EDA on Gapminder Dataset

Importing Necessary Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
```

Dataset

```
gapminder_world = pd.read_csv('Downloads/gapminder_full.csv')
gapminder_world.head()
```

	country	year	population	continent	life_exp	gdp_cap
0	Afghanistan	1952	8425333	Asia	28.801	779.445314
1	Afghanistan	1957	9240934	Asia	30.332	820.853030
2	Afghanistan	1962	10267083	Asia	31.997	853.100710
3	Afghanistan	1967	11537966	Asia	34.020	836.197138
4	Afghanistan	1972	13079460	Asia	36.088	739.981106

Information about Dataset

```
gapminder_world.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1704 entries, 0 to 1703
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   country     1704 non-null   object 
 1   year        1704 non-null   int64  
 2   population  1704 non-null   int64  
 3   continent   1704 non-null   object 
 4   life_exp    1704 non-null   float64
 5   gdp_cap     1704 non-null   float64
dtypes: float64(2), int64(2), object(2)
memory usage: 80.0+ KB
```

Descriptive Statistics

```
gapminder_world.describe()
```

	year	population	life_exp	gdp_cap
count	1704.00000	1.704000e+03	1704.000000	1704.000000
mean	1979.50000	2.960121e+07	59.474439	7215.327081
std	17.26533	1.061579e+08	12.917107	9857.454543
min	1952.00000	6.001100e+04	23.599000	241.165876
25%	1965.75000	2.793664e+06	48.198000	1202.060309
50%	1979.50000	7.023596e+06	60.712500	3531.846988
75%	1993.25000	1.958522e+07	70.845500	9325.462346
max	2007.00000	1.318683e+09	82.603000	113523.132900

Total Countries

```
num_countries = gapminder_world['country'].nunique()
print(f'Total number of countries: {num_countries}')
```

```
Total number of countries: 142
```

```
gapminder_world.isnull().sum().sum()
```

```
0
```

Pivot table that shows the average life expectancy for each continent and year.

```
average_life_expectancy_for_continent = gapminder_world.pivot_table(index = 'continent', columns = 'year', values = 'life_exp', aggfunc
average_life_expectancy_for_continent
```

year	1952	1957	1962	1967	1972	1977	1982	1987	1992	1997	2002	2
continent												
Africa	39.135500	41.266346	43.319442	45.334538	47.450942	49.580423	51.592865	53.344788	53.629577	53.598269	53.325231	54.80
Americas	53.279840	55.960280	58.398760	60.410920	62.394920	64.391560	66.228840	68.090720	69.568360	71.150480	72.422040	73.60
Asia	46.314394	49.318544	51.563223	54.663640	57.319269	59.610556	62.617939	64.851182	66.537212	68.020515	69.233879	70.72
Europe	64.408500	66.703067	68.539233	69.737600	70.775033	71.937767	72.806400	73.642167	74.440100	75.505167	76.700600	77.64
Oceania	69.255000	70.295000	71.085000	71.310000	71.910000	72.855000	74.290000	75.320000	76.945000	78.190000	79.740000	80.71

Countries with a GDP per capita higher than the 75th percentile in 2007

```
gapminder_world_2007 = gapminder_world[gapminder_world['year'] == 2007]
gdp_cap_75_percentile = gapminder_world_2007['gdp_cap'].quantile(0.75)
high_gdp_countries = gapminder_world_2007[gapminder_world_2007['gdp_cap'] > gdp_cap_75_percentile]['country']
high_gdp_countries.tolist()
```

```
['Australia',
 'Austria',
 'Bahrain',
 'Belgium',
 'Canada',
 'Czech Republic',
 'Denmark',
 'Finland',
 'France',
 'Germany',
 'Greece',
 'Hong Kong, China',
 'Hungary',
 'Iceland',
 'Ireland',
 'Israel',
 'Italy',
 'Japan',
 'Korea, Rep.',
 'Kuwait',
 'Netherlands',
 'New Zealand',
 'Norway',
 'Oman',
 'Portugal',
 'Puerto Rico',
 'Saudi Arabia',
 'Singapore',
 'Slovak Republic',
 'Slovenia',
 'Spain',
 'Sweden',
 'Switzerland',
 'Taiwan',
 'United Kingdom',
 'United States']
```

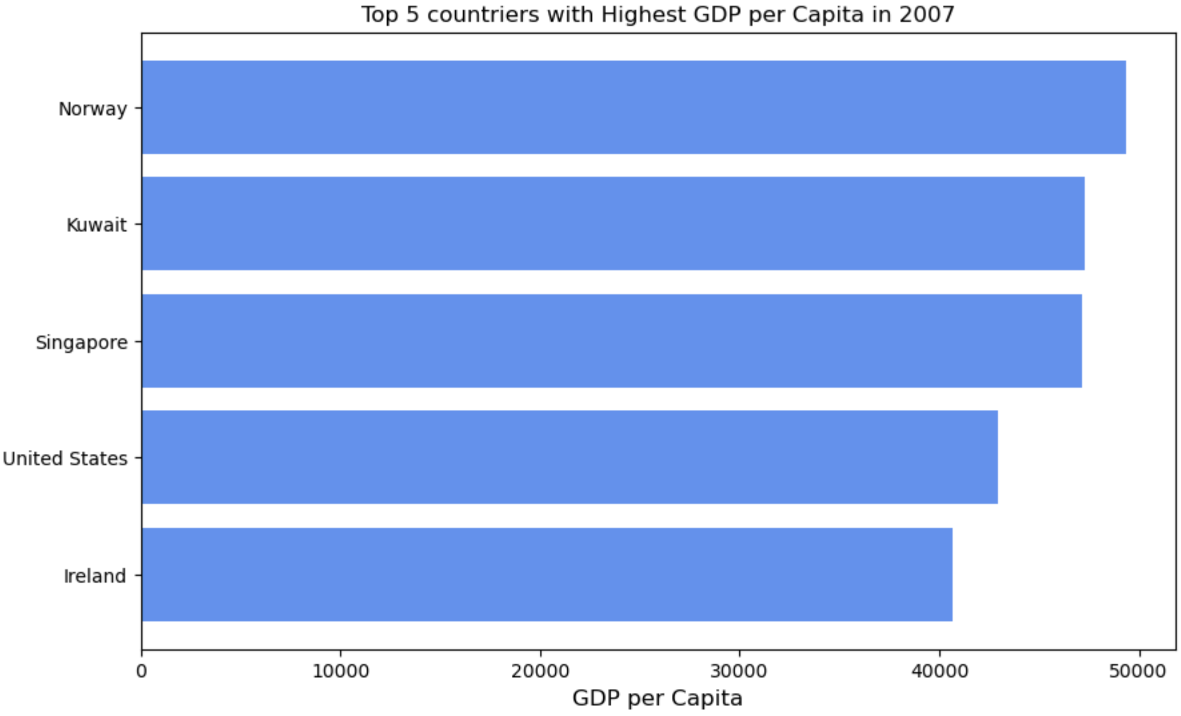
Life Expectancy category from Low to Very High

```
gapminder_world['Life_Exp_Range'] = pd.cut(gapminder_world['life_exp'], bins = 4, labels = ['Low', 'Mediun', 'High', 'Very High'])
gapminder_world.head()
```

	country	year	population	continent	life_exp	gdp_cap	Life_Exp_Range
0	Afghanistan	1952	8425333	Asia	28.801	779.445314	Low
1	Afghanistan	1957	9240934	Asia	30.332	820.853030	Low
2	Afghanistan	1962	10267083	Asia	31.997	853.100710	Low
3	Afghanistan	1967	11537966	Asia	34.020	836.197138	Low
4	Afghanistan	1972	13079460	Asia	36.088	739.981106	Low

Top 5 countries with the highest GDP per capita in 2007.

```
top_5_gdp_countries = gapminder_world_2007.nlargest(5, 'gdp_cap')

x = top_5_gdp_countries['country']
y = top_5_gdp_countries['gdp_cap']

plt.figure(figsize = (10, 6))
plt.barh(x, y, color = 'cornflowerblue')
plt.xlabel('GDP per Capita', fontsize = 12)
plt.title('Top 5 countriers with Highest GDP per Capita in 2007')
plt.gca().invert_yaxis()
plt.show()
```



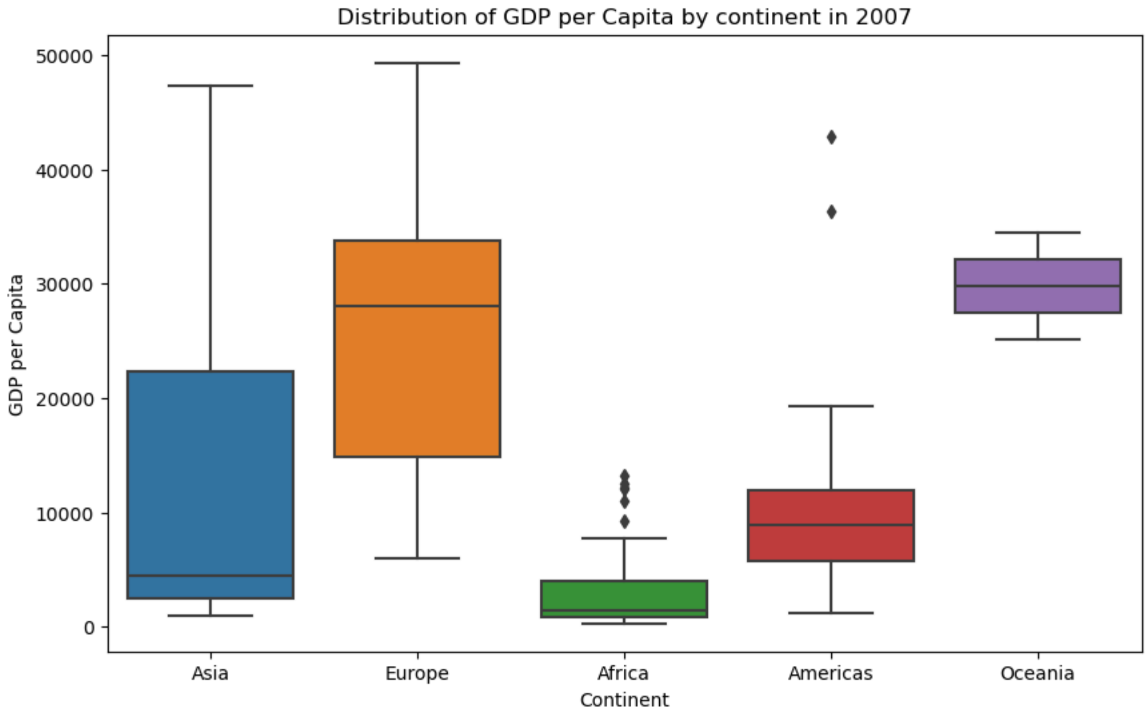
Country names that start with "I" and end with "a" using regex.

```
regex = r'^I.*a$'

countries_with_Ia = gapminder_world[gapminder_world['country'].str.contains(regex, regex = True)]['country'].unique()

print(countries_with_Ia)
```

```
['India' 'Indonesia']
```

Box plot showing the distribution of the GDP per capita for each continent in 2007.

```
plt.figure(figsize = (10, 6))
sns.boxplot(x = 'continent', y = 'gdp_cap', data = gapminder_world_2007)
plt.title('Distribution of GDP per Capita by continent in 2007')
plt.xlabel('Continent')
plt.ylabel('GDP per Capita')
plt.show()
```



Countries with a life expectancy of over 80 years in 2007 with their respective continents.

```
high_life_exp_countries = gapminder_world_2007[gapminder_world_2007['life_exp'] > 80]
high_life_exp_countries.loc[:, ['country', 'continent']]
```

	country	continent
71	Australia	Oceania
251	Canada	Americas
539	France	Europe
671	Hong Kong, China	Asia
695	Iceland	Europe
767	Israel	Asia
779	Italy	Europe
803	Japan	Asia
1103	New Zealand	Oceania
1151	Norway	Europe
1427	Spain	Europe
1475	Sweden	Europe
1487	Switzerland	Europe

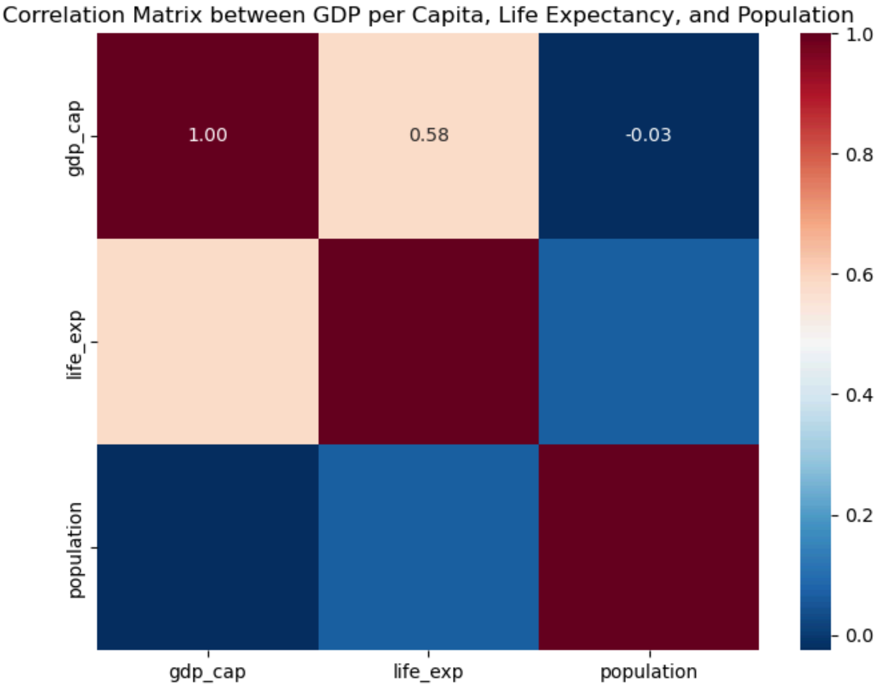
Converted the 'year' column to a datetime type and extracted the decade. Created a new column 'Decade' that groups the years into decades (e.g., the 1950s, 1960s).

```
gapminder_world['year'] = pd.to_datetime(gapminder_world['year'], format = '%Y')

gapminder_world['Decade'] = (gapminder_world['year'].dt.year // 10) * 10
gapminder_world['Decade'] = gapminder_world['Decade'].astype(str) + 's'
gapminder_world.head()
```

	country	year	population	continent	life_exp	gdp_cap	Life_Exp_Range	Decade
0	Afghanistan	1952-01-01	8425333	Asia	28.801	779.445314	Low	1950s
1	Afghanistan	1957-01-01	9240934	Asia	30.332	820.853030	Low	1950s
2	Afghanistan	1962-01-01	10267083	Asia	31.997	853.100710	Low	1960s
3	Afghanistan	1967-01-01	11537966	Asia	34.020	836.197138	Low	1960s
4	Afghanistan	1972-01-01	13079460	Asia	36.088	739.981106	Low	1970s

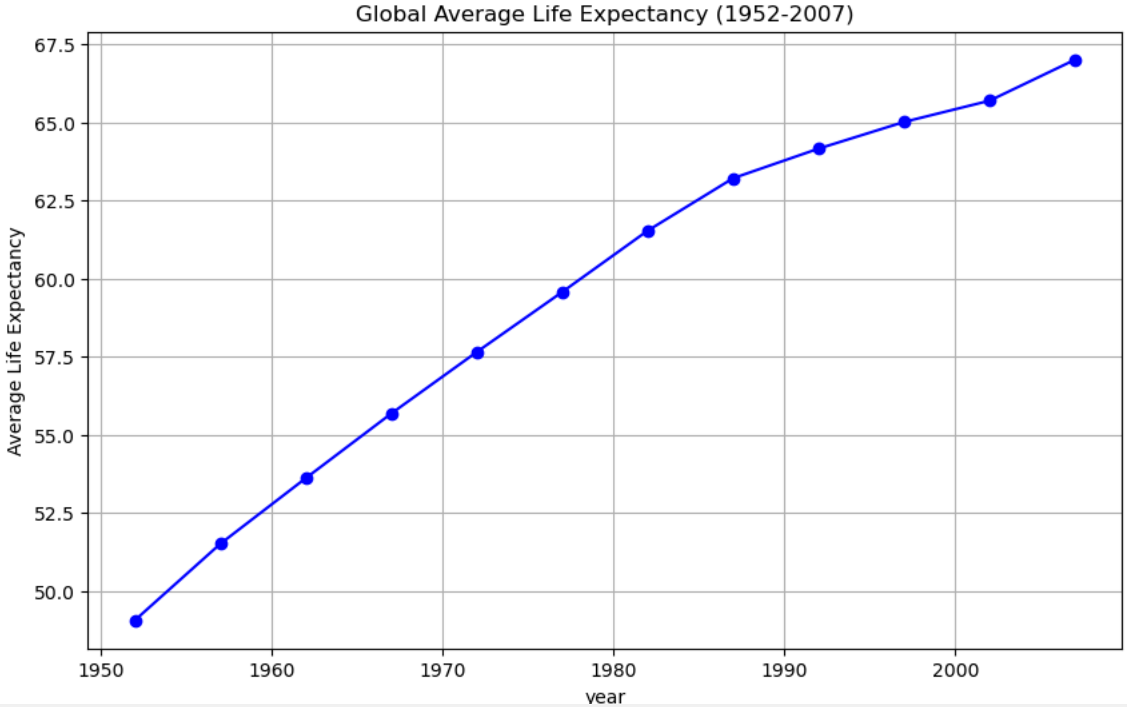
Heat map showing the correlation matrix between GDP per capita, life expectancy, and population.

```
correlation_data = gapminder_world.loc[:,['gdp_cap', 'life_exp', 'population']]

correlation_matrix = correlation_data.corr()

plt.figure(figsize = (8, 6))
sns.heatmap(data = correlation_matrix, annot = True, cmap = 'RdBu_r', fmt = '.2f')
plt.title('Correlation Matrix between GDP per Capita, Life Expectancy, and Population')
plt.show()
```



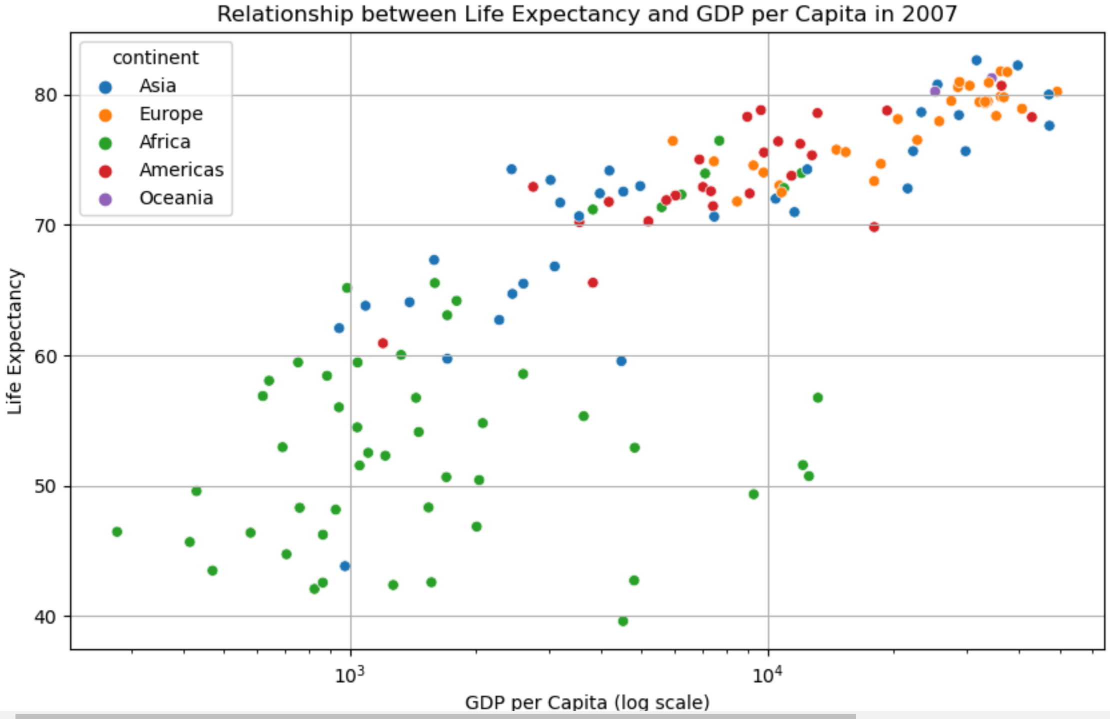
Line graph showing how the global average life expectancy changed from 1952 to 2007.

```
global_avg_life_exp = gapminder_world.groupby('year')['life_exp'].mean()
x = global_avg_life_exp.index
y = global_avg_life_exp.values

plt.figure(figsize = (10, 6))
plt.plot(x, y, marker = 'o', linestyle = '-', color = 'b')
plt.title('Global Average Life Expectancy (1952-2007)')
plt.xlabel('year')
plt.ylabel('Average Life Expectancy')
plt.grid(True)
plt.show()
```



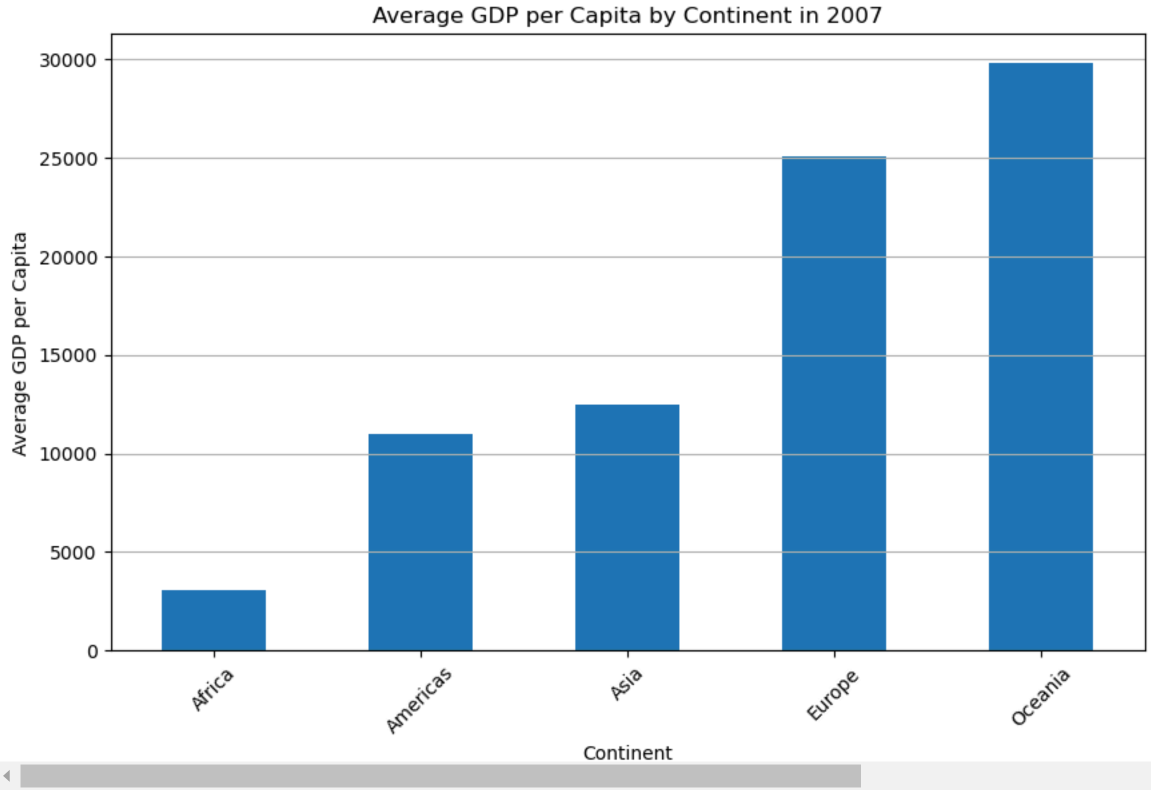
The relationship between life expectancy and GDP per capita for the year 2007.

```
plt.figure(figsize = (10, 6))
sns.scatterplot(x = 'gdp_cap', y = 'life_exp', data = gapminder_world_2007, hue = 'continent')
plt.title('Relationship between Life Expectancy and GDP per Capita in 2007')
plt.xscale('log')
plt.xlabel('GDP per Capita (log scale)')
plt.ylabel('Life Expectancy')
plt.grid(True)
plt.show()
```



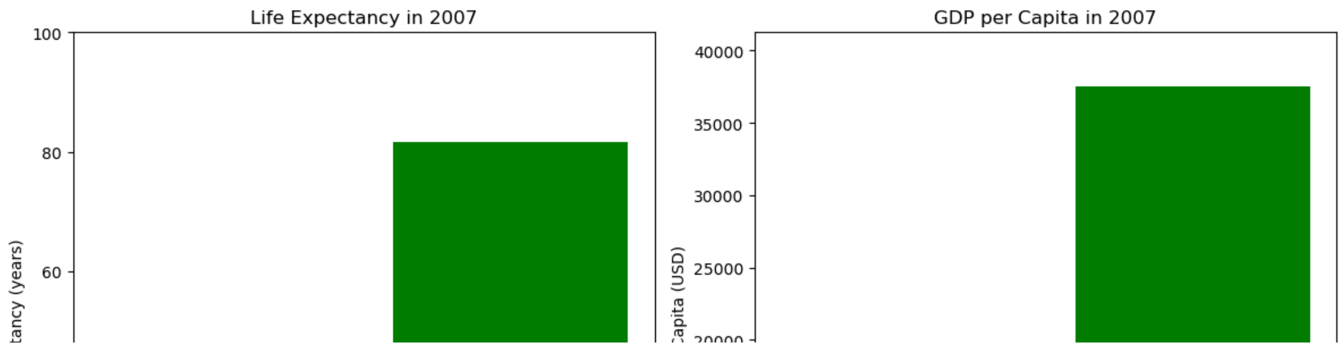
Bar chart showing Comparition of the average GDP per capita for each continent in the year 2007.

```
avg_gdp_per_continent = gapminder_world_2007.groupby('continent')['gdp_cap'].mean()

plt.figure(figsize = (10, 6))
avg_gdp_per_continent.plot(kind = 'bar')
plt.title('Average GDP per Capita by Continent in 2007')
plt.xlabel('Continent')
plt.ylabel('Average GDP per Capita')
plt.xticks(rotation = 45)
plt.grid(axis = 'y')
plt.show()
```



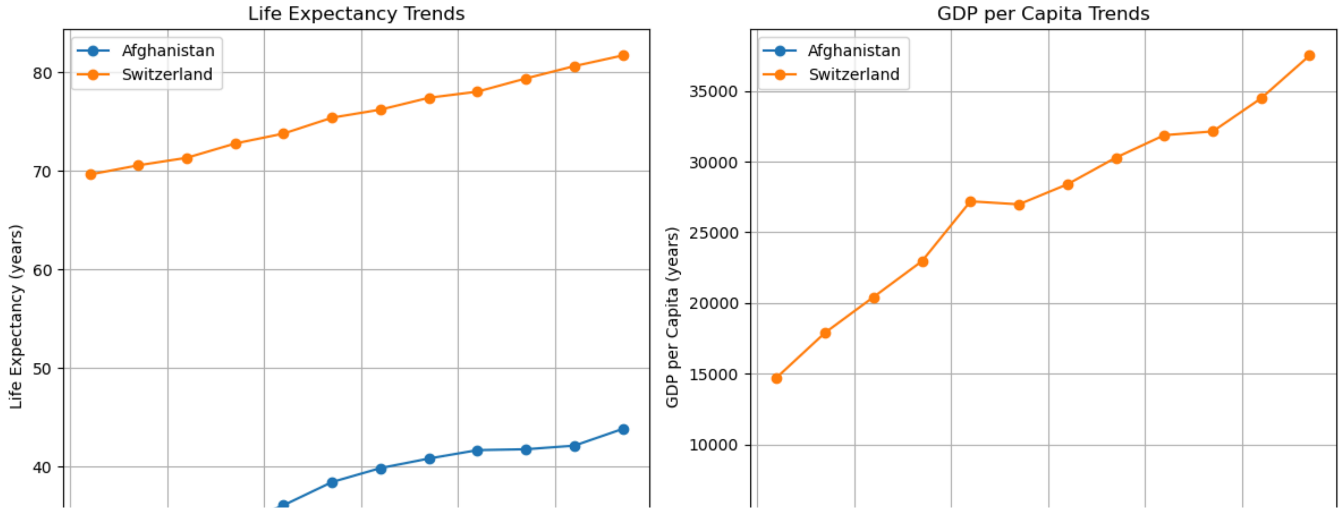
Bar graphs showing the comparison of the the life expectancy and GDP per capita of Afghanistan (a country known for its historical conflicts) and Switzerland (representing a peaceful and economically prosperous country) in the year 2007.

```
gapminder_world_2007_afg_swiss = gapminder_world_2007[gapminder_world_2007['country'].isin(['Afghanistan', 'Switzerland'])]

plt.figure(figsize = (12, 6))

plt.subplot(1, 2, 1)
plt.bar(gapminder_world_2007_afg_swiss['country'], gapminder_world_2007_afg_swiss['life_exp'], color = ['red', 'green'])
plt.title('Life Expectancy in 2007')
plt.ylabel('Life Expectancy (years)')
plt.ylim(0, 100)

plt.subplot(1, 2, 2)
plt.bar(gapminder_world_2007_afg_swiss['country'], gapminder_world_2007_afg_swiss['gdp_cap'], color = ['red', 'green'])
plt.title('GDP per Capita in 2007')
plt.ylabel('GDP per Capita (USD)')
plt.ylim(0, max(gapminder_world_2007_afg_swiss['gdp_cap']) * 1.1)

plt.tight_layout()
plt.show()
```



Line graphs showing the trends of life expectancy and GDP per capita of Afghanistan and Switzerland over all available years in the dataset.

```
gapminder_world_afg_swiss = gapminder_world[gapminder_world['country'].isin(['Afghanistan', 'Switzerland'])]

plt.figure(figsize = (12, 6))
plt.subplot(1, 2, 1)
for country in ['Afghanistan', 'Switzerland']:
    country_data = gapminder_world_afg_swiss[gapminder_world_afg_swiss['country'] == country]
    plt.plot(country_data['year'], country_data['life_exp'], marker = 'o', label = country)
plt.title('Life Expectancy Trends')
plt.xlabel('Year')
plt.ylabel('Life Expectancy (years)')
plt.legend()
plt.grid(True)

plt.subplot(1, 2, 2)
for country in ['Afghanistan', 'Switzerland']:
    country_data = gapminder_world_afg_swiss[gapminder_world_afg_swiss['country'] == country]
    plt.plot(country_data['year'], country_data['gdp_cap'], marker = 'o', label = country)
plt.title('GDP per Capita Trends')
plt.xlabel('Year')
plt.ylabel('GDP per Capita (years)')
plt.legend()
plt.grid(True)

plt.tight_layout()
plt.show()
```


Life Expectancy Trends
Afghanistan
Switzerland
Life Expectancy (years)
GDP per Capita Trends
Afghanistan
Switzerland
GDP per Capita (years)