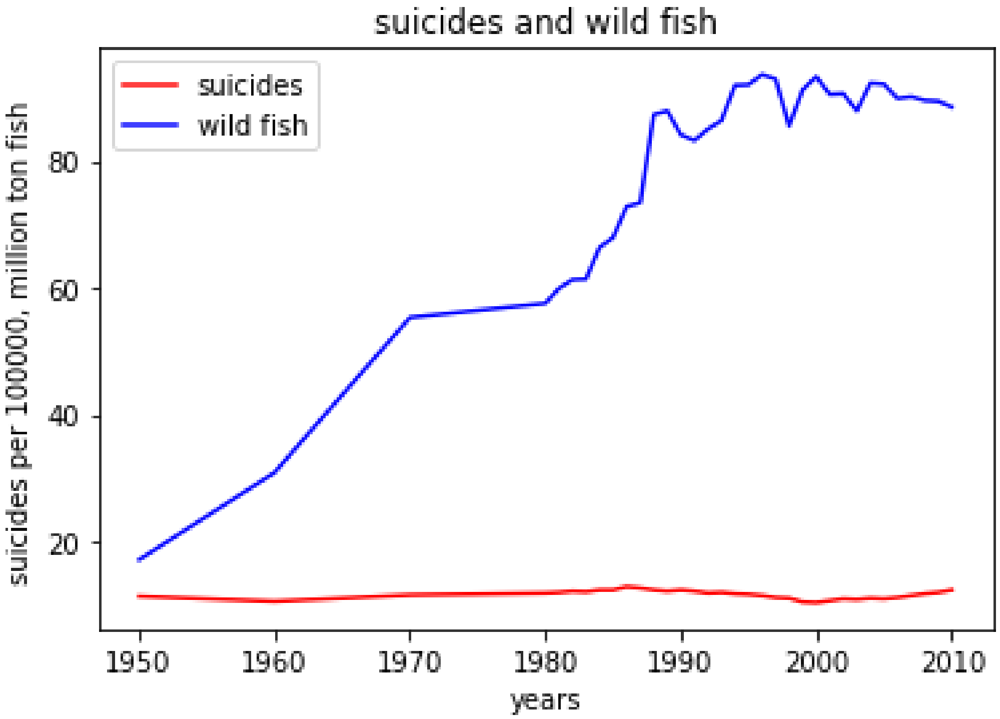
```
In [10]: df5 = pd.concat([df3, df4], axis=1, join='inner')
In [11]: df5
```

```
Out[11]:       All ages, crude  Wild Catch
         1950             11.4   17.157267
         1960             10.6   30.901492
         1970             11.6   55.350784
         1980             11.9   57.579454
         1981             12.0   59.931941
         1982             12.2   61.324728
         1983             12.1   61.409689
         1984             12.4   66.367655
         1985             12.4   67.939544
         1986             12.9   72.810218
         1987             12.7   73.443081
         1988             12.4   87.357491
         1989             12.2   87.924960
         1990             12.4   84.149669
         1991             12.2   83.247335
         1992             11.9   85.062590
         1993             12.0   86.406878
         1994             11.8   91.969659
         1995             11.7   92.052943
         1996             11.5   93.633925
         1997             11.2   92.926515
         1998             11.1   85.543098
         1999             10.5   91.259461
         2000             10.4   93.306179
         2001             10.7   90.536416
         2002             11.0   90.647461
         2003             10.9   87.934364
         2004             11.1   92.304240
         2005             11.0   92.145097
         2006             11.2   89.878707
         2007             11.5   90.170168
         2008             11.8   89.579537
         2009             12.0   89.461456
         2010             12.4   88.544684
```

```
In [12]: import matplotlib.pyplot as plt
         plt.plot(df5.index, df5['All ages, crude'], label='suicides', c='red')
         plt.plot(df5.index,  df5['Wild Catch'], label='wild fish', c='blue')
         plt.title('suicides and wild fish')
         plt.xlabel('years')
         plt.ylabel('suicides per 100000, million ton fish')
         plt.legend()
```

```
Out[12]: <matplotlib.legend.Legend at 0x7fa6c8711128>
```

```
In [13]: plt.show()
```



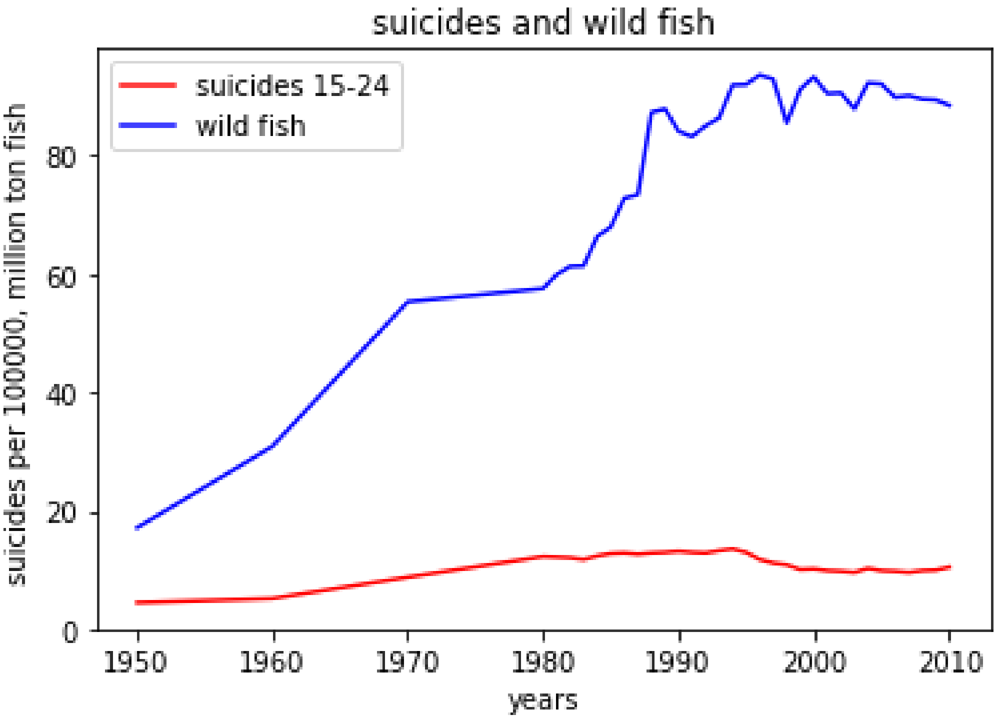
```
In [14]: from scipy.stats.stats import pearsonr
         pearsonr(df5['All ages, crude'], df5['Wild Catch'])

Out[14]: (-0.12143321323444307, 0.49389637606857117)

In [15]: df6 = df.transpose().loc['1950':'2010', ['15-24 years']]

In [17]: import scipy
         scipy.stats.mstats.normaltest(df6['15-24 years'])

Out[17]: NormaltestResult(statistic=15.14765003550238, pvalue=0.00051372368352340075)

In [18]: scipy.stats.mstats.normaltest(df5['Wild Catch'])

Out[18]: NormaltestResult(statistic=20.925754884213799, pvalue=2.8577883539474535e-05)

In [19]: pearsonr(df6['15-24 years'], df5['Wild Catch'])

Out[19]: (0.47095136820787403, 0.0049360369670134428)

In [20]: scipy.stats.mstats.normaltest(df5['All ages, crude'])

Out[20]: NormaltestResult(statistic=2.4409416080931656, pvalue=0.29509120408145417)

In [157]: pearsonr(df5['Wild Catch'], df6['15-24 years'])
```

```
Out[157]: (0.47095136820787403, 0.004936036967013442 8)
```

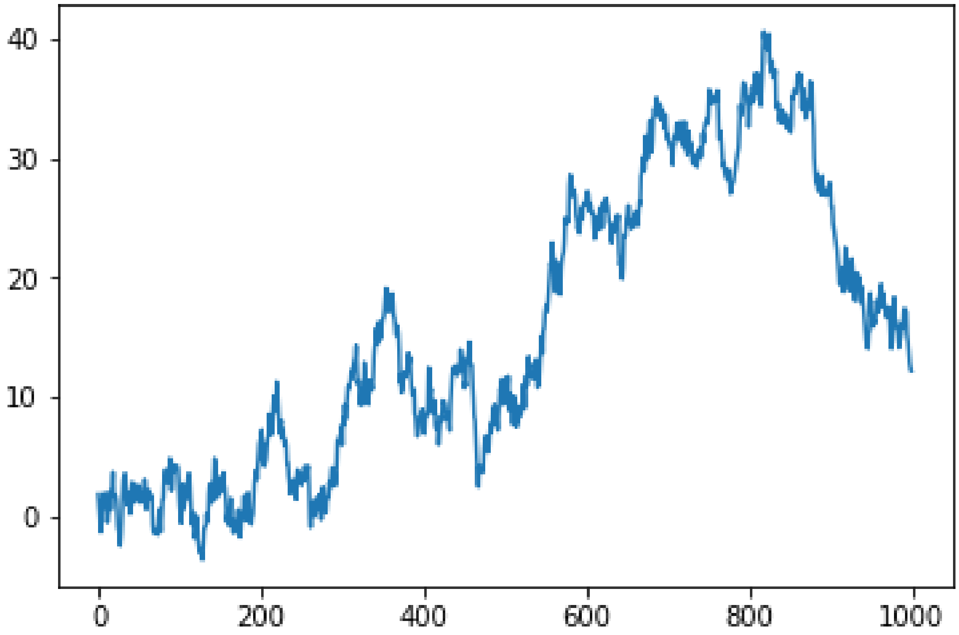
```
In [154]: plt.plot(df5.index, df6['15-24 years'], label='suicides 15-24', c='red')
          plt.plot(df5.index,  df5['Wild Catch'], label='wild fish', c='blue')
          plt.title('suicides and wild fish')
          plt.xlabel('years')
          plt.ylabel('suicides per 100000, million ton fish')
          plt.legend()
```

```
Out[154]: <matplotlib.legend.Legend at 0x7f4fa5b72a90>
```

```
In [155]: plt.show()
```



```
In [159]:
```



```
In [ ]:
```