Assignment 1 - Introduction to Data Science and Python

November 1, 2024

Stefan Dimitrov Velev, 0MI3400521, Big Data Technologies

Faculty of Mathematics and Informatics, Sofia University

1. Import required Python packages

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

2. Read the CSV file - Life expectancy vs. GDP per capita, 2022 *Data source: UN, World Population Prospects (2024); World Bank (2023)*

https://ourworldindata.org/grapher/life-expectancy-un-vs-gdp-per-capita-wb

```
[2]: df = pd.read_csv('./data/life-expectancy-un-vs-gdp-per-capita-wb.csv',␣
      ↪delimiter=',')
```

```
[3]: df.head()
```

```
[3]:        Entity      Code   Year  \
     0     Abkhazia  OWID_ABK   2015
     1  Afghanistan       AFG -10000
     2  Afghanistan       AFG  -9000
     3  Afghanistan       AFG  -8000
     4  Afghanistan       AFG  -7000

        Life expectancy - Sex: all - Age: 0 - Variant: estimates  \
     0                                                NaN
     1                                                NaN
     2                                                NaN
     3                                                NaN
     4                                                NaN

        GDP per capita, PPP (constant 2017 international $)  \
     0                                                NaN
     1                                                NaN
     2                                                NaN
     3                                                NaN
     4                                                NaN
```

```
   Population (historical) Continent
0                      NaN      Asia
1                  14737.0       NaN
2                  20405.0       NaN
3                  28253.0       NaN
4                  39120.0       NaN
```

[4]:
```
print("The number of rows in the data frame is:", len(df))
```

```
The number of rows in the data frame is: 59858
```

3. Data Cleaning

[5]:
```
# Remove the unnecessary columns in the data frame
df = df[['Entity', 'Year', 'Life expectancy - Sex: all - Age: 0 - Variant:␣
 ↪estimates', 'GDP per capita, PPP (constant 2017 international $)',␣
 ↪'Population (historical)']]
```

[6]:
```
# Rename the applicable columns
df = df.rename(columns={'Entity': 'Country', 'Life expectancy - Sex: all - Age:␣
 ↪0 - Variant: estimates': 'Life expectancy', 'GDP per capita, PPP (constant␣
 ↪2017 international $)': 'GDP per capita', 'Population (historical)':␣
 ↪'Population'})
```

[7]:
```
# Leaving only rows for year 2022
df = df[df['Year'] == 2022]
```

[8]:
```
# Remove rows with missing values
df = df.dropna()
```

[9]:
```
# Remove not-country-specific entries
df = df[df['Country'] != 'High-income countries']
df = df[df['Country'] != 'Low-income countries']
df = df[df['Country'] != 'Lower-middle-income countries']
df = df[df['Country'] != 'Upper-middle-income countries']
df = df[df['Country'] != 'World']
```

[10]:
```
df[df['Country'] == 'Central African Republic']
```

[10]:
```
                        Country  Year  Life expectancy  GDP per capita  \
10140  Central African Republic  2022           18.818        823.9822

       Population
10140   5098038.0
```

[11]:
```
# Correct the Central African Republic life expectancy according the World Bank␣
 ↪Report for 2022
# Source: https://data.worldbank.org/indicator/SP.DYN.LE00.IN
```

```
df.loc[df['Country'] == 'Central African Republic', 'Life expectancy'] = 54.48
```

[12]:

```
df.count()
```

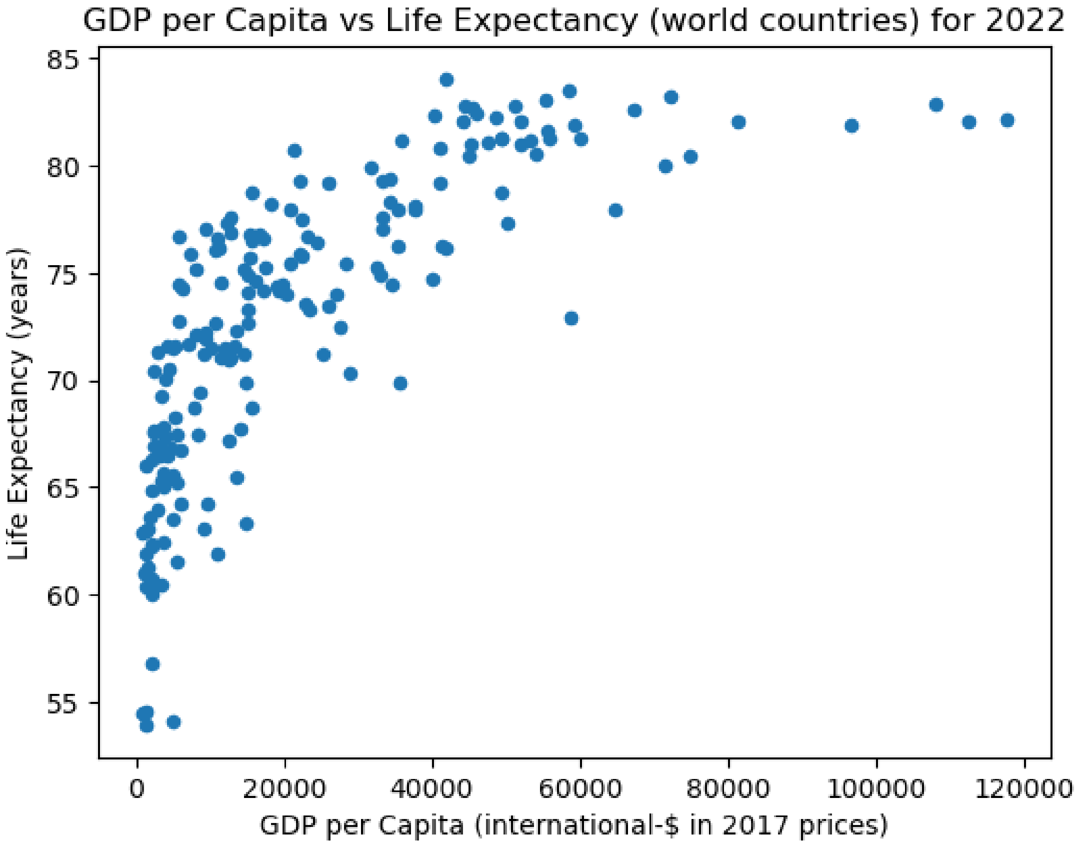
[12]:

```
Country            188
Year               188
Life expectancy    188
GDP per capita     188
Population         188
dtype: int64
```

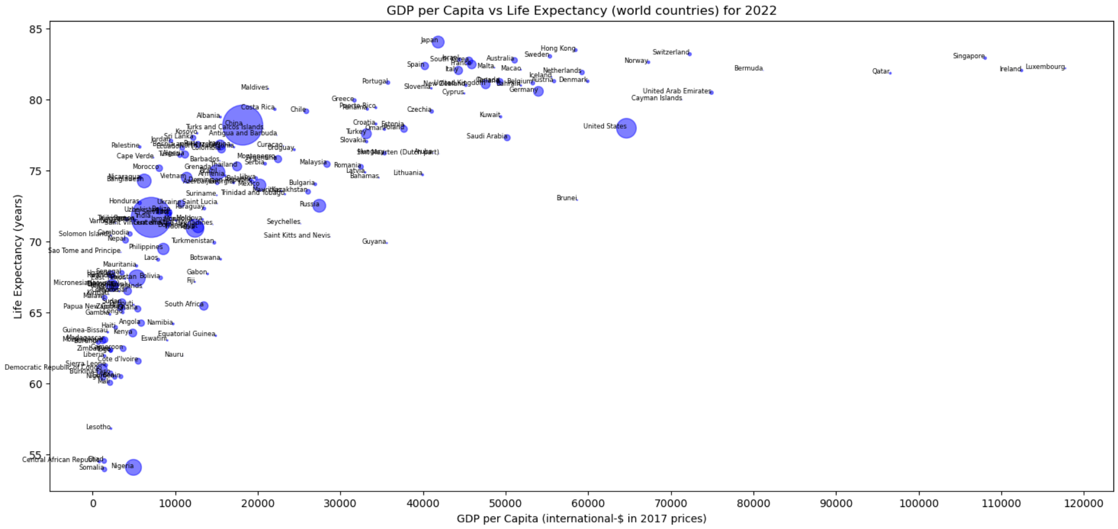
4. Draw a scatter plot of GDP per capita vs Life expectancy for 2022

[13]:

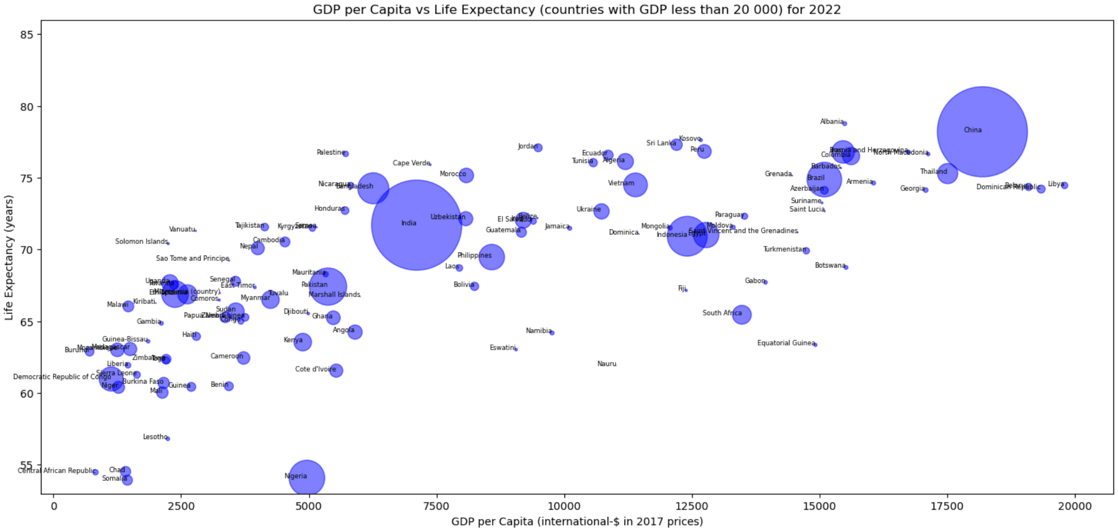
```
df.plot.scatter(x='GDP per capita', y='Life expectancy')
plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (world countries) for 2022')
plt.show()
```



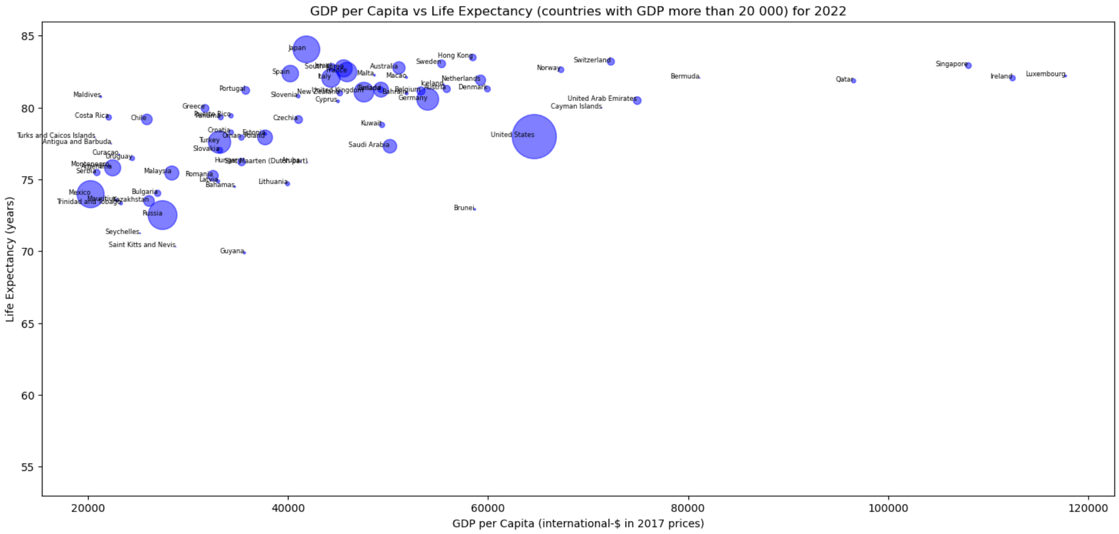
[14]:
```
plt.figure(figsize=(18, 8))
plt.scatter(df['GDP per capita'], df['Life expectancy'], color='blue', s =␣
 ↪df['Population']/1000000, alpha=0.5)

for i, country in enumerate(df['Country']):
    plt.text(df['GDP per capita'].iloc[i], df['Life expectancy'].iloc[i],␣
 ↪df['Country'].iloc[i], fontsize=6, ha='right')

plt.xticks([0, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000,␣
 ↪100000, 110000, 120000])
plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (world countries) for 2022')
plt.show()
```



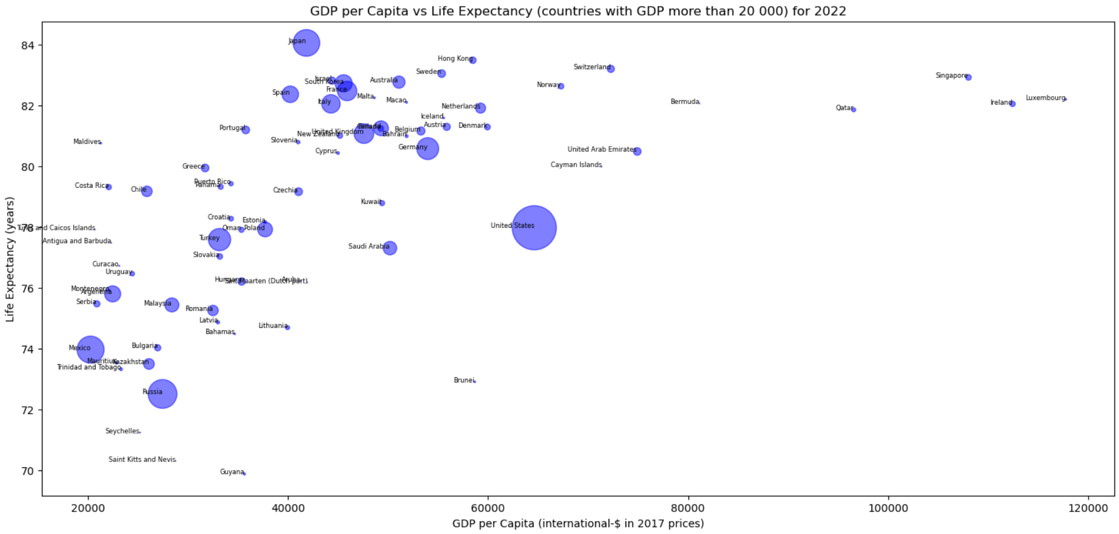
[15]:
```
plt.figure(figsize=(18, 8))
df_lower_GDP = df[df['GDP per capita'] <= 20000]
plt.scatter(df_lower_GDP['GDP per capita'], df_lower_GDP['Life expectancy'],␣
 ↪color='blue', s = df_lower_GDP['Population']/200000, alpha=0.5)

for i, country in enumerate(df_lower_GDP['Country']):
    plt.text(df_lower_GDP['GDP per capita'].iloc[i], df_lower_GDP['Life␣
 ↪expectancy'].iloc[i], df_lower_GDP['Country'].iloc[i], fontsize=6,␣
 ↪ha='right')

plt.ylim(53, 86)

plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (countries with GDP less than 20␣
  ↪000) for 2022')
plt.show()
```



[16]:
```
plt.figure(figsize=(18, 8))
df_higher_GDP = df[df['GDP per capita'] >= 20000]
plt.scatter(df_higher_GDP['GDP per capita'], df_higher_GDP['Life expectancy'],␣
  ↪color='blue', s = df_higher_GDP['Population']/200000, alpha=0.5)

for i, country in enumerate(df_higher_GDP['Country']):
    plt.text(df_higher_GDP['GDP per capita'].iloc[i], df_higher_GDP['Life␣
  ↪expectancy'].iloc[i], df_higher_GDP['Country'].iloc[i], fontsize=6,␣
  ↪ha='right')

plt.ylim(53, 86)

plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (countries with GDP more than 20␣
  ↪000) for 2022')
plt.show()
```



[17]:
```python
plt.figure(figsize=(18, 8))
plt.scatter(df_higher_GDP['GDP per capita'], df_higher_GDP['Life expectancy'],␣
 ↪color='blue', s = df_higher_GDP['Population']/200000, alpha=0.5)

for i, country in enumerate(df_higher_GDP['Country']):
    plt.text(df_higher_GDP['GDP per capita'].iloc[i], df_higher_GDP['Life␣
 ↪expectancy'].iloc[i], df_higher_GDP['Country'].iloc[i], fontsize=6,␣
 ↪ha='right')

plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (countries with GDP more than 20␣
 ↪000) for 2022')
plt.show()
```



5. Find out which countries have a life expectancy higher than one standard deviation above the mean?

[18]:
```
# Get statistic for the whole DataFrame with Pandas
df.describe()
```

[18]:
```
         Year  Life expectancy  GDP per capita    Population
count   188.0       188.000000      188.000000  1.880000e+02
mean   2022.0        72.752750    22643.151954  4.151910e+07
std       0.0         7.039319    22790.551663  1.522105e+08
min    2022.0        53.931000      708.178300  1.001200e+04
25%    2022.0        67.404750     5134.252600  1.831064e+06
50%    2022.0        74.079500    14829.153500  7.917908e+06
75%    2022.0        77.937000    34831.845750  3.040557e+07
max    2022.0        84.054000   117746.990000  1.425423e+09
```

[19]:
```
# Get the average life expectancy and the standard deviation with NumPy
mean_life_expectancy = np.mean(df['Life expectancy'])
standard_deviation_life_expectancy = np.std(df['Life expectancy'])
print('Mean life expectancy is', mean_life_expectancy)
print('Standard deviation of life expectancy is',␣
  ↪standard_deviation_life_expectancy)
```

```
Mean life expectancy is 72.75275
Standard deviation of life expectancy is 7.020572922933241
```

[20]:
```
# Find lower boundary for the searched countries - one standard deviation above␣
  ↪the mean
lower_boundary_high_life_expectancy = mean_life_expectancy +␣
  ↪standard_deviation_life_expectancy
```

```
print('The searched lower boundary for high life expectancy is',␣
 ↪lower_boundary_high_life_expectancy)
```

```
The searched lower boundary for high life expectancy is 79.77332292293325
```

[21]:
```
# List countries with higher life expectancy (one standard deviation above the␣
 ↪mean)
df_higher_life_expectancy = df[df['Life expectancy'] >␣
 ↪lower_boundary_high_life_expectancy]
print('The number of countries with higher life expectancy is',␣
 ↪len(df_higher_life_expectancy))
print('The list of countries with higher life expectancy is',  ', '.
 ↪join(df_higher_life_expectancy['Country']))
```

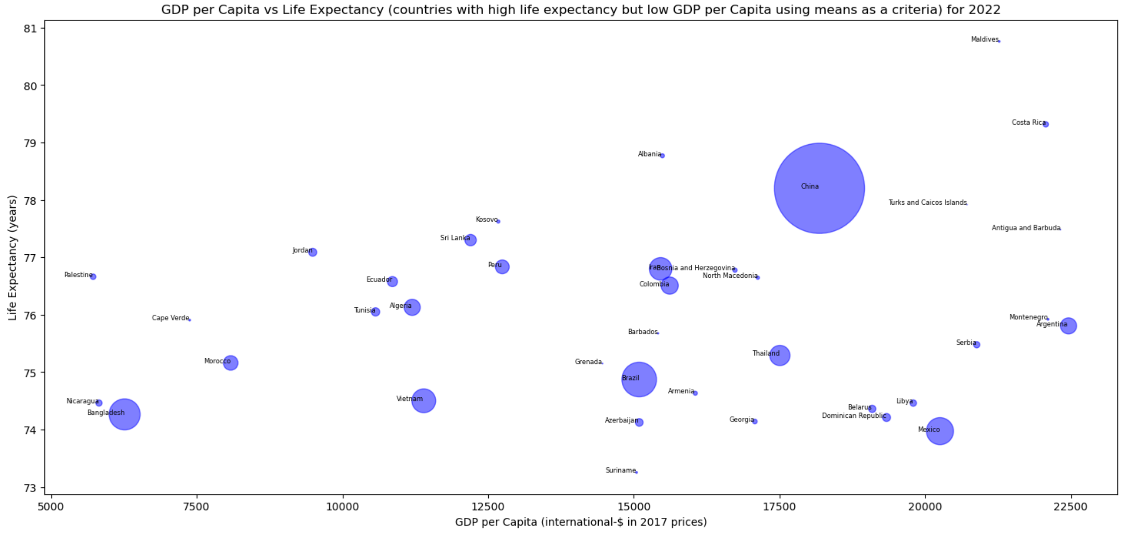
```
The number of countries with higher life expectancy is 36
The list of countries with higher life expectancy is Australia, Austria,
Bahrain, Belgium, Bermuda, Canada, Cayman Islands, Cyprus, Denmark, Finland,
France, Germany, Greece, Hong Kong, Iceland, Ireland, Israel, Italy, Japan,
Luxembourg, Macao, Maldives, Malta, Netherlands, New Zealand, Norway, Portugal,
Qatar, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Arab
Emirates, United Kingdom
```

6. Check which countries have high life expectancy but have low GDP?

[22]:
```
# List countries with high life expectancy but low GDP using means
mean_gdp = np.mean(df['GDP per capita'])
print('Mean GDP per capita is', mean_gdp)
df_higher_life_expectancy_lower_GDP1 = df[(df['Life expectancy'] >␣
 ↪mean_life_expectancy) & (df['GDP per capita'] < mean_gdp)]
print('The number of countries with high life expectancy but low GDP per capita␣
 ↪using means is', len(df_higher_life_expectancy_lower_GDP1))
print('The list of countries with high life expectancy but low GDP per capita␣
 ↪using means is:',  ', '.
 ↪join(df_higher_life_expectancy_lower_GDP1['Country']))
```

```
Mean GDP per capita is 22643.151954255318
The number of countries with high life expectancy but low GDP per capita using
means is 38
The list of countries with high life expectancy but low GDP per capita using
means is: Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan,
Bangladesh, Barbados, Belarus, Bosnia and Herzegovina, Brazil, Cape Verde,
China, Colombia, Costa Rica, Dominican Republic, Ecuador, Georgia, Grenada,
Iran, Jordan, Kosovo, Libya, Maldives, Mexico, Montenegro, Morocco, Nicaragua,
North Macedonia, Palestine, Peru, Serbia, Sri Lanka, Suriname, Thailand,
Tunisia, Turks and Caicos Islands, Vietnam
```

[23]:
```
plt.figure(figsize=(18, 8))
plt.scatter(df_higher_life_expectancy_lower_GDP1['GDP per capita'],␣
 ↪df_higher_life_expectancy_lower_GDP1['Life expectancy'], color='blue', s =␣
 ↪df_higher_life_expectancy_lower_GDP1['Population']/200000, alpha=0.5)

for i, country in enumerate(df_higher_life_expectancy_lower_GDP1['Country']):
    plt.text(df_higher_life_expectancy_lower_GDP1['GDP per capita'].iloc[i],␣
 ↪df_higher_life_expectancy_lower_GDP1['Life expectancy'].iloc[i],␣
 ↪df_higher_life_expectancy_lower_GDP1['Country'].iloc[i], fontsize=6,␣
 ↪ha='right')

plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (countries with high life␣
 ↪expectancy but low GDP per Capita using means as a criteria) for 2022')
plt.show()
```



```
[24]: # List countries with high life expectancy but low GDP using 75th percentile␣
       ↪for high life expectancy and the 25th percentile for low GDP
      upper_boundary_low_gdp = df['GDP per capita'].quantile(0.25)
      print('The upper boundary for low GDP per capita is', upper_boundary_low_gdp)
      lower_boundary_high_life_expectancy = df['Life expectancy'].quantile(0.75)
      print('The lower boundary for high life expectancy is',␣
       ↪lower_boundary_high_life_expectancy)
      df_higher_life_expectancy_lower_GDP2 = df[(df['Life expectancy'] >=␣
       ↪lower_boundary_high_life_expectancy) & (df['GDP per capita'] <=␣
       ↪upper_boundary_low_gdp)]
      print('The number of countries with high life expectancy but low GDP per capita␣
       ↪using percentiles is', len(df_higher_life_expectancy_lower_GDP2))
```

```
The upper boundary for low GDP per capita is 5134.2526
The lower boundary for high life expectancy is 77.937
The number of countries with high life expectancy but low GDP per capita using
percentiles is 0
```

7. Find whether each strong economy (normally indicated by GDP) have high life expectancy?

[25]:
```
lower_boundary_high_gdp = df['GDP per capita'].quantile(0.75)
print('The lower boundary for high GDP per capita is', lower_boundary_high_gdp)
df_higher_gdp = df[df['GDP per capita'] >= lower_boundary_high_gdp]
print('The number of countries with high GDP per capita using 75th percentile␣
  ↪is', len(df_higher_gdp))
```

```
The lower boundary for high GDP per capita is 34831.84575
The number of countries with high GDP per capita using 75th percentile is 47
```

[26]:
```
df_higher_gdp
```

[26]:
```
                      Country  Year  Life expectancy  GDP per capita  \
2853                    Aruba  2022           76.226       41273.613
3579                Australia  2022           82.766       51090.260
3840                  Austria  2022           81.296       55867.184
4624                  Bahrain  2022           80.992       51854.715
5661                  Belgium  2022           81.159       53287.152
6273                  Bermuda  2022           82.062       81165.650
7998                   Brunei  2022           72.917       58669.902
9564                   Canada  2022           81.249       49296.380
9879           Cayman Islands  2022           79.984       71353.890
13156                  Cyprus  2022           80.434       44996.316
13417                 Czechia  2022           79.165       41052.348
14024                 Denmark  2022           81.291       59935.120
16671                 Estonia  2022           78.167       37711.820
18745                 Finland  2022           81.243       49275.152
19006                  France  2022           82.475       45904.410
20258                 Germany  2022           80.580       53969.625
22553                  Guyana  2022           69.888       35634.688
23603               Hong Kong  2022           83.485       58478.883
23864                 Hungary  2022           76.212       35356.777
24125                 Iceland  2022           81.588       55567.438
25430                 Ireland  2022           82.050      112445.420
25767                  Israel  2022           82.814       44393.300
26028                   Italy  2022           82.052       44292.190
26542                   Japan  2022           84.054       41837.910
27963                  Kuwait  2022           78.788       49400.355
30715               Lithuania  2022           74.696       39955.246
31506              Luxembourg  2022           82.201      117746.990
31592                   Macao  2022           82.103       51840.140
33118                   Malta  2022           82.250       48641.850
```

```
37552                Netherlands  2022           81.912       59249.168
38016                New Zealand  2022           81.006       45185.312
40106                     Norway  2022           82.631       67296.160
40706                       Oman  2022           77.911       35336.895
42983                     Poland  2022           77.923       37706.605
43244                   Portugal  2022           81.194       35767.723
43624                      Qatar  2022           81.857       96557.810
46611               Saudi Arabia  2022           77.310       50188.297
47945                  Singapore  2022           82.921      108036.110
48019  Sint Maarten (Dutch part)  2022           76.180       41942.918
48541                   Slovenia  2022           80.793       41015.227
50005                South Korea  2022           82.727       45560.125
50493                      Spain  2022           82.366       40223.010
51673                     Sweden  2022           83.046       55359.344
51934                Switzerland  2022           83.200       72278.210
56201       United Arab Emirates  2022           80.487       74917.670
56462             United Kingdom  2022           81.074       47587.168
56723              United States  2022           77.979       64623.125

        Population
2853      107792.0
3579    26200987.0
3840     9064679.0
4624     1533459.0
5661    11641813.0
6273       64772.0
7998      455374.0
9564    38821260.0
9879       71609.0
13156    1331376.0
13417   10673216.0
14024    5902898.0
16671    1350092.0
18745    5569299.0
19006   66277412.0
20258   84086228.0
22553     821636.0
23603    7465914.0
23864    9684306.0
24125     380368.0
25430    5110013.0
25767    9103144.0
26028   59619106.0
26542  124997586.0
27963    4589514.0
30715    2816922.0
31506     653316.0
```

```
31592      704359.0
33118      528194.0
37552    17904422.0
38016     5131733.0
40106     5456795.0
40706     4730227.0
42983    38385734.0
43244    10417075.0
43624     2892465.0
46611    32175352.0
47945     5649886.0
48019       42163.0
48541     2115230.0
50005    51782514.0
50493    47828386.0
51673    10487333.0
51934     8792180.0
56201    10242085.0
56462    68179315.0
56723   341534041.0
```

[27]:
```
print('Mean life expectancy is', mean_life_expectancy)
print('75th Percentile life expectancy is', lower_boundary_high_life_expectancy)
```

```
Mean life expectancy is 72.75275
75th Percentile life expectancy is 77.937
```

[28]:
```
df_higher_gdp[df_higher_gdp['Life expectancy'] >␣
 ↪lower_boundary_high_life_expectancy]
```

[28]:
```
                   Country  Year  Life expectancy  GDP per capita  \
3579             Australia  2022           82.766       51090.260
3840               Austria  2022           81.296       55867.184
4624               Bahrain  2022           80.992       51854.715
5661               Belgium  2022           81.159       53287.152
6273               Bermuda  2022           82.062       81165.650
9564                Canada  2022           81.249       49296.380
9879        Cayman Islands  2022           79.984       71353.890
13156               Cyprus  2022           80.434       44996.316
13417              Czechia  2022           79.165       41052.348
14024              Denmark  2022           81.291       59935.120
16671              Estonia  2022           78.167       37711.820
18745              Finland  2022           81.243       49275.152
19006               France  2022           82.475       45904.410
20258              Germany  2022           80.580       53969.625
23603            Hong Kong  2022           83.485       58478.883
24125              Iceland  2022           81.588       55567.438
25430              Ireland  2022           82.050      112445.420
```

```
25767                Israel  2022           82.814        44393.300
26028                 Italy  2022           82.052        44292.190
26542                 Japan  2022           84.054        41837.910
27963                Kuwait  2022           78.788        49400.355
31506            Luxembourg  2022           82.201       117746.990
31592                 Macao  2022           82.103        51840.140
33118                 Malta  2022           82.250        48641.850
37552           Netherlands  2022           81.912        59249.168
38016           New Zealand  2022           81.006        45185.312
40106                Norway  2022           82.631        67296.160
43244              Portugal  2022           81.194        35767.723
43624                 Qatar  2022           81.857        96557.810
47945             Singapore  2022           82.921       108036.110
48541              Slovenia  2022           80.793        41015.227
50005           South Korea  2022           82.727        45560.125
50493                 Spain  2022           82.366        40223.010
51673                Sweden  2022           83.046        55359.344
51934           Switzerland  2022           83.200        72278.210
56201  United Arab Emirates  2022           80.487        74917.670
56462        United Kingdom  2022           81.074        47587.168
56723         United States  2022           77.979        64623.125

        Population
3579    26200987.0
3840     9064679.0
4624     1533459.0
5661    11641813.0
6273       64772.0
9564    38821260.0
9879       71609.0
13156    1331376.0
13417   10673216.0
14024    5902898.0
16671    1350092.0
18745    5569299.0
19006   66277412.0
20258   84086228.0
23603    7465914.0
24125     380368.0
25430    5110013.0
25767    9103144.0
26028   59619106.0
26542  124997586.0
27963    4589514.0
31506     653316.0
31592     704359.0
33118     528194.0
```

```
37552   17904422.0
38016    5131733.0
40106    5456795.0
43244   10417075.0
43624    2892465.0
47945    5649886.0
48541    2115230.0
50005   51782514.0
50493   47828386.0
51673   10487333.0
51934    8792180.0
56201   10242085.0
56462   68179315.0
56723  341534041.0
```

```
[29]: df_higher_gdp[(df_higher_gdp['Life expectancy'] >= mean_life_expectancy) &␣
       ↪(df_higher_gdp['Life expectancy'] <= lower_boundary_high_life_expectancy)]
```

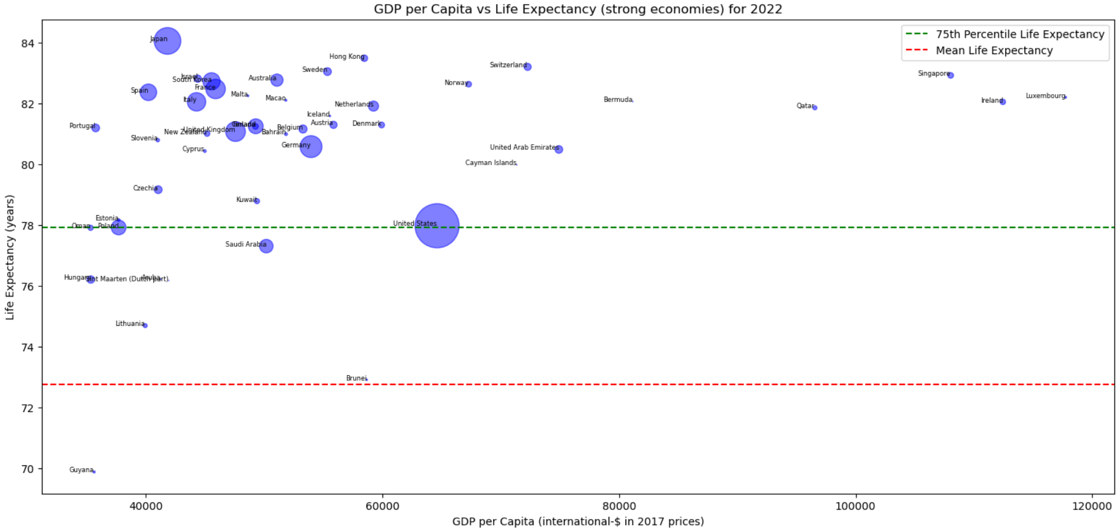
```
[29]:                       Country  Year  Life expectancy  GDP per capita  \
      2853                    Aruba  2022           76.226       41273.613
      7998                   Brunei  2022           72.917       58669.902
      23864                 Hungary  2022           76.212       35356.777
      30715               Lithuania  2022           74.696       39955.246
      40706                    Oman  2022           77.911       35336.895
      42983                  Poland  2022           77.923       37706.605
      46611            Saudi Arabia  2022           77.310       50188.297
      48019  Sint Maarten (Dutch part)  2022           76.180       41942.918

              Population
      2853      107792.0
      7998      455374.0
      23864    9684306.0
      30715    2816922.0
      40706    4730227.0
      42983   38385734.0
      46611   32175352.0
      48019      42163.0
```

```
[30]: df_higher_gdp[df_higher_gdp['Life expectancy'] < mean_life_expectancy]
```

```
[30]:       Country  Year  Life expectancy  GDP per capita  Population
      22553  Guyana  2022           69.888       35634.688    821636.0
```

```
[31]: plt.figure(figsize=(18, 8))
      plt.scatter(df_higher_gdp['GDP per capita'], df_higher_gdp['Life expectancy'],␣
       ↪color='blue', s = df_higher_gdp['Population']/200000, alpha=0.5)
```

```
for i, country in enumerate(df_higher_gdp['Country']):
    plt.text(df_higher_gdp['GDP per capita'].iloc[i], df_higher_gdp['Life␣
 ↪expectancy'].iloc[i], df_higher_gdp['Country'].iloc[i], fontsize=6,␣
 ↪ha='right')

plt.axhline(y=lower_boundary_high_life_expectancy, color='green',␣
 ↪linestyle='--', label='75th Percentile Life Expectancy')

plt.axhline(y=mean_life_expectancy, color='red', linestyle='--', label='Mean␣
 ↪Life Expectancy')

plt.xlabel('GDP per Capita (international-$ in 2017 prices)')
plt.ylabel('Life Expectancy (years)')
plt.title('GDP per Capita vs Life Expectancy (strong economies) for 2022')
plt.legend()
plt.show()
```


GDP per Capita vs Life Expectancy (strong economies) for 2022
75th Percentile Life Expectancy
Mean Life Expectancy
Life Expectancy (years)
GDP per Capita (international-$ in 2017 prices)
Japan
Hong Kong
Switzerland
Sweden
Israel
South Korea
Australia
Norway
Singapore
Spain
France
Malta
Italy
Macao
Netherlands
Bermuda
Qatar
Ireland
Luxembourg
Iceland
Austria
Denmark
Portugal
Canada
Finland
New Zealand
United Kingdom
Belgium
Bahrain
Slovenia
Germany
Cyprus
United Arab Emirates
Cayman Islands
Czechia
Kuwait
Estonia
Oman
Poland
United States
Saudi Arabia
Hungary
Sint Maarten (Dutch part)
Aruba
Lithuania
Brunei
Guyana