life_expectancy

September 14, 2021

1 Life Expectancy and GDP

1.1 Introduction

Is there a correlation between a country's output and the general life expectancy of its citizens? This is a fundamental question in macroeconomics and metrics such as the WHOQOL (World Health Organization Quality of Life).

In this project we will prepare our data, analyze it with visuals, and explore our findings.

We'll endeavor to answer the following questions: - How has life expectancy changed over time? - How has GDP changed over time? - Is there a correlation between GDP and life expectancy? - What are the distributions of life expectency and GDP?

1.1.1 Imports

Our imports are straightforward - we require only pandas and two graphing utilities.

[32]:
```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

1.1.2 Importing Data

First we will load our datasets into pandas so they can be more easily visualized. We'll also look at the first few values and overall shape of the dataframe to get an idea of the data we will be working with.

[33]:
```
df = pd.read_csv("all_data.csv")
print(df.head())
print(df.shape)
```

```
  Country  Year  Life expectancy at birth (years)          GDP
0   Chile  2000                             77.3  7.786093e+10
1   Chile  2001                             77.3  7.097992e+10
2   Chile  2002                             77.8  6.973681e+10
3   Chile  2003                             77.9  7.564346e+10
```

```
4   Chile  2004                                  78.0  9.921039e+10
(96, 4)
```

1.1.3 Data Exploration

There are six countries in our set: Chile, China, Germany, Mexico, the US, and Zimbabwe whose data spans from 2000-2015.

[34]:
```
print(df.Country.unique())
print(df.Year.unique())
```

```
['Chile' 'China' 'Germany' 'Mexico' 'United States of America' 'Zimbabwe']
[2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
 2014 2015]
```

1.2 Cleaning up the Column Names

When we look closer at the data there are inconsistencies. For example, the column `Life expectancy at birth (years)` is a bit too descriptive and makes the coding more difficult to understand. Let's change the column to something more simple: `LEABY`.

[35]:
```
df = df.rename({"Life expectancy at birth (years)":"LEABY"}, axis = "columns")
df.head()
```

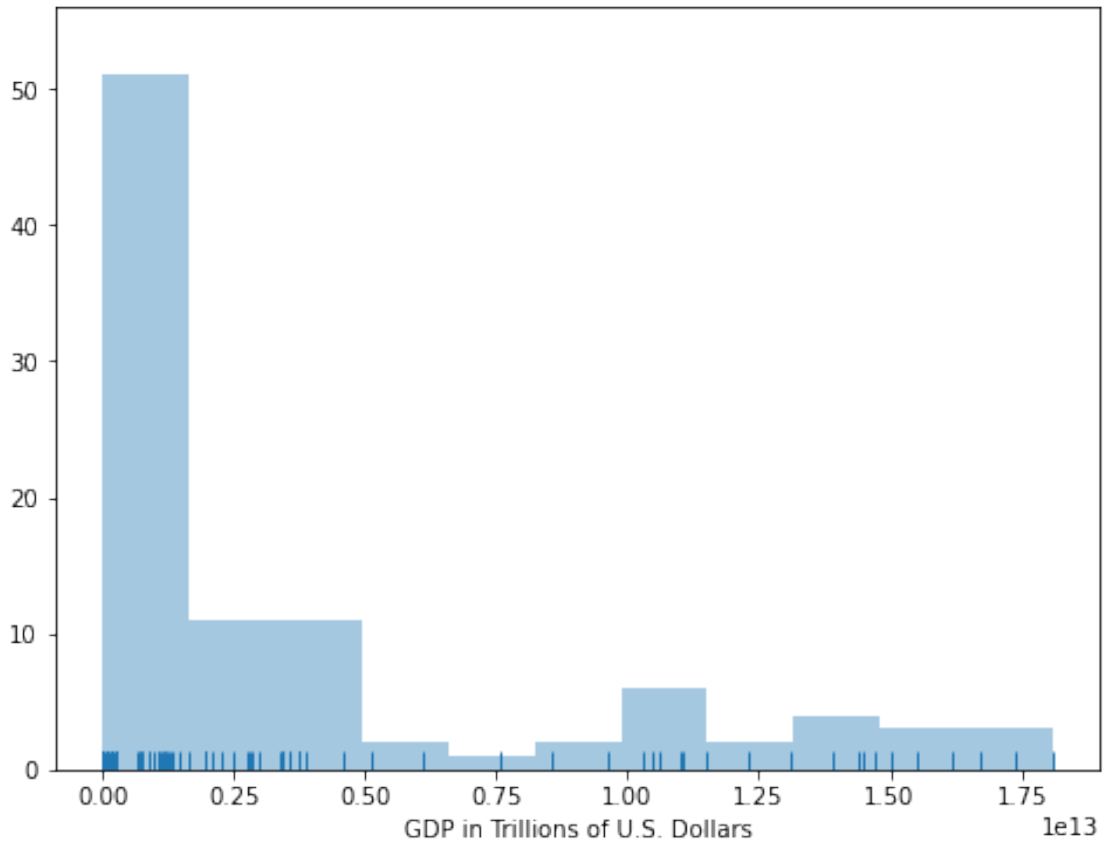
[35]:

	Country	Year	LEABY	GDP
0	Chile	2000	77.3	7.786093e+10
1	Chile	2001	77.3	7.097992e+10
2	Chile	2002	77.8	6.973681e+10
3	Chile	2003	77.9	7.564346e+10
4	Chile	2004	78.0	9.921039e+10

1.3 Exploratory Plots

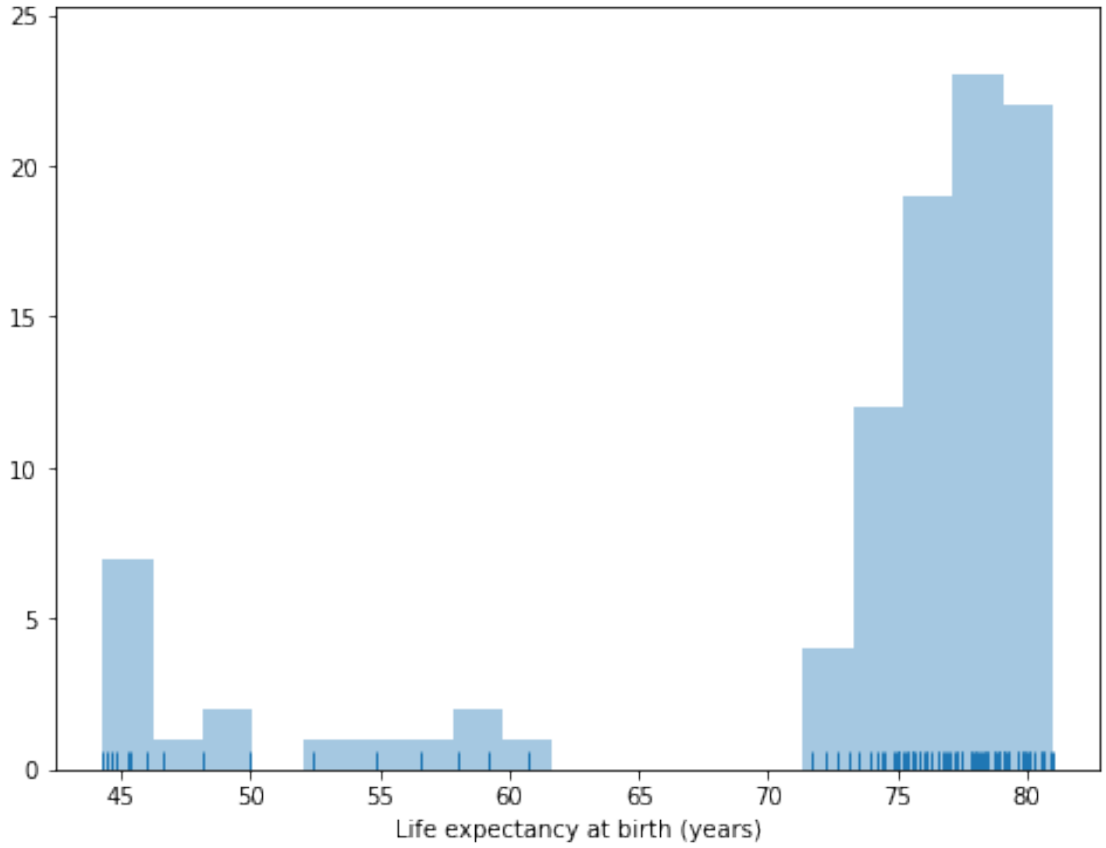
First we will look at the distribution of `GDP` to get a better understanding of the overall distribution. With a strong right skew, this is representative of a power law distribution where one quality varies with another on a proportionally smaller/larger scale. In other words, a minority of countries have large GDPs.

[36]:
```
plt.figure(figsize=(8,6))
sns.distplot(df.GDP, rug = True, kde=False)
plt.xlabel("GDP in Trillions of U.S. Dollars");
```



Now let's take a look at the `LEABY` distribution. We see a left-skewed plot indicating some unique groupings.

[37]:
```
plt.figure(figsize=(8,6))
sns.distplot(df.LEABY, rug = True, kde=False)
plt.xlabel("Life expectancy at birth (years)");
```



Broken up by country, the data is a bit easier to follow and it's clear that Zimbabwe in particular is likely responsible for the odd groupings in the previous chart.

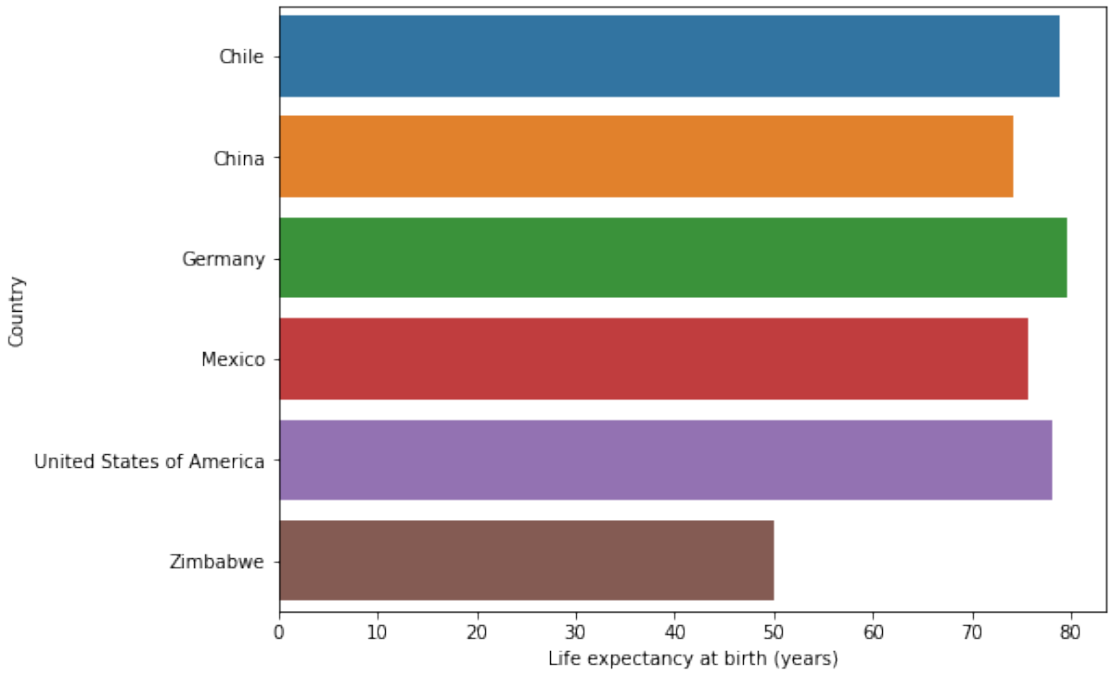
[38]:
```
dfMeans = df.drop("Year", axis = 1).groupby("Country").mean().reset_index()
```

[39]:
```
print(dfMeans)
```

```
                    Country     LEABY           GDP
0                     Chile  78.94375  1.697888e+11
1                     China  74.26250  4.957714e+12
2                   Germany  79.65625  3.094776e+12
3                    Mexico  75.71875  9.766506e+11
4  United States of America  78.06250  1.407500e+13
5                  Zimbabwe  50.09375  9.062580e+09
```

Let's make some bar plots of this new data. As we expected Zimbabwe is the outlier.

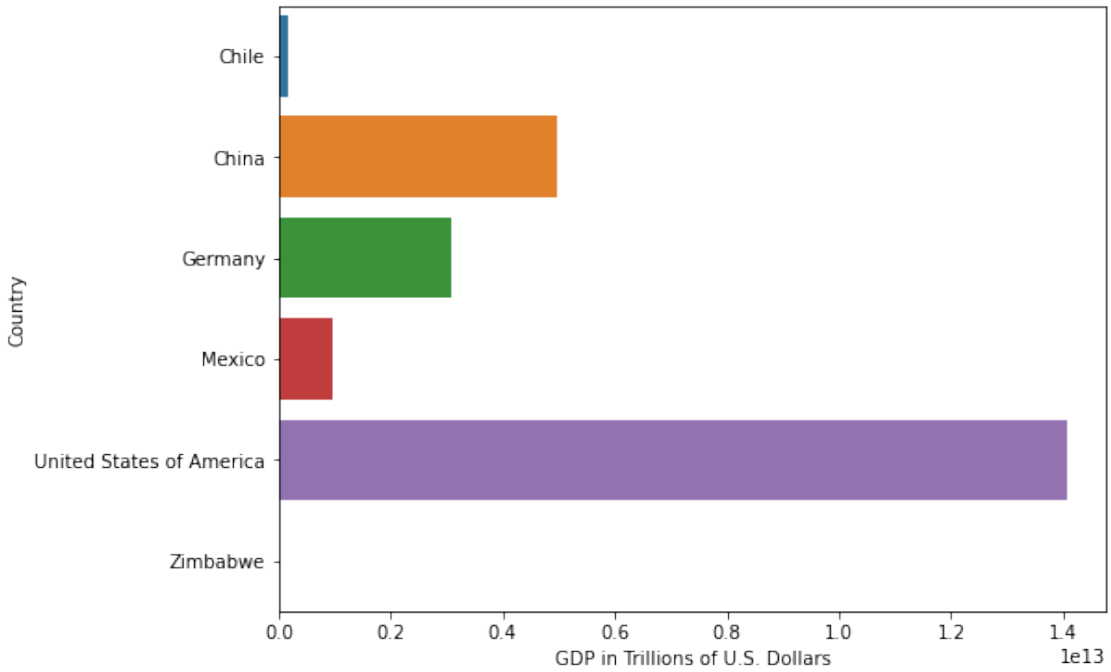
[40]:
```
plt.figure(figsize=(8,6))
sns.barplot(x="LEABY", y="Country", data=dfMeans)
plt.xlabel("Life expectancy at birth (years)");
```



Now we can perform a similar breakdown of GDP by country. The United States has a significantly higher GDP than every other country, yet its life expectancy is on par with the others.

[41]:
```
plt.figure(figsize=(8,6))
sns.barplot(x="GDP", y="Country", data=dfMeans)
plt.xlabel("GDP in Trillions of U.S. Dollars")
```

[41]:
```
Text(0.5, 0, 'GDP in Trillions of U.S. Dollars')
```

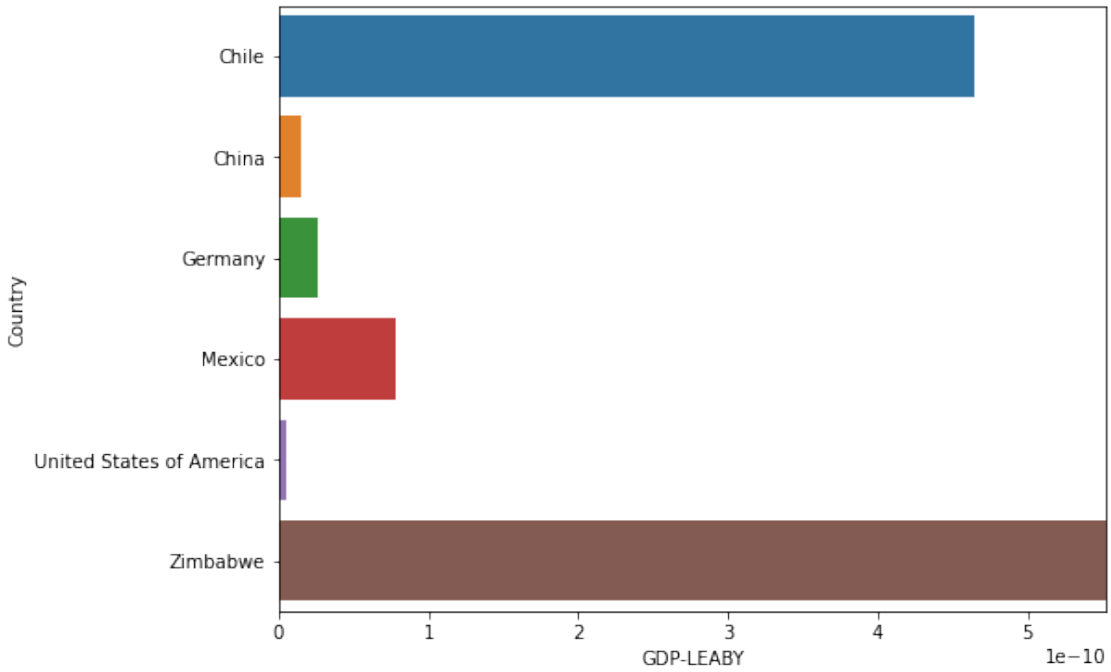


If we break down life expectancy in relation to GDP the results are intriguing. Once the scale is lowered to account for Zimbabwe, the U.S. actually performs significantly worse in life expectancy compared to other nations with dramatically smaller GDPs.

[42]:
```
dfMeans['GDP_LEABY'] = dfMeans.LEABY / dfMeans.GDP

plt.figure(figsize=(8,6))
sns.barplot(data = dfMeans, x='GDP_LEABY', y='Country')
plt.xlim(0, (max(dfMeans.GDP_LEABY)/10))
plt.xlabel('GDP-LEABY')
```

[42]:
```
Text(0.5, 0, 'GDP-LEABY')
```



1.4 Violin Plots

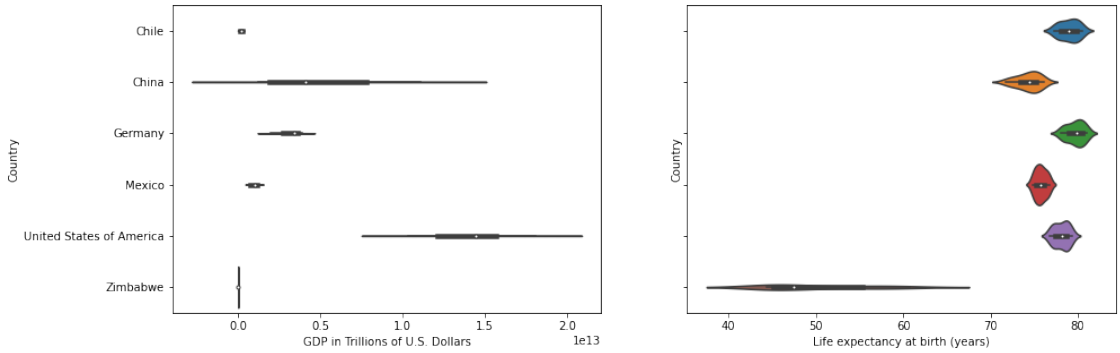
We can also use violin plots to visualize these distributions, using their shapes to compare patterns in the data.

In the left graph (`GDP`), the U.S. and China have wide ranges, while the other countries are much narrower.

In the right plot (`LEABY`), every country apart from Zimbabwe has a similarly narrow range of life expectancy.

[43]:
```
fig, axes = plt.subplots(1, 2, sharey=True, figsize=(15, 5))
axes[0] = sns.violinplot(ax=axes[0], x=df.GDP, y=df.Country)
axes[0].set_xlabel("GDP in Trillions of U.S. Dollars")
axes[1] = sns.violinplot(ax=axes[1], x=df.LEABY, y=df.Country)
axes[1].set_xlabel("Life expectancy at birth (years)")
```

[43]:
```
Text(0.5, 0, 'Life expectancy at birth (years)')
```



1.5 Swarm Plots

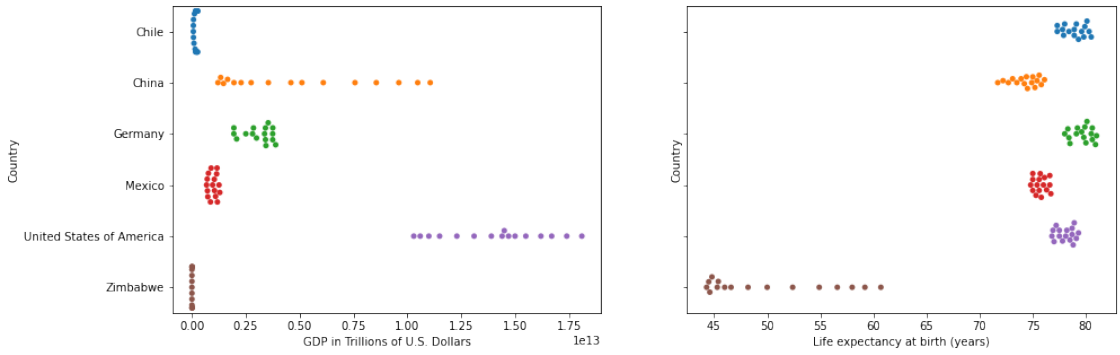
We can also show this data with a swarm plot. Although similar to a violin plot, a swarm plot can show clustering around values and more easily identify outliers in the set. The top graphs are pure swarm plots, and then the bottom graphs show them overlayed on to violin plots.

The left plot (`GDP`) shows more centered lines of dots around Chile and Zimbabwe, indicating their stable GDP over time.

The right plot (`LEABY`) shows much better the variance in Zimbabwe's life expectancy.

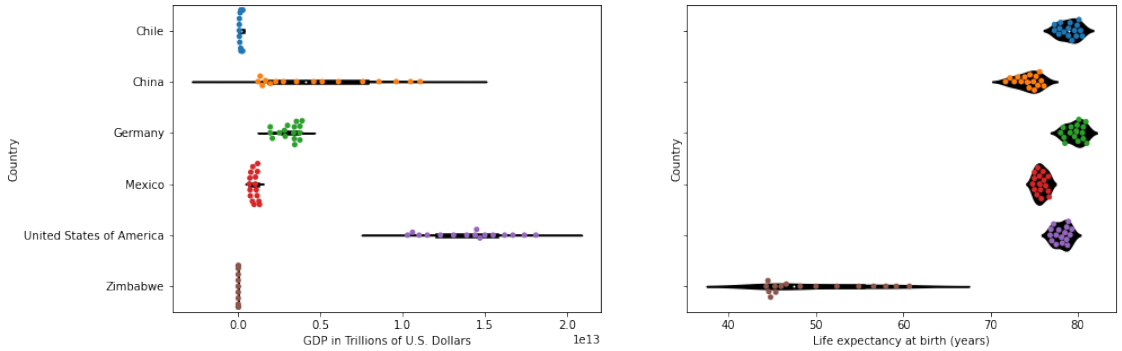
[44]:
```
fig, axes = plt.subplots(1, 2, sharey=True, figsize=(15, 5))
axes[0] = sns.swarmplot(ax=axes[0], x=df.GDP, y=df.Country)
axes[0].set_xlabel("GDP in Trillions of U.S. Dollars")
axes[1] = sns.swarmplot(ax=axes[1], x=df.LEABY, y=df.Country)
axes[1].set_xlabel("Life expectancy at birth (years)")
```

[44]:
```
Text(0.5, 0, 'Life expectancy at birth (years)')
```



[45]:
```
fig, axes = plt.subplots(1, 2, sharey=True, figsize=(15, 5))
axes[0] = sns.violinplot(ax=axes[0], x=df.GDP, y=df.Country,color = "black")
axes[0] = sns.swarmplot(ax=axes[0], x=df.GDP, y=df.Country)
```

```
axes[0].set_xlabel("GDP in Trillions of U.S. Dollars")
axes[1] = sns.violinplot(ax=axes[1], x=df.LEABY, y=df.Country, color = "black")
axes[1] = sns.swarmplot(ax=axes[1], x=df.LEABY, y=df.Country)
axes[1].set_xlabel("Life expectancy at birth (years)");
```

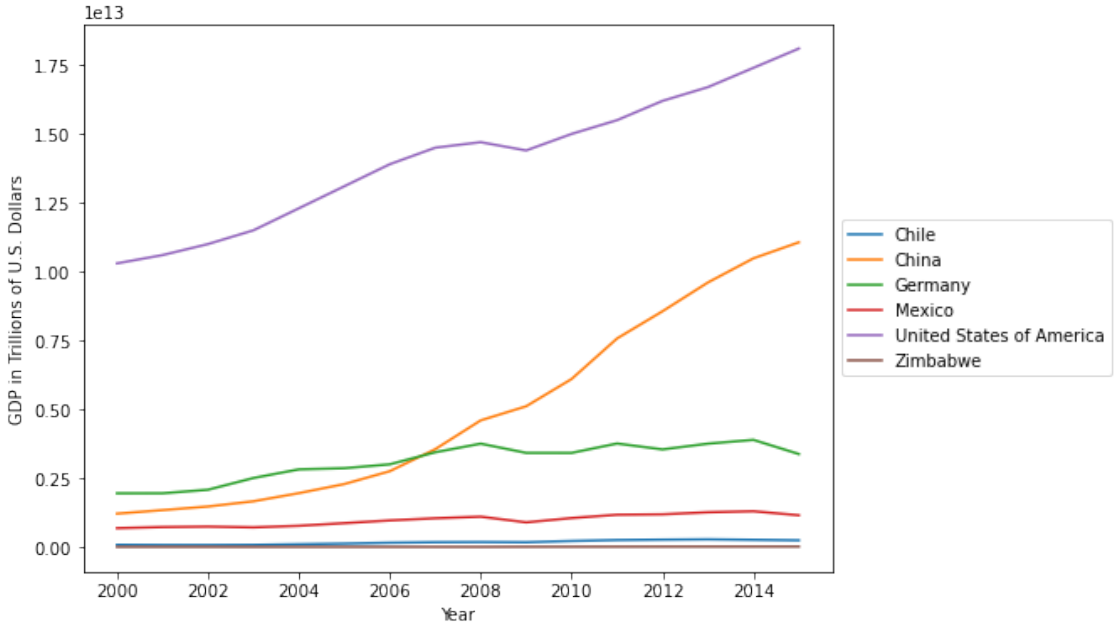


1.6 Line Charts

Now we will use more traditional line graphs to explore the growth of GDP over the dataset's timeframe. Although this was also present in the previous violin and swarm plots, it is easier to interpret the rapid growth of both China and U.S. over this 15 year period.

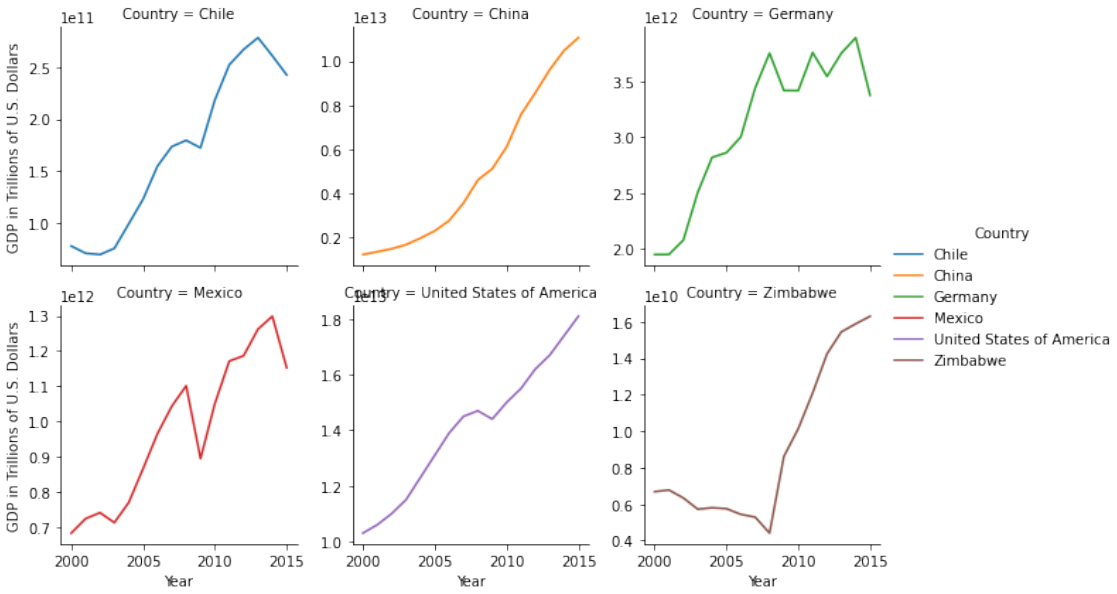
[46]:
```
plt.figure(figsize=(8,6))
sns.lineplot(x=df.Year, y=df.GDP, hue=df.Country)
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1)
plt.ylabel("GDP in Trillions of U.S. Dollars")
```

[46]:
```
Text(0, 0.5, 'GDP in Trillions of U.S. Dollars')
```



We can also break down each country's GDP growth within their own scale. This shows that while all countries have seen increases, the scales of growth in China and the U.S. is striking.

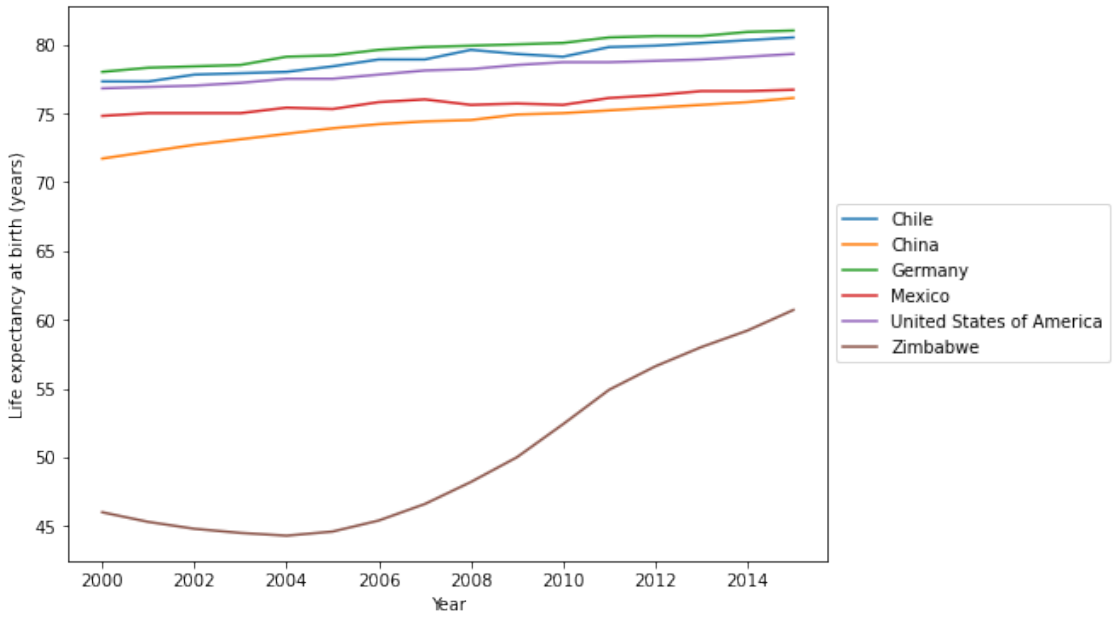
[47]:
```
graphGDP = sns.FacetGrid(df, col="Country", col_wrap=3,
                         hue = "Country", sharey = False)
graphGDP = (graphGDP.map(sns.lineplot,"Year","GDP")
          .add_legend()
          .set_axis_labels("Year","GDP in Trillions of U.S. Dollars"))
```



We can also inspect the rate of LEABY over this timeframe. When comparing the sudden jump in GDP for Zimbabwe it seems correlated with a similar increase in LEABY at that timepoint.

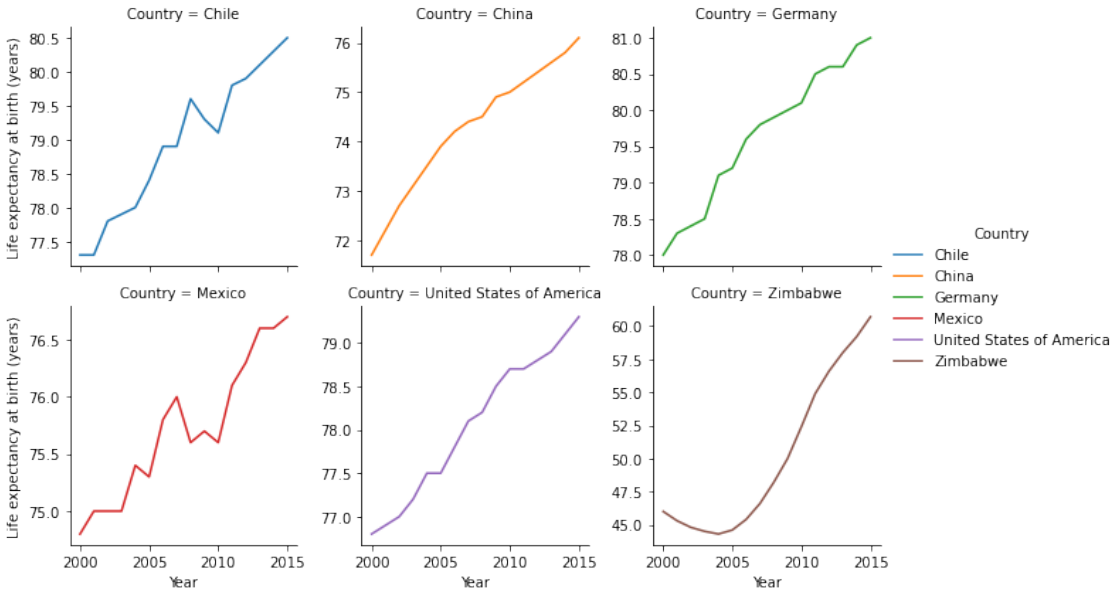
[48]:
```
plt.figure(figsize=(8,6))
sns.lineplot(x=df.Year, y=df.LEABY, hue=df.Country)
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1)
plt.ylabel("Life expectancy at birth (years)")
```

[48]:
```
Text(0, 0.5, 'Life expectancy at birth (years)')
```



Below is LEABY broken down by country, similar to the graphs on GDP above. The dips in Chile and Mexico might be explained by the global finance collapse around that timeframe, and the data further supports the increase in living standards for Zimbabwe.

[49]:
```
graphLEABY = sns.FacetGrid(df, col="Country", col_wrap=3,
                           hue = "Country", sharey = False)
graphLEABY = (graphLEABY.map(sns.lineplot,"Year","LEABY")
              .add_legend()
              .set_axis_labels("Year","Life expectancy at birth (years)"))
```

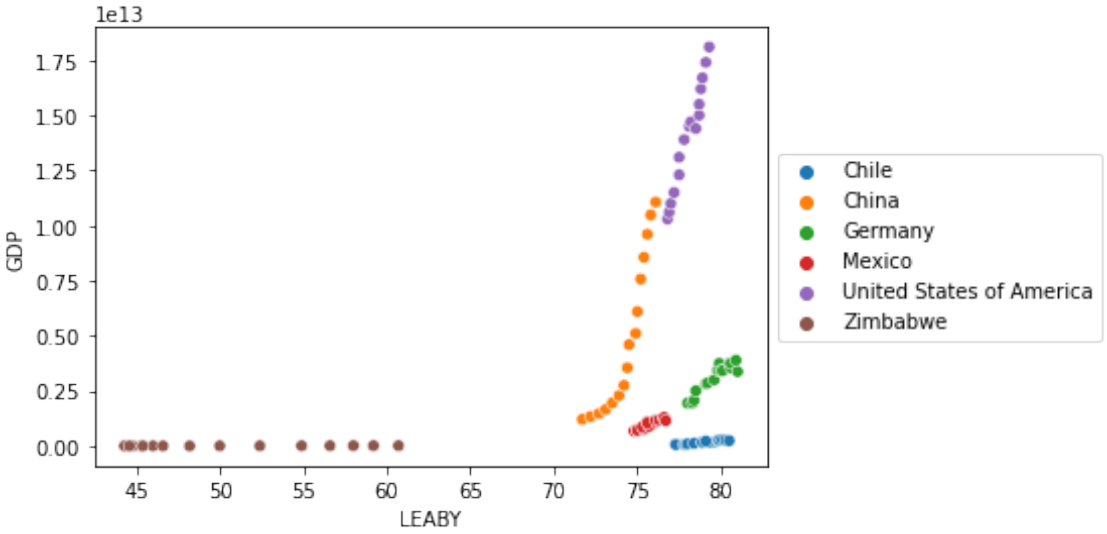


1.7 Scatter Plots

Like the bar plot above, showing `LEABY` in proportion to `GDP`, we can explore their relationship through scatter plots.

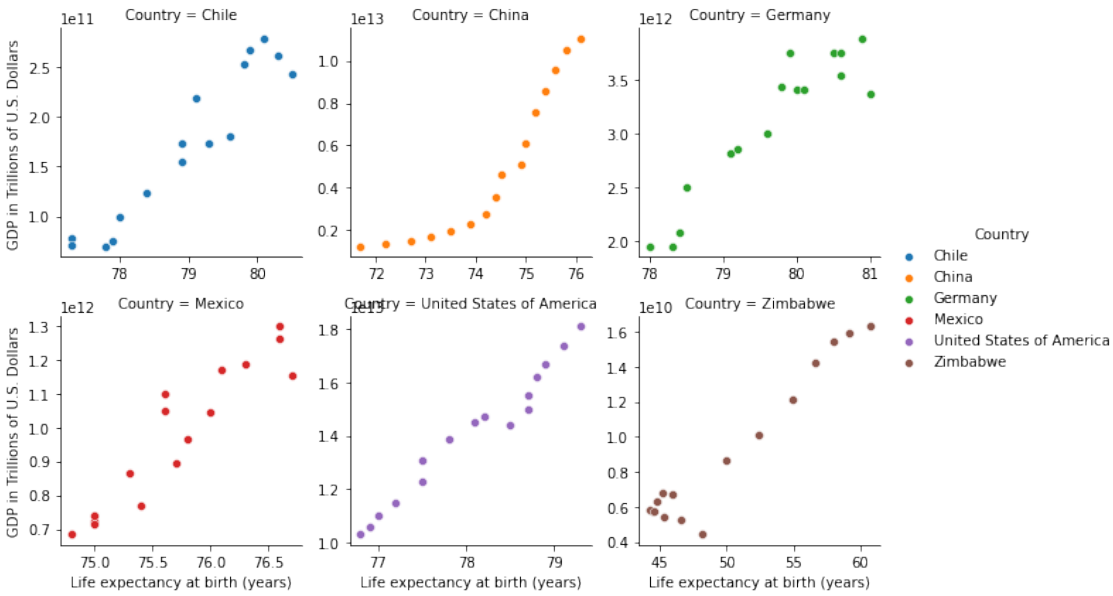
[50]:
```
sns.scatterplot(x=df.LEABY, y=df.GDP, hue=df.Country)\
    .legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1)
```

[50]:
```
<matplotlib.legend.Legend at 0x211b17fa040>
```



We can also break these plots down by country.

[51]:
```
graph = sns.FacetGrid(df, col="Country", col_wrap=3,
                      hue = "Country", sharey = False, sharex = False)
graph = (graph.map(sns.scatterplot,"LEABY", "GDP")
         .add_legend()
         .set_axis_labels("Life expectancy at birth (years)", "GDP in Trillions␣
 ↪of U.S. Dollars"))
```



1.8 Conclusions

In this project, we visually explored and compared the GDP and Life expectancy at birth in six countries. Even though our dataset had limited observations, we were able to derive answers to our earlier questions: - How has life expectancy changed over time? - Life expectancy has overall increased, with Zimbabwe seeing the biggest change. - How has GDP changed over time? - Overall GDP has increased for all of the countries, notably in the U.S. and China. - Is there a correlation between GDP and life expectancy? - There appears to a upward and direct correlation between GDP and LEABY, most easily seen in Zimbabwe. - What are the distributions of life expectancy and GDP? - GDP had a strong right skew, and Life expectancy demonstrated a left skew.

1.9 Further Research

From our investigation there are several areas to explore further. One possible area of further research is the interesting correlation between GDP and LEABY. There are two instances that warrant this. First, Zimbabwe had a strong upward trend with its increase in GDP towards the latter half of our dataset. It would behoove us to do a wider study on countries similar to Zimbabwe and how GDP fluctuations have impacted their LEABY. Second, The U.S. demonstrates a disproportionately low effect of GDP on LEABY in comparison to other nations. Further investigation

is warranted to determine if other confounding factors are causing this and if it would be wise to exclude the U.S. in further investigations of the phenomenon.

Data Sources: The data in this project comes from the World Bank (GDP) and the World Health Organization (life expectancy)