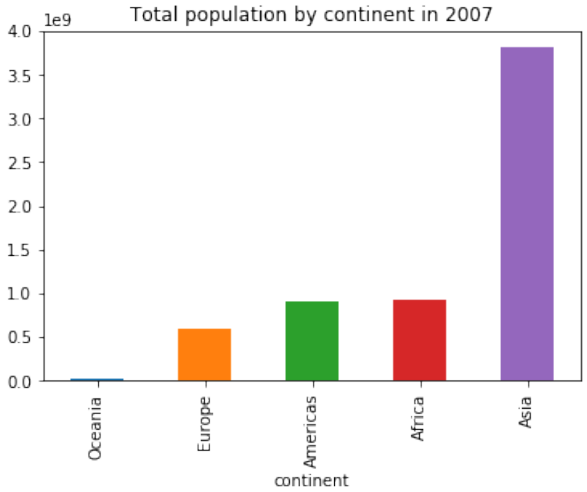
Session 14: Data Aggregation using Groupby (Solutions Only)

Q1: Create a bar plot comparing the total populations of each continent in 2007, as below.
(Hint: First use "query" to filter for year being 2007, then group by the continent and compute the sum of the "pop" column. Then sort the result using "sort_values" and plot using "kind='bar'". All this can be chained together into one line.)

```
[7]: gapminder.query('year==2007').groupby('continent')['pop'].sum()\
         .sort_values().plot(kind='bar',title='Total population by continent in 2007')
```

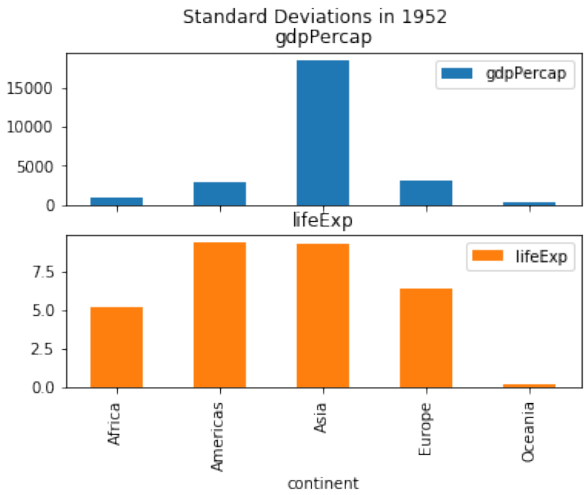
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3e224bbe0>
```



Q2: Create the following plots to compare the standard deviation in GDP per capita and life expectancy across countries within each continent, in 1952 and in 2007.
(Hint: for each graph, first use "query" to filter for the year, then group by the continent, and compute the standard deviation of both "gdpPercap" and "lifeExp". Then plot using "subplots=True". Each plot can be created using one line by chaining together commands.)

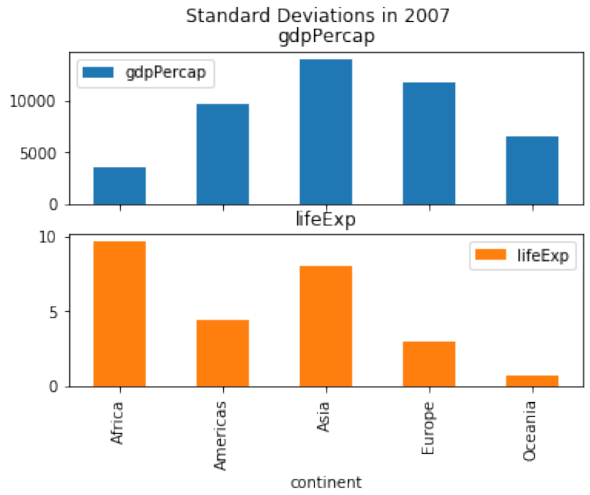
```
[8]: gapminder.query('year==1952').groupby('continent')[['gdpPercap','lifeExp']].std()\
         .plot(kind='bar',subplots=True,title='Standard Deviations in 1952')
```

```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7fe3e224be10>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7fe3e21a4908>],
      dtype=object)
```



```
[9]: gapminder.query('year==2007').groupby('continent')[['gdpPercap','lifeExp']].std()\
        .plot(kind='bar',subplots=True,title='Standard Deviations in 2007')
```

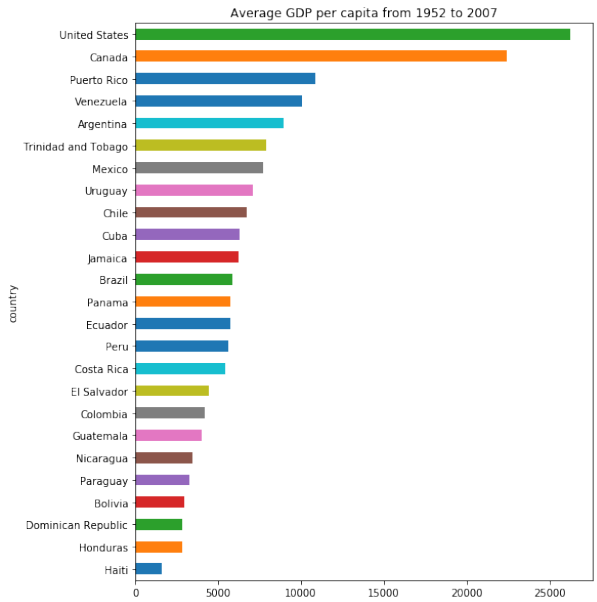
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7fe3e21f8668>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7fe3e20fbf60>],
      dtype=object)
```



Q3: Plot the average GDP per capita over the years in the dataset for all countries in the continent "Americas", as below.

```
[10]: gapminder.query('continent=="Americas"').groupby('country')['gdpPercap'].mean()\
         .sort_values()\
         .plot(kind='barh',title='Average GDP per capita from 1952 to 2007',figsize=(8,10))
```

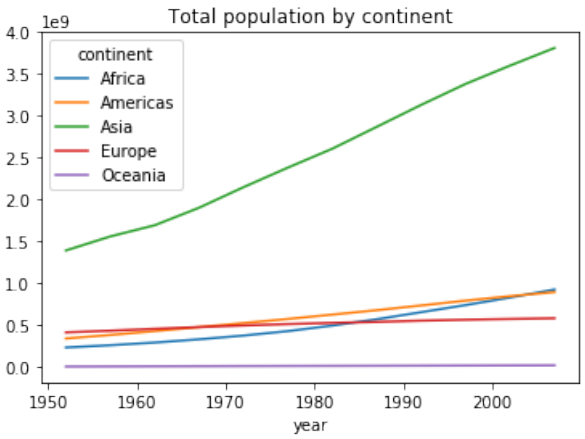
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3e2025320>
```



Q4: Plot the trend in total population of each continent as below.

```
[20]: gapminder.groupby(['year','continent'])['pop'].sum()\
         .unstack().plot(title="Total population by continent")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3e1ebfb38>
```



Q5: Compute the average GDP per capita for each continent in 1952 and 2007, and plot the ratio.

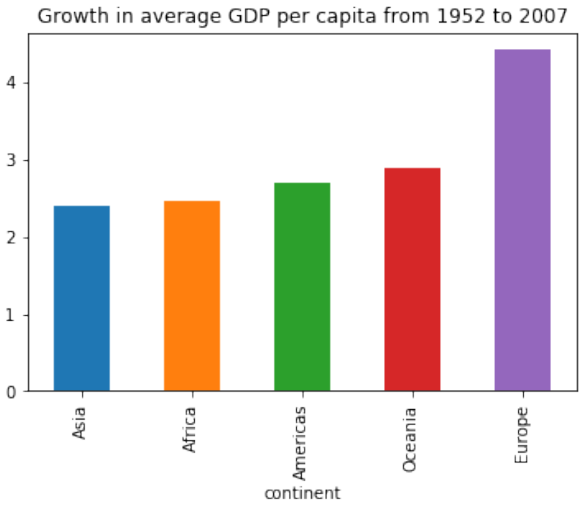
(Hint: a quick way is to first group by the continent and year and compute the average GDP Per capita for each combination, then unstack it so that the years are the columns, similar to in Out[30]. Then you can compute the desired ratio by dividing the column for 2007 by the column for 1952.)

```
[21]: df=gapminder.groupby(['year','continent'])['gdpPercap'].mean().unstack(0)
      df[[1952,2007]].head()
```

year	1952	2007
continent		
Africa	1252.572466	3089.032605
Americas	4079.062552	11003.031625
Asia	5195.484004	12473.026870
Europe	5661.057435	25054.481636
Oceania	10298.085650	29810.188275

```
[22]: (df[2007]/df[1952]).sort_values()\
          .plot(kind='bar',title='Growth in average GDP per capita from 1952 to 2007')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3e1dceb38>
```



Q6. Create a plot over time of the difference in total GDP between the richest and the poorest continent, as below.

(Hint: first add a "GDP" column in the gapminder DataFrame by multiplying the "gdpPercap" and "pop" columns. Then create a "gdpSum" DataFrame by grouping by the year and continent, and summing the GDPs. See below for what this DataFrame looks like. Using this DataFrame, you can compute a Series called "maxGDP" by grouping by the year and finding the max GDP, and similarly compute a Series called "minGDP". Both of these are indexed by year. Finally, subtract maxGDP by minGDP and plot the result.)

```
[23]: gapminder['GDP']=gapminder['gdpPercap']*gapminder['pop']
      gapminder.groupby(['year','continent'])['GDP'].sum().reset_index().head()
```

	year	continent	GDP
0	1952	Africa	3.115993e+11
1	1952	Americas	2.943475e+12
2	1952	Asia	1.125160e+12
3	1952	Europe	2.549140e+12
4	1952	Oceania	1.083144e+11

```
[24]: gapminder.groupby(['year','continent'])['GDP'].sum().reset_index()\
          .groupby('year')['GDP'].max()
```

```
year
1952    2.943475e+12
1957    3.520427e+12
1962    4.228827e+12
1967    5.446688e+12
1972    6.703979e+12
1977    8.102135e+12
1982    9.082850e+12
1987    1.098619e+13
1992    1.224750e+13
1997    1.456733e+13
2002    1.653122e+13
2007    2.070795e+13
Name: GDP, dtype: float64
```

```
[26]: gdpSum=gapminder.groupby(['year','continent'])['GDP'].sum().reset_index()
      maxGDP=gdpSum.groupby('year')['GDP'].max()
      minGDP=gdpSum.groupby('year')['GDP'].min()
      diff=maxGDP-minGDP
      diff.plot(title='Difference in GDP between the riches and the poorest continents')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3e1c51278>
```


Difference in GDP between the riches and the poorest continents
1e13
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
1950
1960
1970
1980
1990
2000
year