GDP List

In [13]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [14]:
```
df = pd.read_csv('..\Data\GDPlistUTF.csv', encoding = 'ISO-8859-1')
```

In [16]:
```
print(df.info(), '\n')
print(df)
df.Country.value_counts()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 3 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   Country                125 non-null    object
 1   Continent              125 non-null    object
 2   GDP (millions of US$)  125 non-null    int64 
dtypes: int64(1), object(2)
memory usage: 3.1+ KB
None

          Country      Continent  GDP (millions of US$)
0     Afghanistan           Asia                  18181
1         Albania         Europe                  12847
2         Algeria         Africa                 190709
3          Angola         Africa                 100948
4       Argentina  South America                 447644
..            ...            ...                    ...
120    Uzbekistan           Asia                  45353
121     Venezuela  South America                 315841
122       Vietnam           Asia                 122722
123         Yemen         Africa                  33675
124        Zambia         Africa                  19206

[125 rows x 3 columns]
```

Out[16]:
```
Country
 Afghanistan          1
 New Zealand          1
 Romania              1
 Qatar                1
 Portugal             1
                     ..
 Equatorial Guinea    1
 El Salvador          1
 Egypt                1
 Ecuador              1
 Zambia               1
Name: count, Length: 125, dtype: int64
```

Vấn đề của bộ dữ liệu

Bộ dữ liệu được mã hóa với tiêu chuẩn mã hóa ISO-8859-1. Trong thuộc tính Country có chứa một ký tự gọi là **Non-breaking space**. Đây là một ký tự đặc biệt mà tiêu chuẩn mã hóa UTF-8 không có. Ký tự này rất dễ bị nhầm lẫn với ký tự **space**. Vì vậy, khi in ra chúng ta sẽ thấy \xa0 (0xA0) theo ISO.... Và khi mã hóa lại sang UTF-8 sẽ có dạng b'\xc2\xa0...'. Chúng ta có thể kiểm tra như đoạn mã dưới.

Bài đọc trên StackOverFlow: How to remove \xa0 from string in Python?

In [9]:
```
df['Country'][0]
```

Out[9]:
```
'\xa0Afghanistan'
```

In [77]:
```
# Ví dụ xử lý lại encoder
g = df['Country'][0]
g = g.replace(u'\xa0', u'')
print(g)
```

```
Afghanistan
```

In [18]:
```
df['Country'] = [c.replace(u'\xa0', u'') for c in df['Country']]
print(df)
df.Country.value_counts()
```

```
         Country      Continent  GDP (millions of US$)
0    Afghanistan           Asia                  18181
1        Albania         Europe                  12847
2        Algeria         Africa                 190709
3         Angola         Africa                 100948
4      Argentina  South America                 447644
..           ...            ...                    ...
120   Uzbekistan           Asia                  45353
121    Venezuela  South America                 315841
122      Vietnam           Asia                 122722
123        Yemen         Africa                  33675
124       Zambia         Africa                  19206

[125 rows x 3 columns]
```

Out[18]:
```
Country
Afghanistan          1
New Zealand          1
Romania              1
Qatar                1
Portugal             1
                    ..
Equatorial Guinea    1
El Salvador          1
Egypt                1
Ecuador              1
Zambia               1
Name: count, Length: 125, dtype: int64
```

Yêu cầu

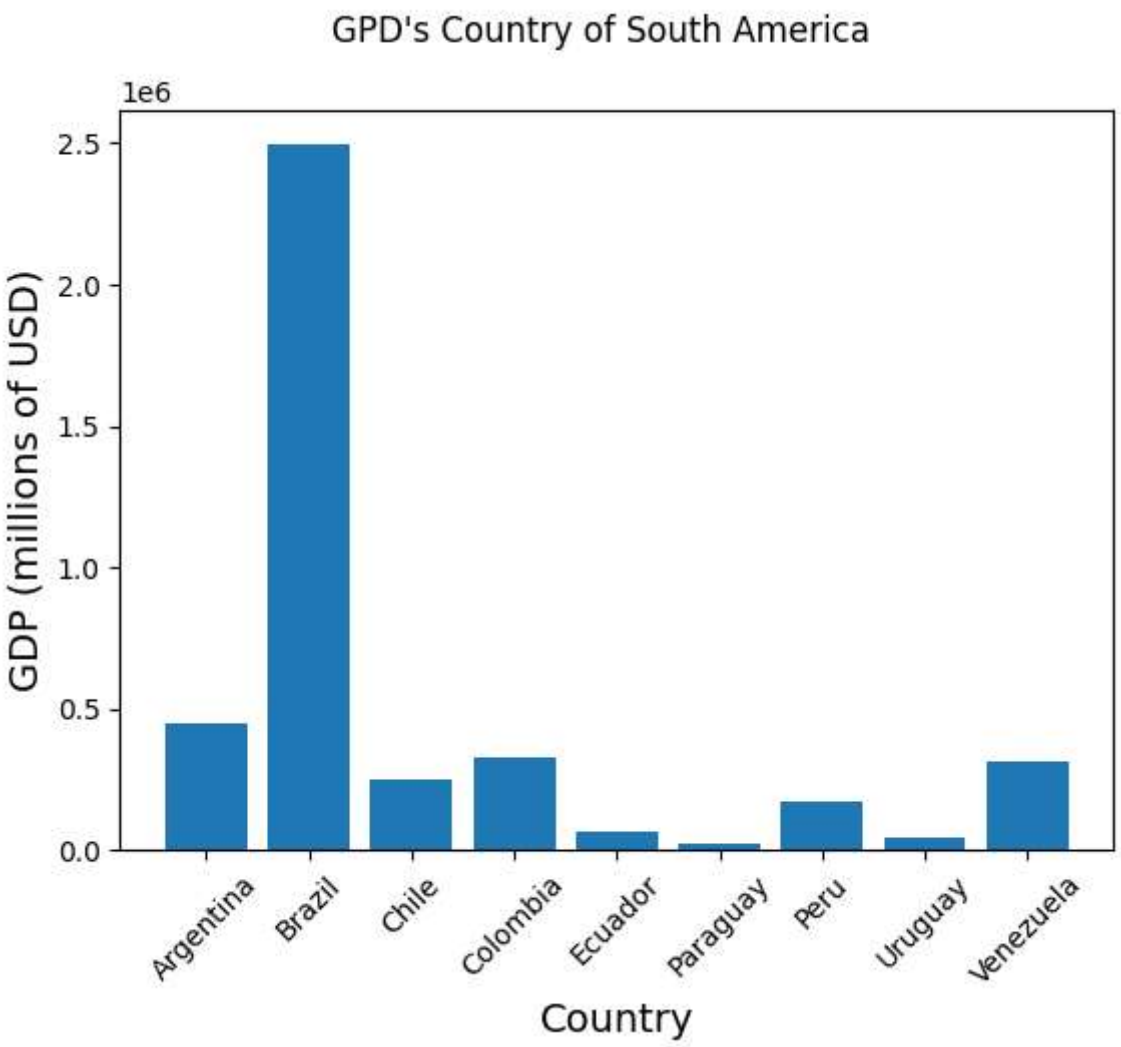
- So sánh GDP các nước ở South America
- Đánh giá tỉ lệ đóng góp GDP của Việt Nam trên tổng số GDP của 5 nước Đông Nam Á là Vietnam, Indonesia, Cambodia, Thailand và Malaysia.

In [19]:
```
df.Continent.value_counts()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 3 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   Country                125 non-null    object
 1   Continent              125 non-null    object
 2   GDP (millions of US$)  125 non-null    int64 
dtypes: int64(1), object(2)
memory usage: 3.1+ KB
```

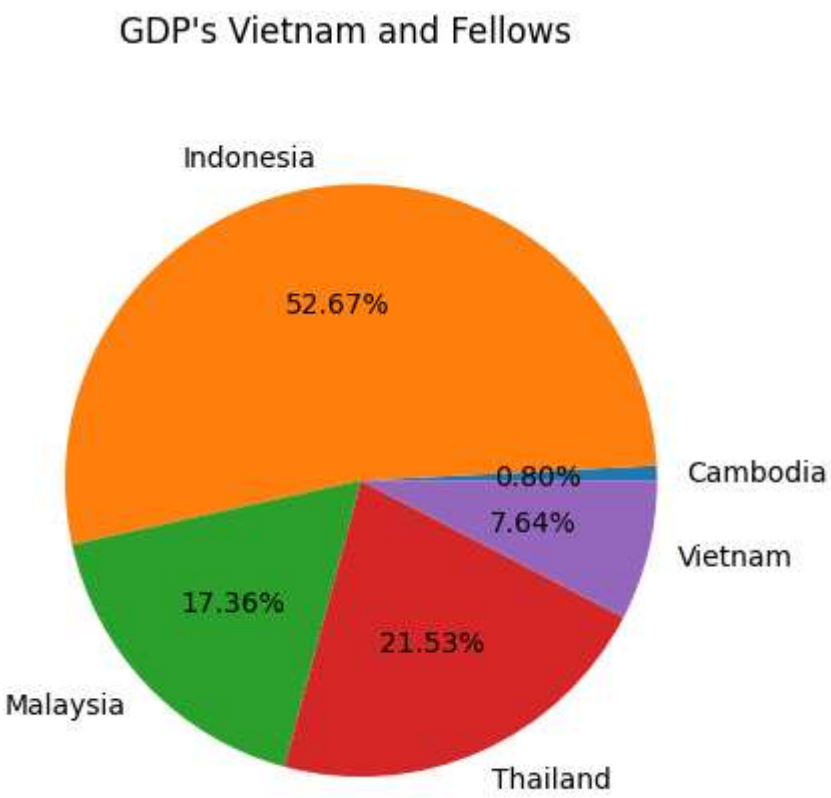
In [101…

```
df_1 = df[df.Continent == 'South America']
plt.bar(df_1.Country, df_1['GDP (millions of US$)'])
plt.suptitle("GPD's Country of South America")
plt.xlabel('Country', fontsize=14)
plt.ylabel('GDP (millions of USD)', fontsize=14)
plt.xticks(rotation=45)
plt.show()
```



In [117…

```
df_2 = df[df.Country.isin(['Vietnam', 'Indonesia', 'Cambodia', 'Thailand', 'Malaysia'])]
plt.pie(df_2['GDP (millions of US$)'], labels=df_2.Country, autopct='%1.2f%%')
plt.suptitle('GDP\'s Vietnam and Fellows')
plt.show()
```


GDP's Vietnam and Fellows
Indonesia
52.67%
0.80%
Cambodia
7.64%
Vietnam
17.36%
21.53%
Malaysia
Thailand