Week_3_ext_project - GGM

December 1, 2017

1 Project 3: GDP and life expectancy

by Gonzalo Gomez Millan, 1st december 2017, based on Michel Wermelinger, 27 August 2015, updated 5 April 2016

This is the project notebook for Week 3 of The Open University's *Learn to code for Data Analysis* course.

Richer countries can afford to invest more on healthcare, on work and road safety, and other measures that reduce mortality. On the other hand, richer countries may have less healthy lifestyles. Is there any relation between the wealth of a country and the life expectancy of its inhabitants?

The following analysis checks whether there is any correlation between the total gross domestic product (GDP) of a country in 2013 and the life expectancy of people born in that country in 2013.

Additionally, we will try to answer the following questions: - To what extent do the ten countries with the highest GDP coincide with the ten countries with the longest life expectancy? - Which are the two countries in the right half of the plot (higher GDP) with life expectancy below 60 years? - What factors could explain their lower life expectancy compared to countries with similar GDP?

1.1 Getting the data

Two datasets of the World Bank are considered. One dataset, available at http://data.worldbank.org/indicator/NY.GDP.MKTP.CD, lists the GDP of the world's countries in current US dollars, for various years. The use of a common currency allows us to compare GDP values across countries. The other dataset, available at http://data.worldbank.org/indicator/SP.DYN.LE00.IN, lists the life expectancy of the world's countries. The datasets were downloaded as CSV files in March 2016.

In [1]:
```
import warnings
warnings.simplefilter('ignore', FutureWarning)

from pandas import *

YEAR = 2013
GDP_INDICATOR = 'NY.GDP.MKTP.CD'
gdpReset = read_csv('WB GDP 2013.csv')
```

```
LIFE_INDICATOR = 'SP.DYN.LE00.IN'
lifeReset = read_csv('WB LE 2013.csv')
lifeReset.head()
```

1.2 Cleaning the data

Inspecting the data with `head()` and `tail()` shows that:

1. the first 34 rows are aggregated data, for the Arab World, the Caribbean small states, and other country groups used by the World Bank;

- GDP and life expectancy values are missing for some countries.

The data is therefore cleaned by: 1. removing the first 34 rows; - removing rows with unavailable values.

```
In [2]: gdpCountries = gdpReset[34:].dropna()
        lifeCountries = lifeReset[34:].dropna()
```

1.3 Transforming the data

The World Bank reports GDP in US dollars and cents. To make the data easier to read, the GDP is converted to millions of British pounds (the author's local currency) with the following auxiliary functions, using the average 2013 dollar-to-pound conversion rate provided by http://www.ukforex.co.uk/forex-tools/historical-rate-tools/yearly-average-rates.

```
In [3]: def roundToMillions (value):
            return round(value / 1000000)

        def usdToGBP (usd):
            return usd / 1.564768

        GDP = 'GDP (čm)'
        gdpCountries[GDP] = gdpCountries[GDP_INDICATOR].apply(usdToGBP).apply(roundToMillions)
        gdpCountries.head()
```

The unnecessary columns can be dropped.

```
In [4]: COUNTRY = 'country'
        headings = [COUNTRY, GDP]
        gdpClean = gdpCountries[headings]
        gdpClean.head()
```

The World Bank reports the life expectancy with several decimal places. After rounding, the original column is discarded.

```
In [5]: LIFE = 'Life expectancy (years)'
        lifeCountries[LIFE] = lifeCountries[LIFE_INDICATOR].apply(round)
        headings = [COUNTRY, LIFE]
        lifeClean = lifeCountries[headings]
        lifeClean.head()
```

1.4 Combining the data

The tables are combined through an inner join on the common 'country' column.

```
In [6]: gdpVsLife = merge(gdpClean, lifeClean, on=COUNTRY, how='inner')
        gdpVsLife.head()
```

1.5 Calculating the correlation

To measure if the life expectancy and the GDP grow together, the Spearman rank correlation coefficient is used. It is a number from -1 (perfect inverse rank correlation: if one indicator increases, the other decreases) to 1 (perfect direct rank correlation: if one indicator increases, so does the other), with 0 meaning there is no rank correlation. A perfect correlation doesn't imply any cause-effect relation between the two indicators. A p-value below 0.05 means the correlation is statistically significant.

```
In [7]: from scipy.stats import spearmanr

        gdpColumn = gdpVsLife[GDP]
        lifeColumn = gdpVsLife[LIFE]
        (correlation, pValue) = spearmanr(gdpColumn, lifeColumn)
        print('The correlation is', correlation)
        if pValue < 0.05:
            print('It is statistically significant.')
        else:
            print('It is not statistically significant.')
```

```
The correlation is 0.501023238967
It is statistically significant.
```

The value shows a direct correlation, i.e. richer countries tend to have longer life expectancy, but it is not very strong.

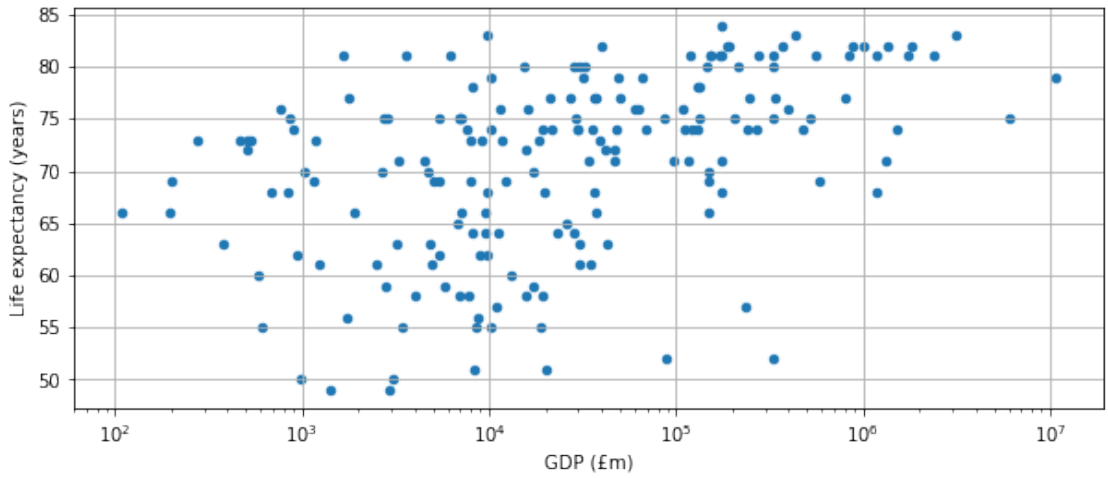
1.6 Showing the data

Measures of correlation can be misleading, so it is best to see the overall picture with a scatterplot. The GDP axis uses a logarithmic scale to better display the vast range of GDP values, from a few million to several billion (million of million) pounds.

```
In [8]: %matplotlib inline
        gdpVsLife.plot(x=GDP, y=LIFE, kind='scatter', grid=True, logx=True, figsize=(10, 4))
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9ecd8b9978>
```

```
Out[8]:
```



The plot shows there is no clear correlation: there are rich countries with low life expectancy, poor countries with high expectancy, and countries with around 10 thousand (104) million pounds GDP have almost the full range of values, from below 50 to over 80 years. Towards the lower and higher end of GDP, the variation diminishes. Above 40 thousand million pounds of GDP (3rd tick mark to the right of 104), most countries have an expectancy of 70 years or more, whilst below that threshold most countries' life expectancy is below 70 years.

```
In [9]: gdpVsLife[(gdpVsLife[LIFE] < 60) & (gdpVsLife[GDP] > 100000)]
```

The plot shows too two countries in the right half of the plot (higher GDP) with life expectancy below 60 years. Those countries are Nigeria and South Africa. The factors that could explain their lower life expectancy compared to countries with similar GDP are related with heath care, because if we review their tuberculosis deaths we can find Nigeria and South Africa as second and eleventh in the highest rank of deaths.

Comparing the 10 poorest countries and the 10 countries with the lowest life expectancy shows that total GDP is a rather crude measure. The population size should be taken into account for a more precise definiton of what 'poor' and 'rich' means. Furthermore, looking at the countries below, droughts and internal conflicts may also play a role in life expectancy.

```
In [12]: # the 10 countries with lowest GDP
         gdpVsLife.sort_values(GDP).head(10)

In [13]: # the 10 countries with lowest life expectancy
         gdpVsLife.sort_values(LIFE).head(10)
```

1.6.1 Comparing richest countries and longest life expectancy countries

```
In [14]: # the 10 countries with highest GDP
         gdpVsLife.sort_values(GDP).tail(10)

In [15]: # the 10 countries with longest life expectancy
         gdpVsLife.sort_values(LIFE).tail(10)
```

```
In [17]: from scipy.stats import spearmanr

         gdpColumn_t = gdpVsLife.sort_values(GDP).tail(10)[GDP]
         lifeColumn_t = gdpVsLife.sort_values(LIFE).tail(10)[LIFE]
         (correlation, pValue) = spearmanr(gdpColumn, lifeColumn)
         print('The correlation is', correlation)
         if pValue < 0.05:
             print('It is statistically significant.')
         else:
             print('It is not statistically significant.')
```

```
The correlation is 0.501023238967
It is statistically significant.
```

Comparing the 10 richest countries and the 10 countries with the longest life expectancy shows that only three countries are in both groups. Threre is a low correlation between both so the total GDP is a rather crude measure. The population size should be taken into account for a more precise definiton of what 'poor' and 'rich' means.

1.6.2 Comparing richest countries and longest life expectancy countries

The population information is got from WB and is cleaned.

```
In [24]: # Getting world's population
         YEAR = 2013
         POPULATION_INDICATOR = 'SP.POP.TOTL'

         # from pandas_datareader import wb

         # populationWB = wb.download(indicator=POPULATION_INDICATOR, country='all', start=YEAR,
         # populationWB.head()

In [25]: #populationReset = populationWB.reset_index()
         populationReset = read_csv('WB POP 2013.csv')
         populationReset.head()

In [26]: headings = [COUNTRY, POPULATION_INDICATOR]
         populationData = populationReset[47:].reset_index()[headings]
         populationData.head()

In [27]: # Getting world's GDP and life expentancy

         import warnings
         warnings.simplefilter('ignore', FutureWarning)

         from pandas import *

         YEAR = 2013
```

```
GDP_INDICATOR = 'NY.GDP.MKTP.CD'
gdpReset = read_csv('WB GDP 2013.csv')

LIFE_INDICATOR = 'SP.DYN.LE00.IN'
lifeReset = read_csv('WB LE 2013.csv')
lifeReset.head()
```

In [28]:
```
# Removing unuseful lines and null data

gdpCountries = gdpReset[34:].dropna()
lifeCountries = lifeReset[34:].dropna()

COUNTRY = 'country'
GDP_INDICATOR = 'NY.GDP.MKTP.CD'

headings = [COUNTRY, GDP_INDICATOR]
gdpClean = gdpCountries[headings]
gdpClean.head()
```

In [29]:
```
LIFE = 'Life expectancy (years)'
lifeCountries[LIFE] = lifeCountries[LIFE_INDICATOR].apply(round)
headings = [COUNTRY, LIFE]
lifeClean = lifeCountries[headings]
lifeClean.head()
```

In [30]:
```
gdpVsLife = merge(gdpClean, lifeClean, on=COUNTRY, how='inner')
gdpVsLife.head()
```

In [31]:
```
# Merging population, gdp and life expectancy

gdpPopulationLife = merge(gdpVsLife, populationData, on=COUNTRY, how='inner')
gdpPopulationLife.head(5)
```

In [32]:
```
# Deleting columns not needed

GDP_X_CAP = 'GDP per capita'
gdpPopulationLife[GDP_X_CAP] = gdpPopulationLife[GDP_INDICATOR]/gdpPopulationLife[POPUL
gdpPopulationLife[GDP_X_CAP] = gdpPopulationLife[GDP_X_CAP].apply(round)
headings = [COUNTRY, LIFE, GDP_X_CAP]
gdpPopulationLifeClean = gdpPopulationLife[headings]
gdpPopulationLifeClean.head()
```

In [34]:
```
gdpPopulationLifeClean.sort_values(GDP_X_CAP).tail(10)
```

In [35]:
```
gdpPopulationLifeClean.sort_values(LIFE).tail(10)
```

1.7 Calculating the correlation

To measure if the life expectancy and the GDP percapita grow together, the Spearman rank correlation coefficient is used. It is a number from -1 (perfect inverse rank correlation: if one indicator

increases, the other decreases) to 1 (perfect direct rank correlation: if one indicator increases, so does the other), with 0 meaning there is no rank correlation. A perfect correlation doesn't imply any cause-effect relation between the two indicators. A p-value below 0.05 means the correlation is statistically significant.

In [36]:
```
from scipy.stats import spearmanr

gdpPerCapColumn = gdpPopulationLife[GDP_X_CAP]
lifeColumn = gdpPopulationLife[LIFE]
(correlation, pValue) = spearmanr(gdpPerCapColumn, lifeColumn)
print('The correlation is', correlation)
if pValue < 0.05:
    print('It is statistically significant.')
else:
    print('It is not statistically significant.')
```

```
The correlation is 0.857342999366
It is statistically significant.
```

The value shows a direct correlation and look like strong, i.e. richer countries tend to have longer life expectancy.

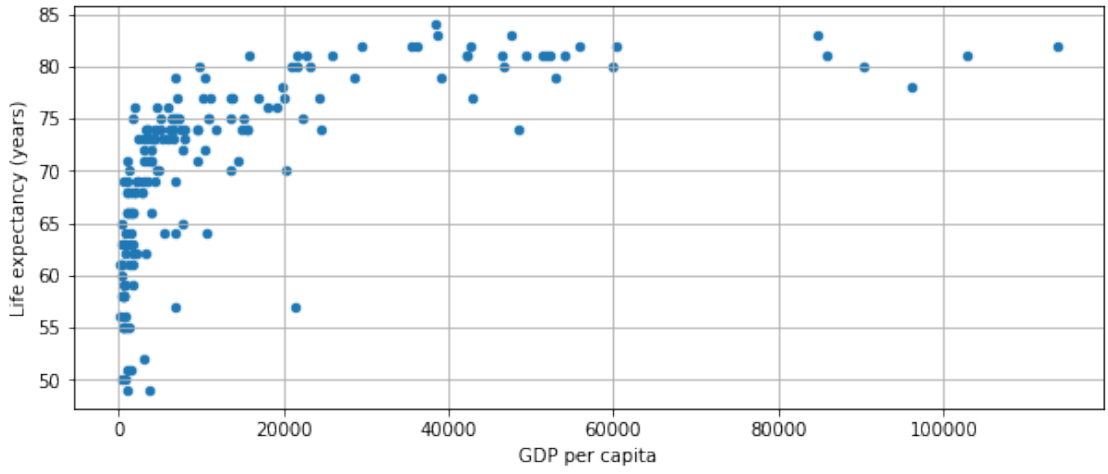
1.8 Showing the data

Measures of correlation can be misleading, so it is best to see the overall picture with a scatterplot. The GDP per capita axis uses USD.

In [37]:
```
%matplotlib inline
gdpPopulationLife.plot(x=GDP_X_CAP, y=LIFE, kind='scatter', grid=True, figsize=(10, 4))
```

Out[37]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f9ec9e5f748>`

Out[37]:



1.9 Conclusions

To sum up, there is no strong correlation between a country's wealth and the life expectancy of its inhabitants: there is often a wide variation of life expectancy for countries with similar GDP, countries with the lowest life expectancy are not the poorest countries, and countries with the highest expectancy are not the richest countries. Nevertheless there is some relationship, because the vast majority of countries with a life expectancy below 70 years is on the left half of the scatterplot.

Using the NY.GDP.PCAP.PP.CD indicator, GDP per capita in current 'international dollars', would make for a better like-for-like comparison between countries, because it would take population and purchasing power into account. Using more specific data, like expediture on health, could also lead to a better analysis. And using this indicatos the correlation beween wealth and life expentancy is stronger but do not definitive.