Affluence and dependency

Sebastian Loth

Dataset(s)

Which dataset did you use of the following:

- World Development Indicators Dataset found at https://www.kaggle.com/worldbank/world-development-indicators

Motivation

Global poverty is decreasing as more countries evolve their economies.

The needs and purchasing behaviour of a society depend on how its affluence is distributed. The degree of financial dependency determines whether earners spend their income on basic food and housing or on communications, household goods or even luxury goods.

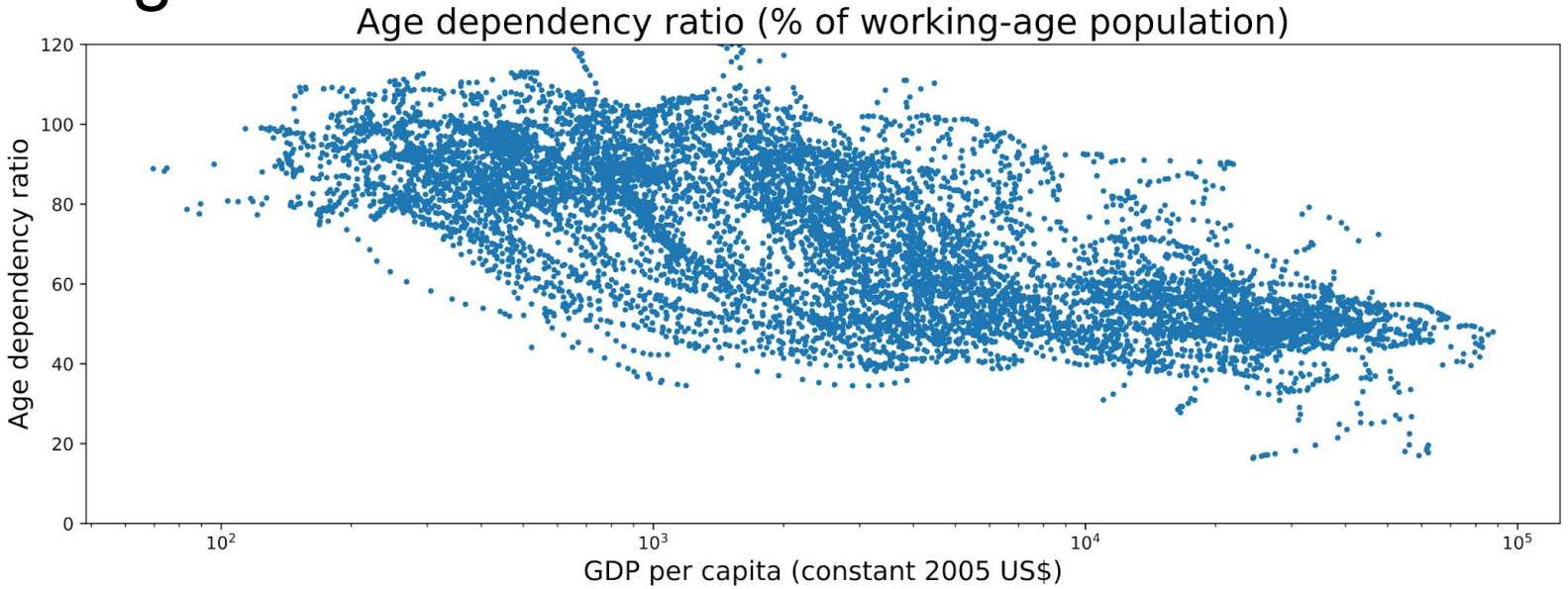
These insights are relevant to companies that wish to understand long-term trends in industrialised countries, expand in the developing world and investors looking into the stock market of developing countries.

Research Question(s)

I investigated the relationship of a country's GDP (per capita) and its ratio of the financially dependent population.

In order to understand this further, I looked at the relation of GDP (per capita) and the ratio of dependency in general, in the elderly and children.

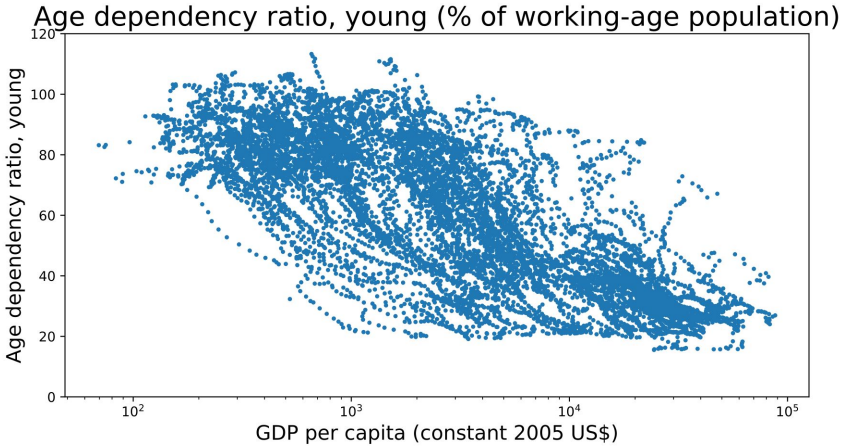
Findings

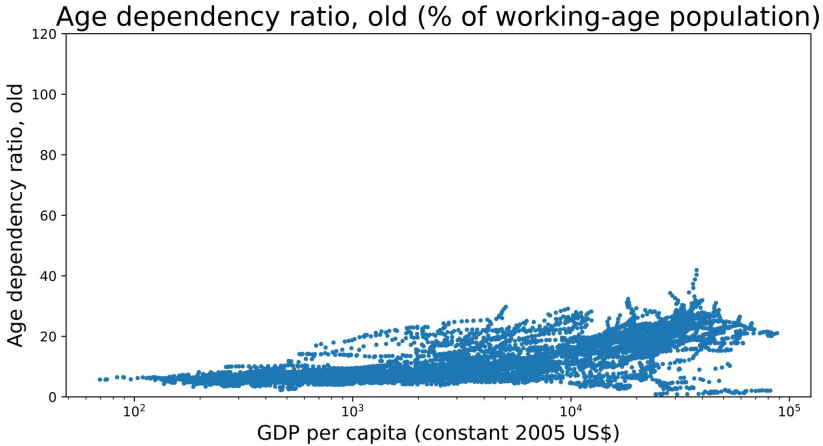


As the GDP per capita increases, the ratio of the financially dependent population decreases despite the fact that higher incomes could support more people.

Country-wise spearman's correlation after averaging with Fisher's z-transformation: $\varrho(219)=-0.49$.

Findings

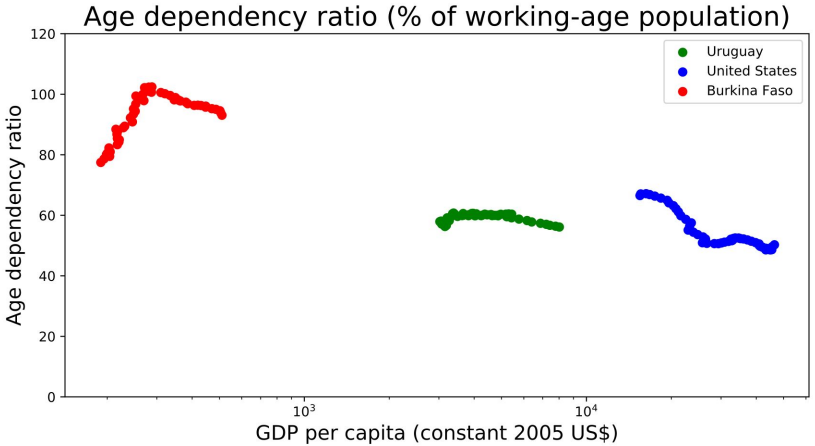


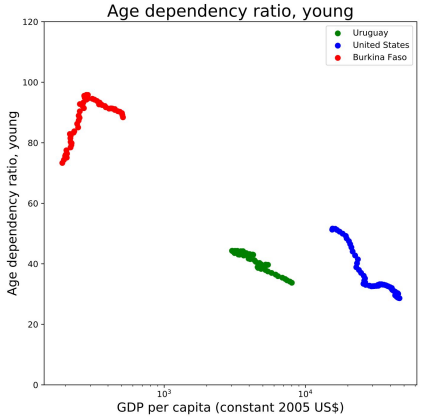


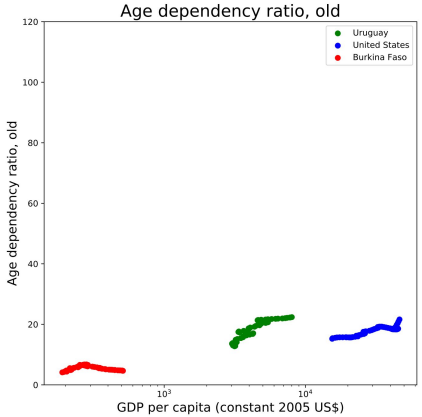
More detailed analysis showed that the ratio of children decreases (left) but ratio dependent elderly increases (right).

Country-wise spearman's correlation after averaging with Fisher's z-transformation: $\varrho(219)=-0.54$ and $\varrho(219)=+0.37$ for children and elderly respectively.

Findings







At the level of single countries, these trends do not always manifest perfectly.

The exemplary industrialised (USA) and developing countries (Uruguay) follow this trend whereas Burkina Faso's data point to influences outside the analysis.

Summary

General trend: The ratio of financially dependent people reduces as the GDP per capita increases. Thus, each earner has more freely spendable income. This might be spent on better or larger housing, communication, or luxury goods. Companies in these sectors could grow over-proportionally compared to GDP.

Children: The ratio of dependent children decreases with growing GDP. Thus, parent earners buy less products but are able to spend more per item. Companies focussing on higher priced products should benefit from this development.

Elderly: The share of elderly dependents increases with GDP. Companies in care and health will benefit as will insurers providing long-term financial products.

Acknowledgements

Did you use other informal analysis to inform your work?

I have drawn from some general knowledge in order to make some analytical conclusions. However, I cannot pinpoint this to any specific resource.

Did you get feedback on your work by friends or colleagues?

I did not receive feedback and completed the assignment alone.

References

If applicable, report any references you used in your work. For example, you may have used a research paper from X to help guide your analysis. You should cite that work here. If you did all the work on your own, please state this.

I have used https://www.kaggle.com/worldbank/world-development-indicators Specifically, the list of the available indicators in the dataset.

I also used stackoverflow and the pandas documentation for programming help.

In [1]:

```
# Data Source: https://www.kaggle.com/worldbank/world-development-indicators
# Folder: 'world-development-indicators'
```

World development indicators

Research question

I would like to explore the relation between GDP per capita and dependency. Dependency refers to parts of the population that cannot support themselves and is divided into children and elderly.

We could expect that the number of dependents increases with GDP growths. This could be attributed simply to the fact that people in work are able to support more family members and other dependents. From this assumption, the number of children would grow. In addition, life expectancy increases with general affluence approximated by the GDP. This implies a growing number of elderly.

In contrast to this, the growing affluence could lead to a decreased number of children. With growing income, less working members of a family are needed in order to support their parents. Thus, the generation of the parents may decide to have less children.

Data and variables

In order to investigate this, I will use four indicators from the World Development Indicators (Data Source: https://www.kaggle.com/worldbank/world-development-indicators (https://www.kaggle.com/worldbank/world-development-indicators)).

- A metric of GDP:
  - NY.GDP.PCAP.KD: GDP per capita (constant 2005 USdollar)
  - The constant US Dollar ratio avoids any effects of inflation and makes the numbers more comparable across years and countries. Furthermore, the GDP per capita makes the numbers more comparable across countries of different sizes.
- Three metrics of dependency:
  - SP.POP.DPND: Age dependency ratio (% of working-age population)
  - SP.POP.DPND.OL: Age dependency ratio, old (% of working-age population)
  - SP.POP.DPND.YG: Age dependency ratio, young (% of working-age population)
  - These metrics cover the general dependency and the two variables that split between children and elderly.

In [2]:

```
import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
```

Data acquisition

In [3]:

```
# read data
data = pd.read_csv('../Week5_Visualization/world-development-indicators/Indicators.csv'
)
data.shape
```

Out[3]:

```
(5656458, 6)
```

In [4]:

```
# check whether this has worked
data.head()
```

Out[4]:

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
0	Arab World	ARB	Adolescent fertility rate (births per 1,000 wo...	SP.ADO.TFRT	1960	1.335609e+02
1	Arab World	ARB	Age dependency ratio (% of working-age populat...	SP.POP.DPND	1960	8.779760e+01
2	Arab World	ARB	Age dependency ratio, old (% of working-age po...	SP.POP.DPND.OL	1960	6.634579e+00
3	Arab World	ARB	Age dependency ratio, young (% of working-age ...	SP.POP.DPND.YG	1960	8.102333e+01
4	Arab World	ARB	Arms exports (SIPRI trend indicator values)	MS.MIL.XPRT.KD	1960	3.000000e+06

First, I have to collect the data such that year and country match for each data point.

Data preparation

In [5]:

```
# collect a data set that only contains the relevant rows from the indicator fields
indicators = ['NY.GDP.PCAP.KD', 'SP.POP.DPND', 'SP.POP.DPND.OL', 'SP.POP.DPND.YG']
gdp = pd.DataFrame(data[data.IndicatorCode.isin(indicators)])
# replace the . with _ for better access of the data
gdp.IndicatorCode = gdp.IndicatorCode.str.replace('.', '_')
indicators = gdp.IndicatorCode.unique()
```

In [6]:

```
# check whether filtering and collecting the data has worked
gdp.head()
```

Out[6]:

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
1	Arab World	ARB	Age dependency ratio (% of working-age populat...	SP_POP_DPND	1960	87.797601
2	Arab World	ARB	Age dependency ratio, old (% of working-age po...	SP_POP_DPND_OL	1960	6.634579
3	Arab World	ARB	Age dependency ratio, young (% of working-age ...	SP_POP_DPND_YG	1960	81.023330
81	Caribbean small states	CSS	Age dependency ratio (% of working-age populat...	SP_POP_DPND	1960	90.542561
82	Caribbean small states	CSS	Age dependency ratio, old (% of working-age po...	SP_POP_DPND_OL	1960	7.874999

In [7]:

```
# collect the indicator's names and codes
labels = pd.DataFrame()
labels['Code'] = gdp.IndicatorCode.unique()
labels['Name'] = gdp.IndicatorName.unique()
```

In [8]:

```
# check whether collecting names and codes has worked as intended
labels
```

Out[8]:

	Code	Name
0	SP_POP_DPND	Age dependency ratio (% of working-age populat...
1	SP_POP_DPND_OL	Age dependency ratio, old (% of working-age po...
2	SP_POP_DPND_YG	Age dependency ratio, young (% of working-age ...
3	NY_GDP_PCAP_KD	GDP per capita (constant 2005 US$)

Data exploration

In [9]:

```
# define a function that returns the name for labeling the graphs below
def get_label(code):
    """Returns the string name of the indicator matching the code provided. Note: this
 is not failsafe if the code is unknown.
    """
    return labels[labels.Code==code].Name.iloc[0]
```

In [10]:

```
# delete columns that make the restructuring/reshaping of the data more difficult
del gdp['CountryCode']
del gdp['IndicatorName']
```

In [11]:

```
# check whether this has worked
gdp.head()
```

Out[11]:

	CountryName	IndicatorCode	Year	Value
1	Arab World	SP_POP_DPND	1960	87.797601
2	Arab World	SP_POP_DPND_OL	1960	6.634579
3	Arab World	SP_POP_DPND_YG	1960	81.023330
81	Caribbean small states	SP_POP_DPND	1960	90.542561
82	Caribbean small states	SP_POP_DPND_OL	1960	7.874999

In order to make this analysis properly, I need a data point that consists of all four indicator values for a given year and country. An easy way of achieving this from the filtered data set was to restructure using the .pivot()-method. However, not all four indicators are provided for each combination of country and year. As a result, the command will not execute.

Alternative approach creating a dataframe for each variable and merge the frames one by one. This way, a NaN-value will be set when ever one of the indicators is missing.

In [12]:

```
# make a dictionary of dataframes such that each entry is a DataFrame of one indicator
helper = {}
for i in indicators:
    helper["{0}".format(i)] = (gdp[gdp['IndicatorCode']==i]).rename({'Value':i}, axis=
'columns')
    del helper[i]['IndicatorCode']
```

In [13]:

```
# merge the indicators in order to get the desired data structure of one column per ind
icator and one row per year-country pair
gdp = pd.DataFrame(helper[indicators[0]])
for i in indicators[1:,]:
    gdp = pd.merge(gdp, helper[i], on=['CountryName', 'Year'], how='outer')
```

In [14]:

```
# check the shape of this
gdp.shape
```

Out[14]:

```
(12877, 6)
```

In [15]:

```
# drop the NaNs and check the shape again
gdp = gdp.dropna()
gdp.shape
```

Out[15]:

```
(9497, 6)
```

In [16]:

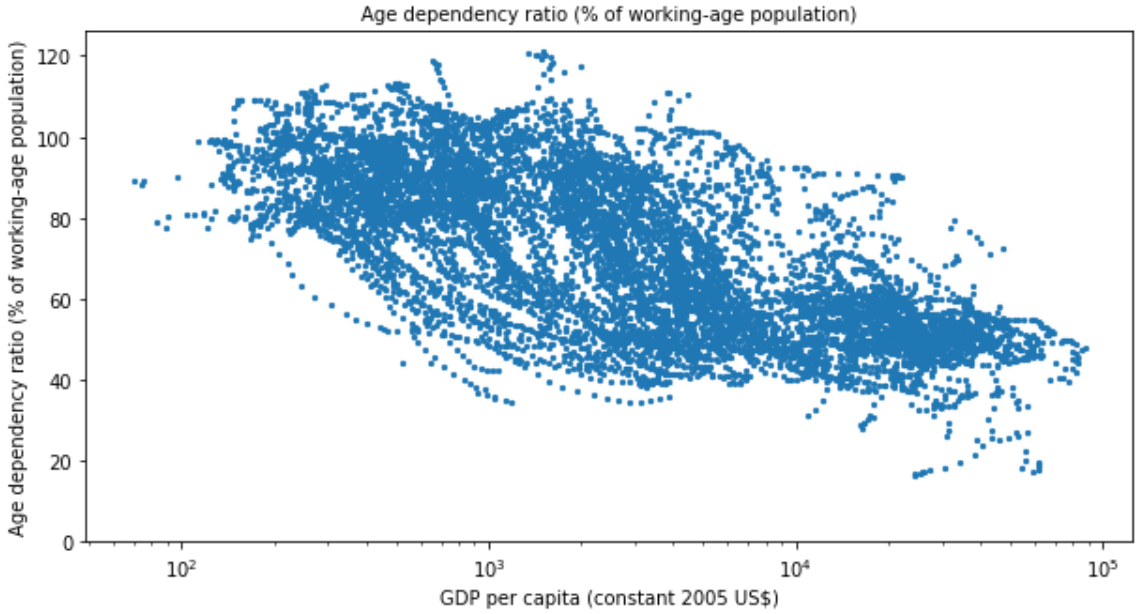
```
# make x and y variables for initial plot
x_variable = 'NY_GDP_PCAP_KD'
y_variable = ['SP_POP_DPND', 'SP_POP_DPND_OL', 'SP_POP_DPND_YG']
```

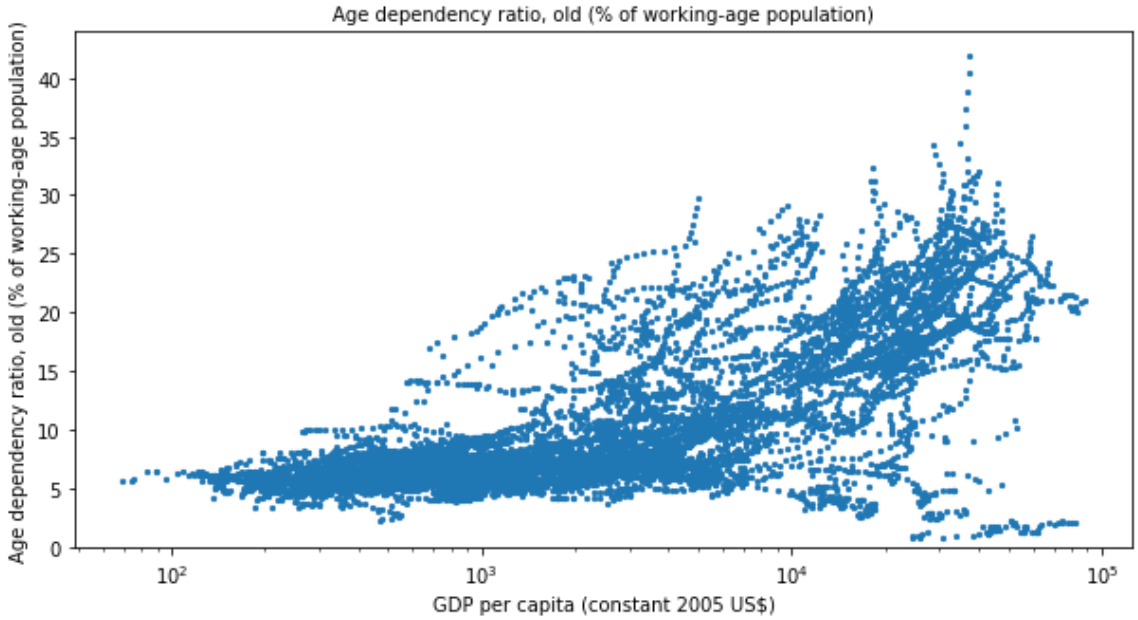
Global view

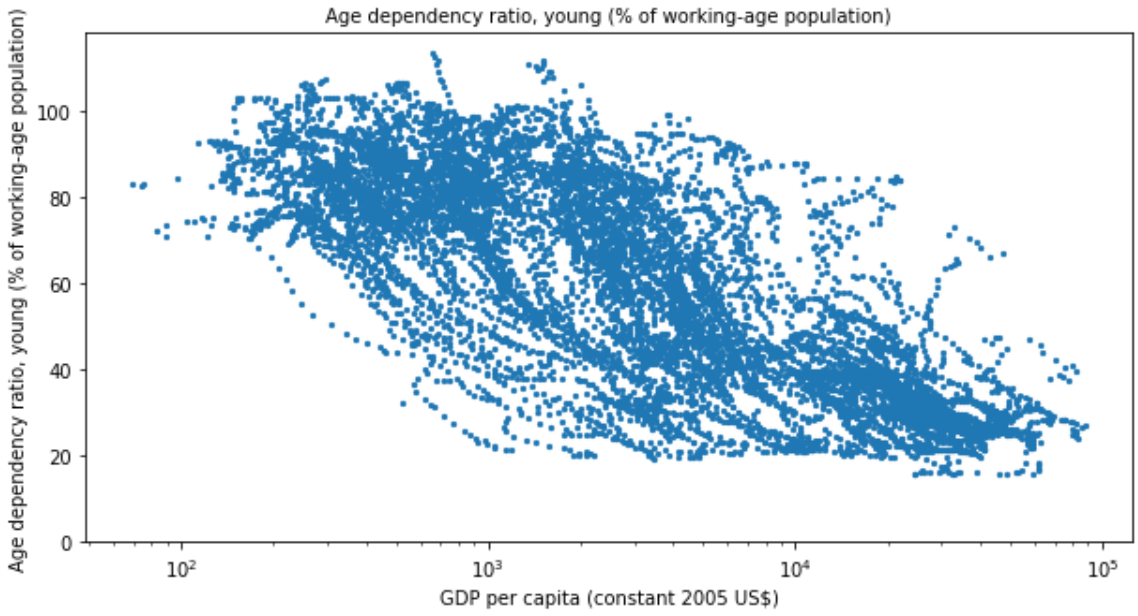
In [17]:

```
# make one scatter plot for the three variables of interest
for i in y_variable:
    fig, axis = plt.subplots(figsize=(10,5))
    axis.set_title(get_label(i),fontsize=10)
    axis.set_xlabel(get_label(x_variable),fontsize=10)
    axis.set_ylabel(get_label(i),fontsize=10)

    axis.scatter(gdp[x_variable], gdp[i], s=5)
    # switch to log-scale of the GDP
    plt.xscale('log')
    plt.ylim(0,)
    plt.show()
```


Age dependency ratio (% of working-age population)
Age dependency ratio (% of working-age population)
0
20
40
60
80
100
120
10^2
10^3
10^4
10^5
GDP per capita (constant 2005 US$)


Age dependency ratio, old (% of working-age population)
Age dependency ratio, old (% of working-age population)
0
5
10
15
20
25
30
35
40
10^2
10^3
10^4
10^5
GDP per capita (constant 2005 US$)



The graphs show that there is a clear trend that the dependency reduces as the per capita GDP increases. That means the greater the income per person the less people are age dependent on these incomes. However, a closer inspection shows that the young age dependency decreases overproportionally whereas old age dependency increases. That means that there are less children in affluent societies but more elderly. Note that this analysis does analyse any temporal relation or causation.

Country-wise exploration

In [18]:

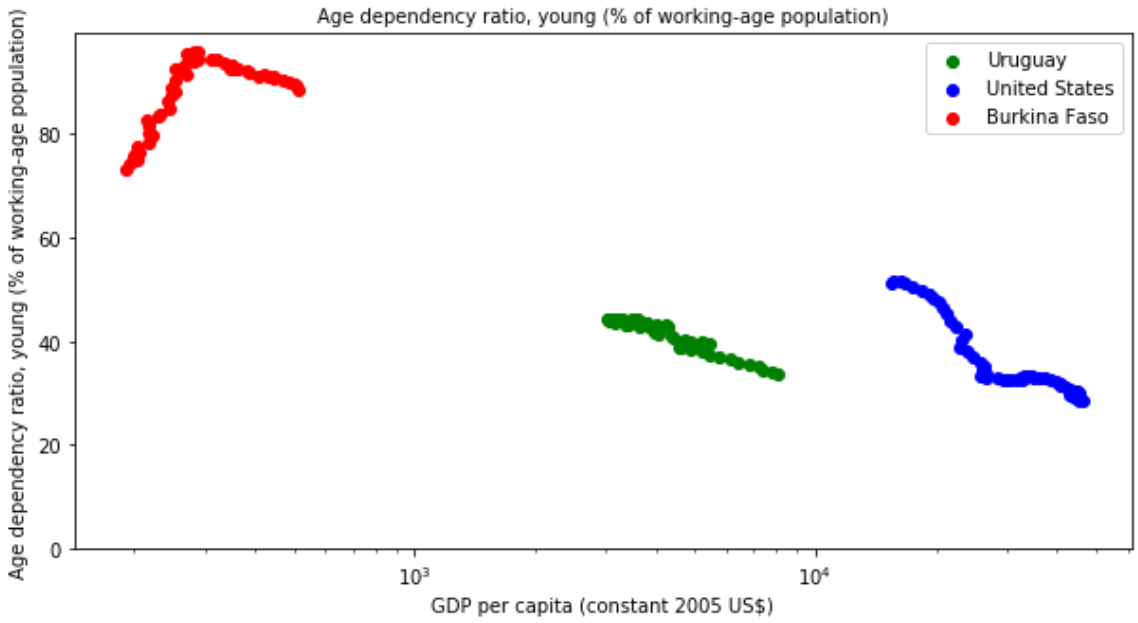
```
# pick three countries with different GDP
countries = ['Uruguay', 'United States', 'Burkina Faso']
colours = ['green', 'blue', 'red']

# plot the graphs with different colours for each country
for i in y_variable:
    fig, axis = plt.subplots(figsize=(10,5))
    axis.set_title(get_label(i),fontsize=10)
    axis.set_xlabel(get_label(x_variable),fontsize=10)
    axis.set_ylabel(get_label(i),fontsize=10)
    plt.xscale('log')

    for country, colour in zip(countries, colours):
        axis.scatter(gdp[gdp.CountryName==country][x_variable], gdp[gdp.CountryName==co
untry][i], c=colour)
    plt.gca().legend(countries)
    plt.ylim(0,)
    plt.show()
```


Age dependency ratio (% of working-age population)
Uruguay
United States
Burkina Faso
Age dependency ratio (% of working-age population)
100
80
60
40
20
0
10^3
10^4
GDP per capita (constant 2005 US$)


Age dependency ratio, old (% of working-age population)
Uruguay
United States
Burkina Faso
Age dependency ratio, old (% of working-age population)
20
15
10
5
0
10^3
10^4
GDP per capita (constant 2005 US$)



A similar finding is found when looking at individual countries. The graphs highlight that the increase in old age dependency is overproportional in the US (industrialised country) and even stronger when countries are developing towards industrialised (e.g., Uruguay). In both countries the dependent younger population decreases with increasing GDP. The more difficult economic situation in Burkina Faso is illustrated by the big gap in GDP per capita. Also, the curves point to seismic change in the society that is typically driven by war, famine and disease. In this case, all of them might hold true. The investigation of this is, however, outside of this project.

Data analysis

In order to assess the correlations appropriately, I use a Spearman correlation. A linear relationship between the GDP and depency is not expected. The graphs show that the relation holds in terms of direction (one increases, the other decreases) but that defining clear term of relation is more complex then a linear function.

The Spearmann correlation sorts the values of each component by rank and tests correlation of the ranks. Thus, it provides the required insights. Kendall's correlation compares whether the pairs of the ranks match between the variables. However, this may underestimate the presence of a general trend.

The analysis has to be conducted by country.

In [19]:

```
# check size of the data frame
correlations = pd.DataFrame(gdp, copy=True)
del correlations['Year']
len(correlations.CountryName.unique())
```

Out[19]:

219

In [20]:

```
# compute correlation coefficients by country
correlations = correlations.groupby('CountryName').corr(method='spearman')
# average correlation coefficients
# 1: make Fisher's z transformation
correlations = correlations.replace(to_replace=1.0, value=np.nan)
correlations[correlations.columns].applymap(lambda x: np.arctanh(x) if x!=0.0 else 0)
# 2: compute mean values per column and correlation type
correlations = correlations.reset_index(level=['CountryName'])
correlations.index.name = 'variable'
correlations = correlations.reset_index(level='variable')
correlations = correlations.groupby('variable').mean()
# 3: transform back standard correlation coefficients
correlations[indicators] = np.arctan(correlations[indicators])
correlations
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:6: Runtim
eWarning: divide by zero encountered in arctanh
```

Out[20]:

variable	NY_GDP_PCAP_KD	SP_POP_DPND	SP_POP_DPND_OL	SP_POP_DPND_Y
NY_GDP_PCAP_KD	NaN	-0.489846	0.366268	-0.5404
SP_POP_DPND	-0.489846	NaN	-0.250702	0.7402
SP_POP_DPND_OL	0.366268	-0.250702	NaN	-0.3644
SP_POP_DPND_YG	-0.540485	0.740238	-0.364451	N

In [21]:

```
target = pd.DataFrame(correlations.NY_GDP_PCAP_KD.dropna()).reset_index(level=['variabl
e'])
target.variable = target.variable.apply(lambda x: get_label(x))
target = target.rename(columns={'NY_GDP_PCAP_KD':get_label('NY_GDP_PCAP_KD')})
target
```

Out[21]:

	variable	GDP per capita (constant 2005 US$)
0	Age dependency ratio (% of working-age populat...	-0.489846
1	Age dependency ratio, old (% of working-age po...	0.366268
2	Age dependency ratio, young (% of working-age ...	-0.540485

The target table presents the findings in terms of correlation coefficients. There is a negative correlation of GDP and dependency indicating that dependency decreases as GDP per capita increases. However, by splitting this into dependence from old age and dependency in childhood, we can two different trends. Dependency from children decreases whereas it increases for elderly.

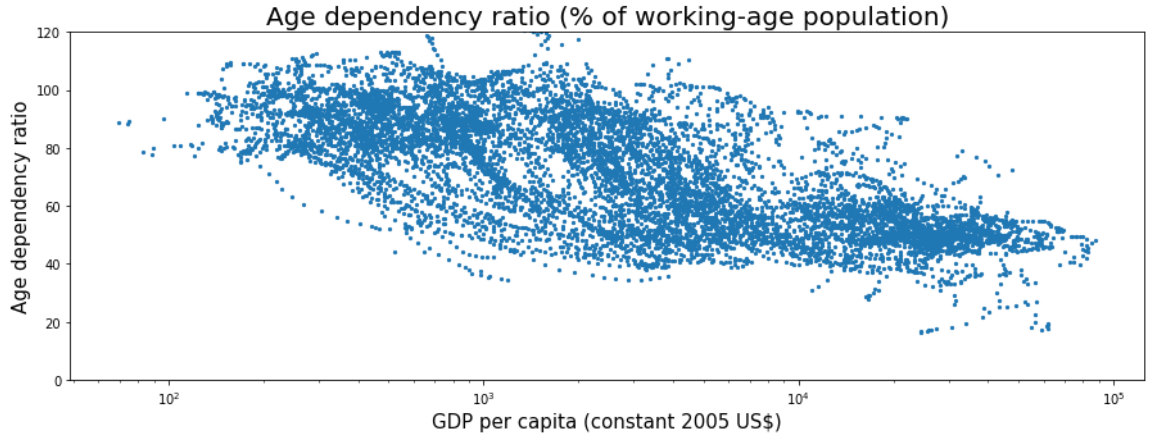
Data presentation

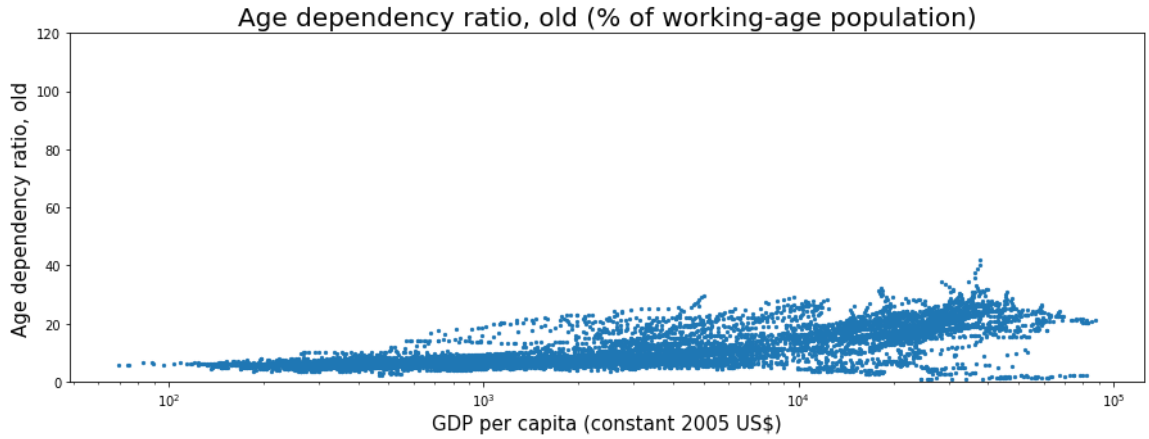
Produce the graphs required for the presentation in an export and import friendly way.

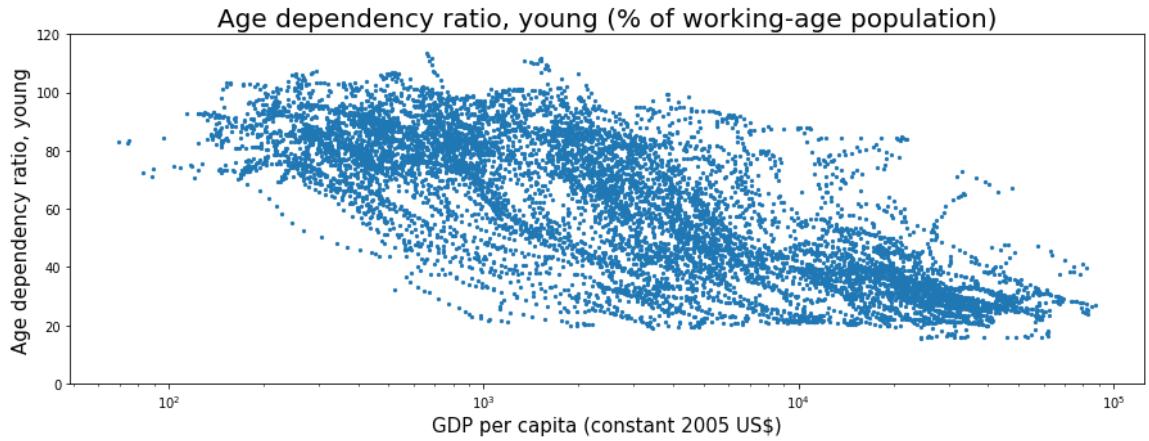
In [22]:

```
# make one scatter plot for the three variables of interest
for i in y_variable:
    fig, axis = plt.subplots(figsize=(15,5))
    axis.set_title(get_label(i),fontsize=20)
    axis.set_xlabel(get_label(x_variable),fontsize=15)
    axis.set_ylabel(get_label(i).split(' (')[0],fontsize=15)

    axis.scatter(gdp[x_variable], gdp[i], s=5)
    # switch to log-scale of the GDP
    plt.xscale('log')
    plt.ylim(0,120)
    fig.savefig("global_{}.svg".format(i))
```


Age dependency ratio (% of working-age population)
Age dependency ratio
GDP per capita (constant 2005 US$)


Age dependency ratio, old (% of working-age population)
Age dependency ratio, old
GDP per capita (constant 2005 US$)

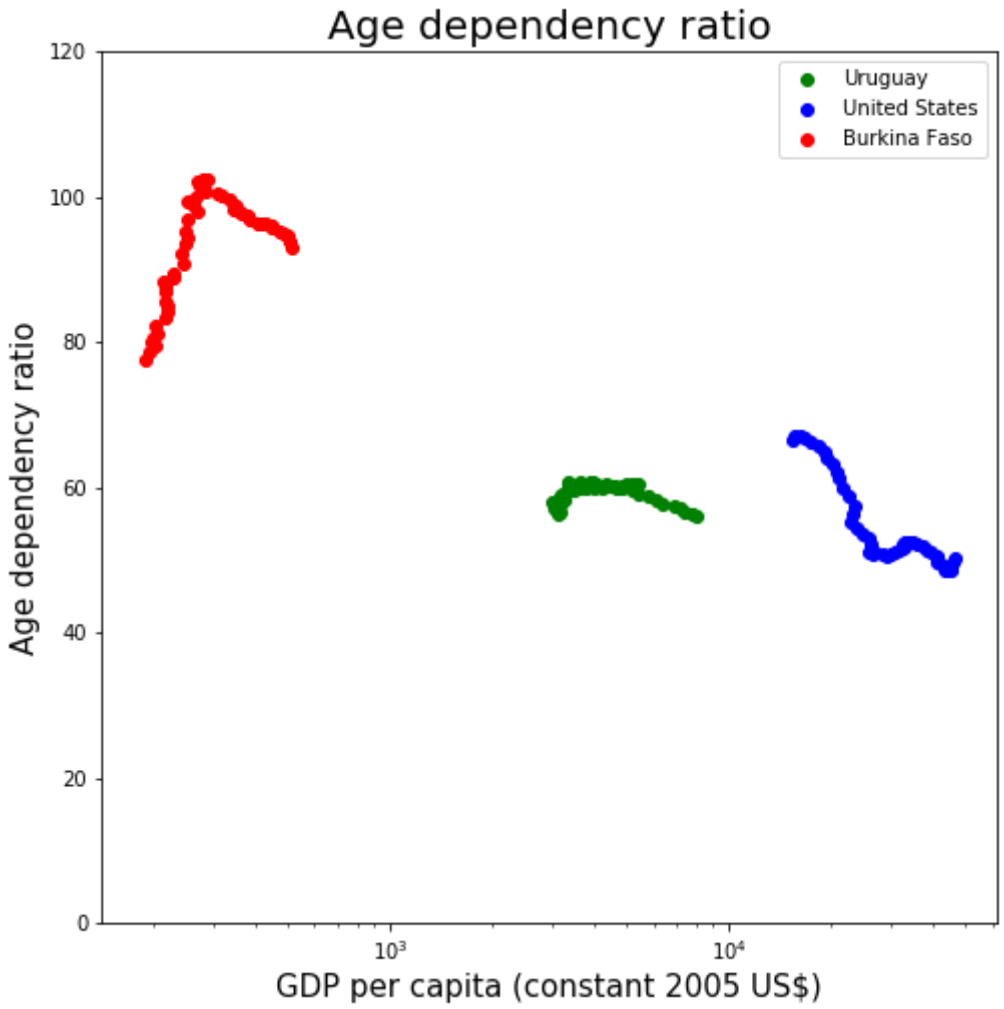

Age dependency ratio, young (% of working-age population)
Age dependency ratio, young
GDP per capita (constant 2005 US$)

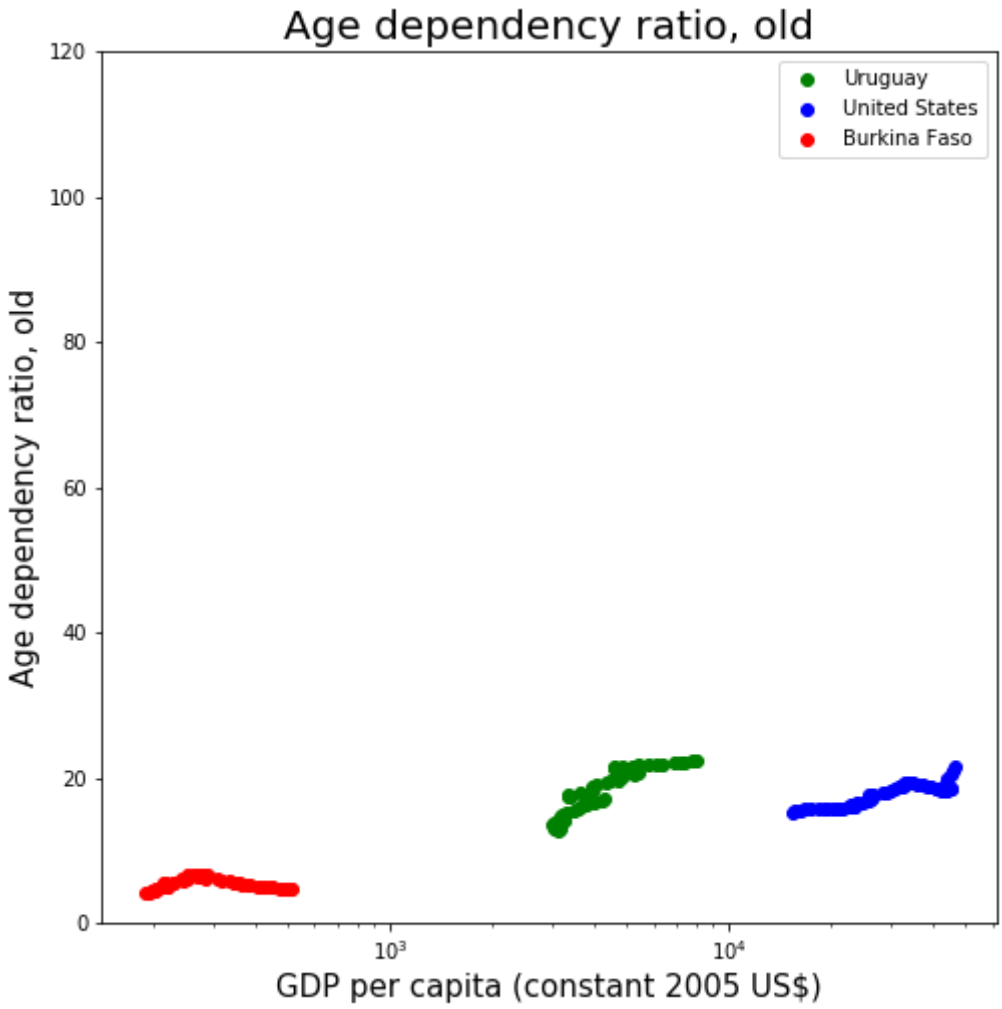
In [23]:

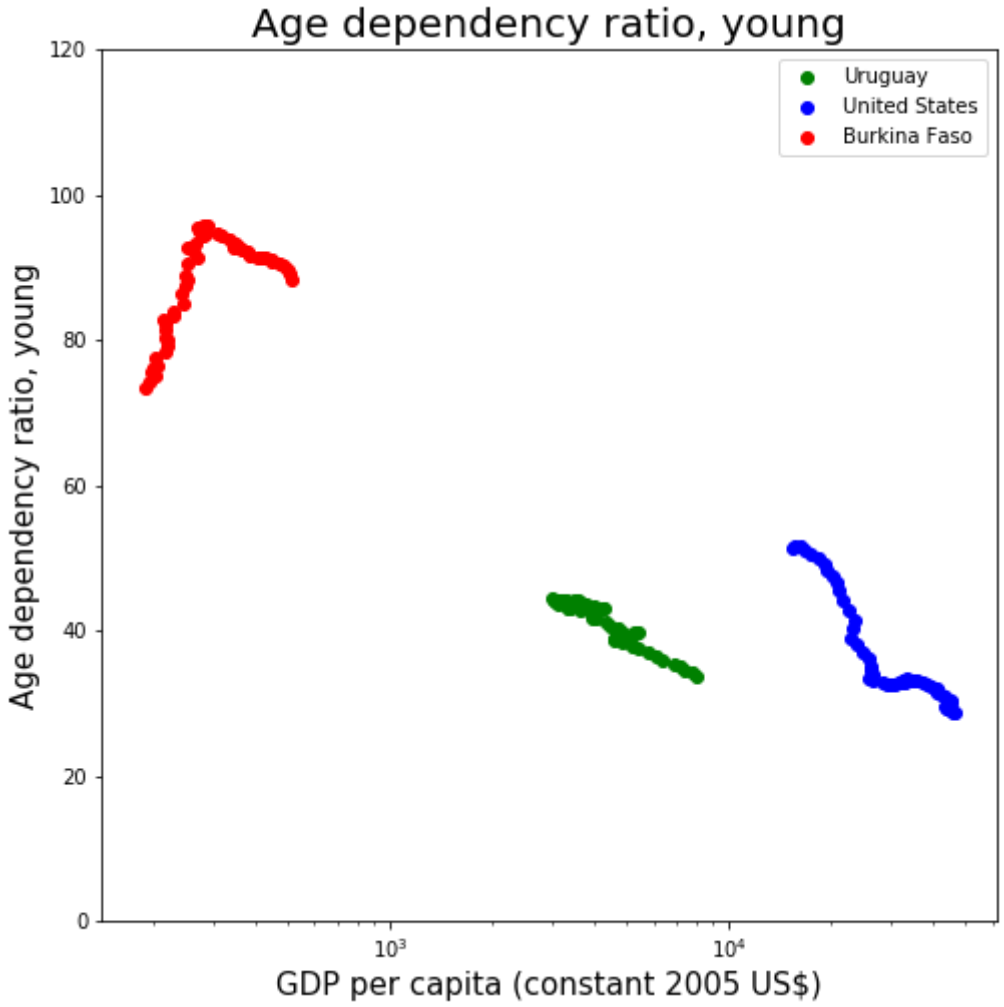
```
# pick three countries with different GDP
countries = ['Uruguay', 'United States', 'Burkina Faso']
colours = ['green', 'blue', 'red']

# plot the graphs with different colours for each country
for i in y_variable:
    fig, axis = plt.subplots(figsize=(8,8))
    axis.set_title(get_label(i).split(' (')[0],fontsize=20)
    axis.set_xlabel(get_label(x_variable),fontsize=15)
    axis.set_ylabel(get_label(i).split(' (')[0],fontsize=15)
    plt.xscale('log')

    for country, colour in zip(countries, colours):
        axis.scatter(gdp[gdp.CountryName==country][x_variable], gdp[gdp.CountryName==co
untry][i], c=colour)
    plt.gca().legend(countries)
    plt.ylim(0,120)
    fig.savefig("country_{}.svg".format(i))
```


Age dependency ratio
Uruguay
United States
Burkina Faso
120
100
80
60
40
20
0
Age dependency ratio
10^3
10^4
GDP per capita (constant 2005 US$)


Age dependency ratio, old
Uruguay
United States
Burkina Faso
120
100
80
60
40
20
0
Age dependency ratio, old
10^3
10^4
GDP per capita (constant 2005 US$)


Age dependency ratio, young
Uruguay
United States
Burkina Faso
120
100
80
60
40
20
0
Age dependency ratio, young
10^3
10^4
GDP per capita (constant 2005 US$)