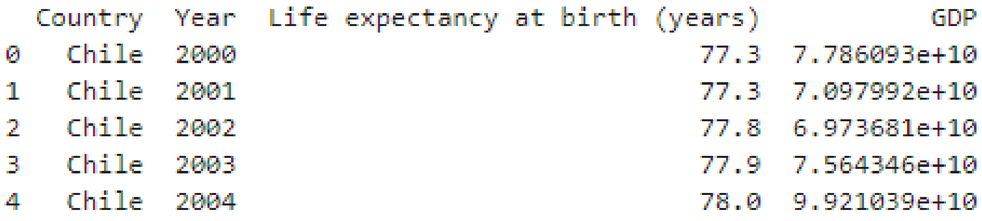
Does GDP affect Life Expectancy?

The purpose of this project is to analyze the effect GDP has on life expectancy. To do so, this project will utilize various Python functions, as well as certain Python libraries (Pandas, NumPy and SciPy), to load and manipulate a dataset - consisting of GDP and life expectancy data for six countries (Zimbabwe, United States, Mexico, Chile, China, and Germany) over the course of 15 years - into a more manageable format, and then calculate numerous statistics regarding that data. The data will then be organized into various charts and graphs using Python libraries such as Matplotlib and Seaborn. Finally, these various statistics and data visualizations will be used to answer the following questions:

- Has life expectancy increased over time in the six nations?
- Has GDP increased over time in the six nations?
- Is there a correlation between GDP and life expectancy of a country?
- If one country has a higher GDP than another country, is that country also likely to have a longer life expectancy?
- What is the average life expectancy in these nations?
- What is the average GDP of these six countries?
- What is the distribution of that life expectancy?
- What is the distribution of GDP among the six countries?

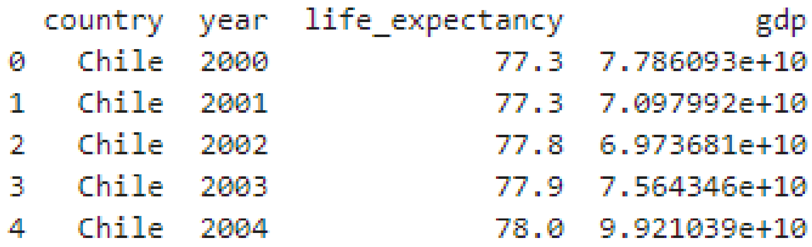
Load and Preview Data



```
  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
```

Data Cleaning

Upon loading and previewing the data, data cleaning and tidying practices are employed by shortening (and otherwise modifying) the names of certain columns while keeping them descriptive. This makes these columns easier to access (e.g. enables the use of dot-notation by replacing all the spaces with underscores, eliminates the need for worrying about capitalization in column names by converting them to lowercase, etc.) without sacrificing readability.



```
  country  year  life_expectancy           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
```

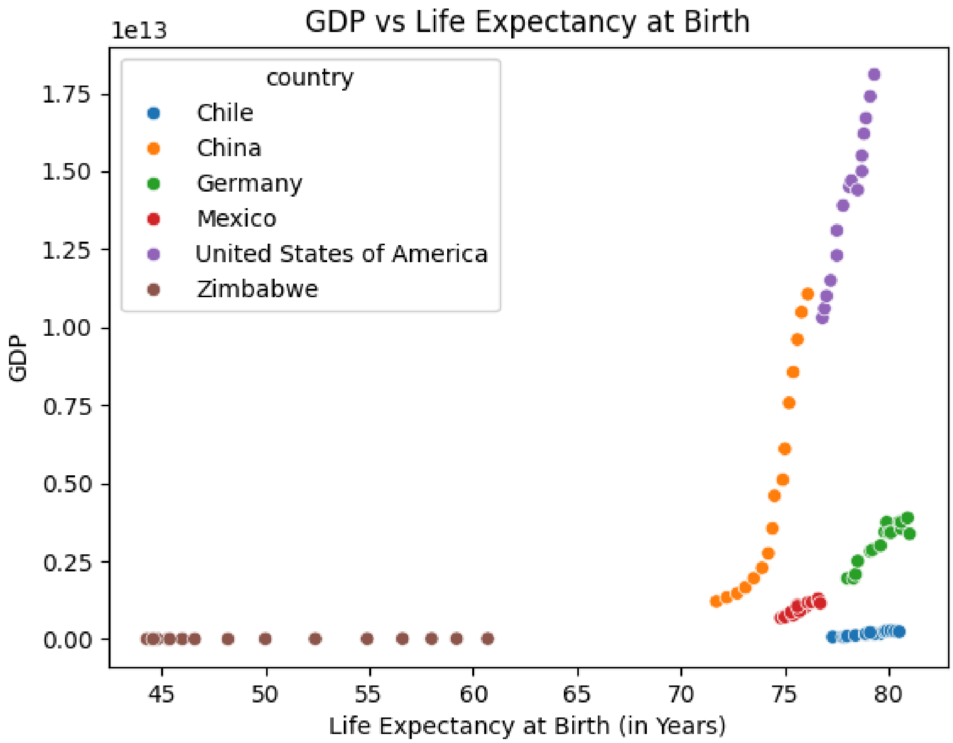
Data Exploration

After the data has been loaded in, previewed and cleaned, the next step is to explore the data visually and statistically. This allows for conclusions to be drawn that would not have been attainable solely through tabular data examination.

Scatter Plot

In the first data visualization, the GDP for each country is plotted against the life expectancy, with each country marked by a different color, in an attempt to discern the impact GDP has on life expectancy. Unfortunately, the graph's legibility is compromised by densely packed point clusters, as some countries have much larger GDPs than others. This indicates the necessity for either creating a separate plot for each country or rescaling the axes. Nonetheless, some observations can still be made from this visualization, as the majority of the point clusters seem to indicate a positive relationship between GDP and life expectancy. Furthermore, from this graph, it can be seen that there is not a strong correlation between the GDP and life expectancy between countries. That is to say that there is strong reason to believe that one country having a higher GDP than another does not indicate a higher life expectancy. This seems to answer the following question (although it's hard to be sure considering the quality of the graph, so later on we'll explore this question further):

- *If one country has a higher GDP than another country, is that country also likely to have a longer life expectancy?*

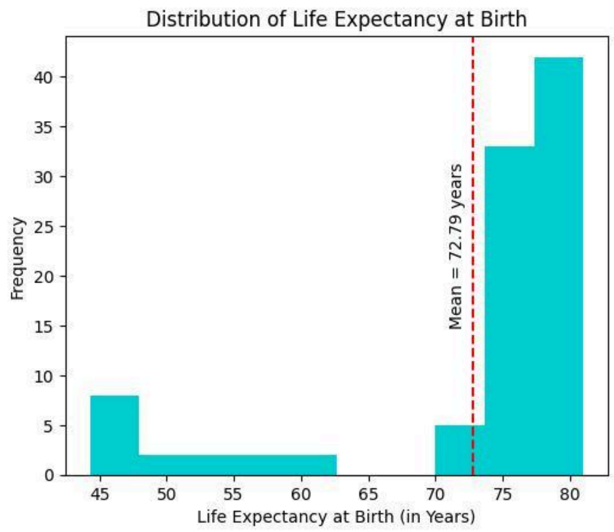


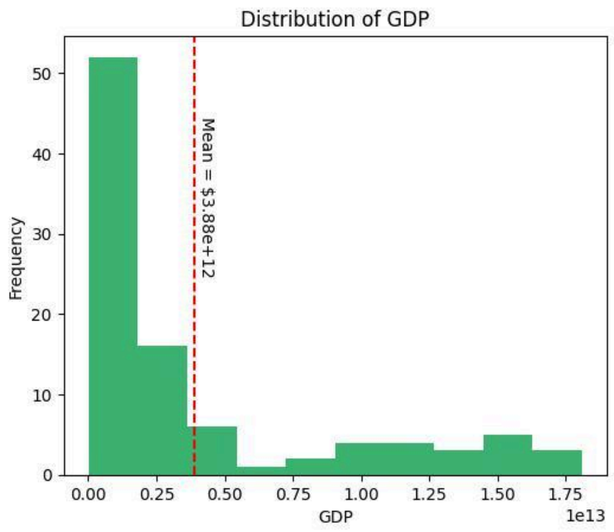
Distribution Plots

Next, histograms of both GDP and Life Expectancy are plotted in an attempt to answer the following questions:

- *What is the distribution of Life Expectancy?*
- *What is the distribution of GDP among the six countries?*

As demonstrated by the plots below, the distribution of Life Expectancy is skewed left, with an average of 72.8 years, while the distribution of GDP is skewed right, with an average of approximately $3.9 Trillion.



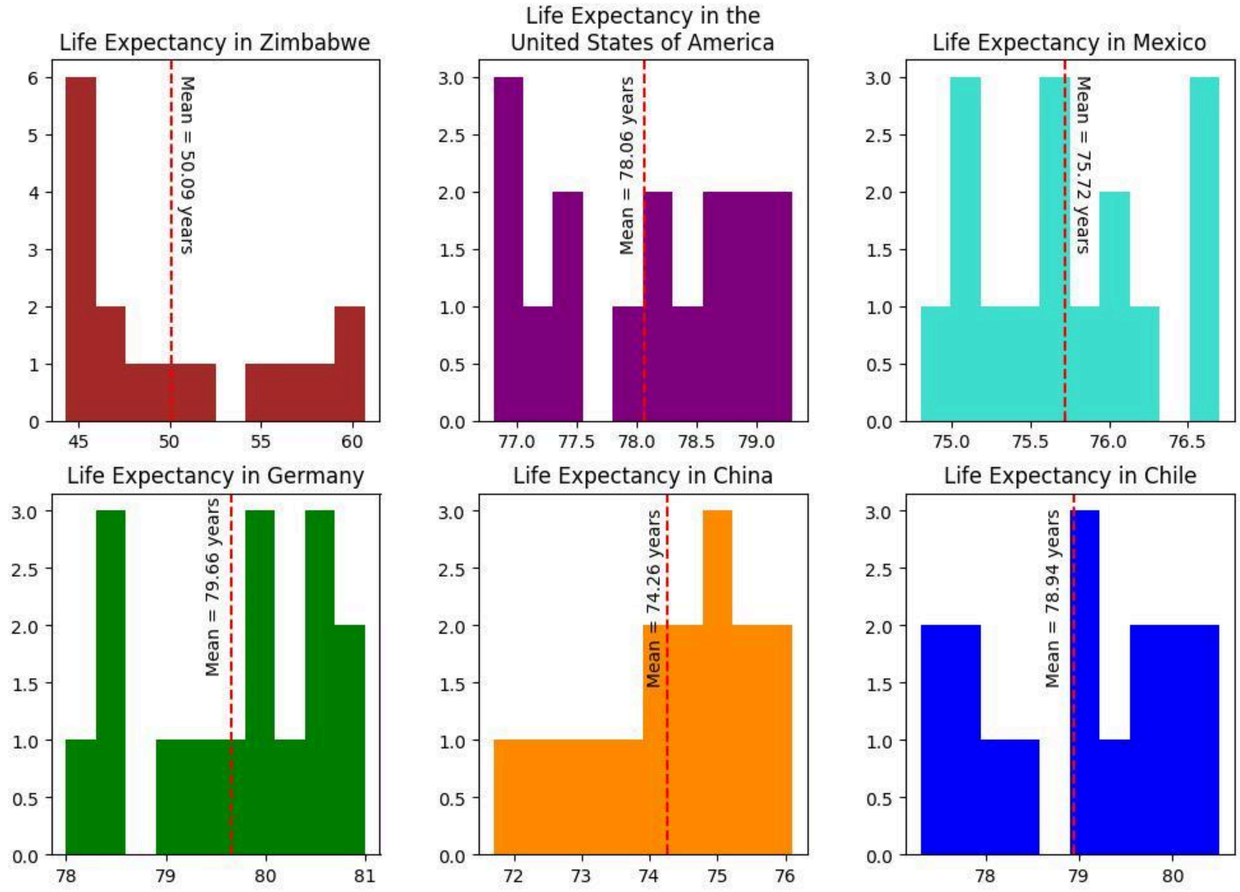


Life Expectancy Distribution Plots per Country

Subsequently, histograms for the Life Expectancy of each country were plotted for the purpose of visualizing each country's Life Expectancy distributions. Furthermore, these distribution plots answer the following question:

- *What is the average life expectancy in these nations?*

As one can see, many of these distributions are relatively normal with a few outliers. However, a few of these plots appear to be rather skewed towards one direction or the other, implying a wider range of life expectancies over the course of 15 years.

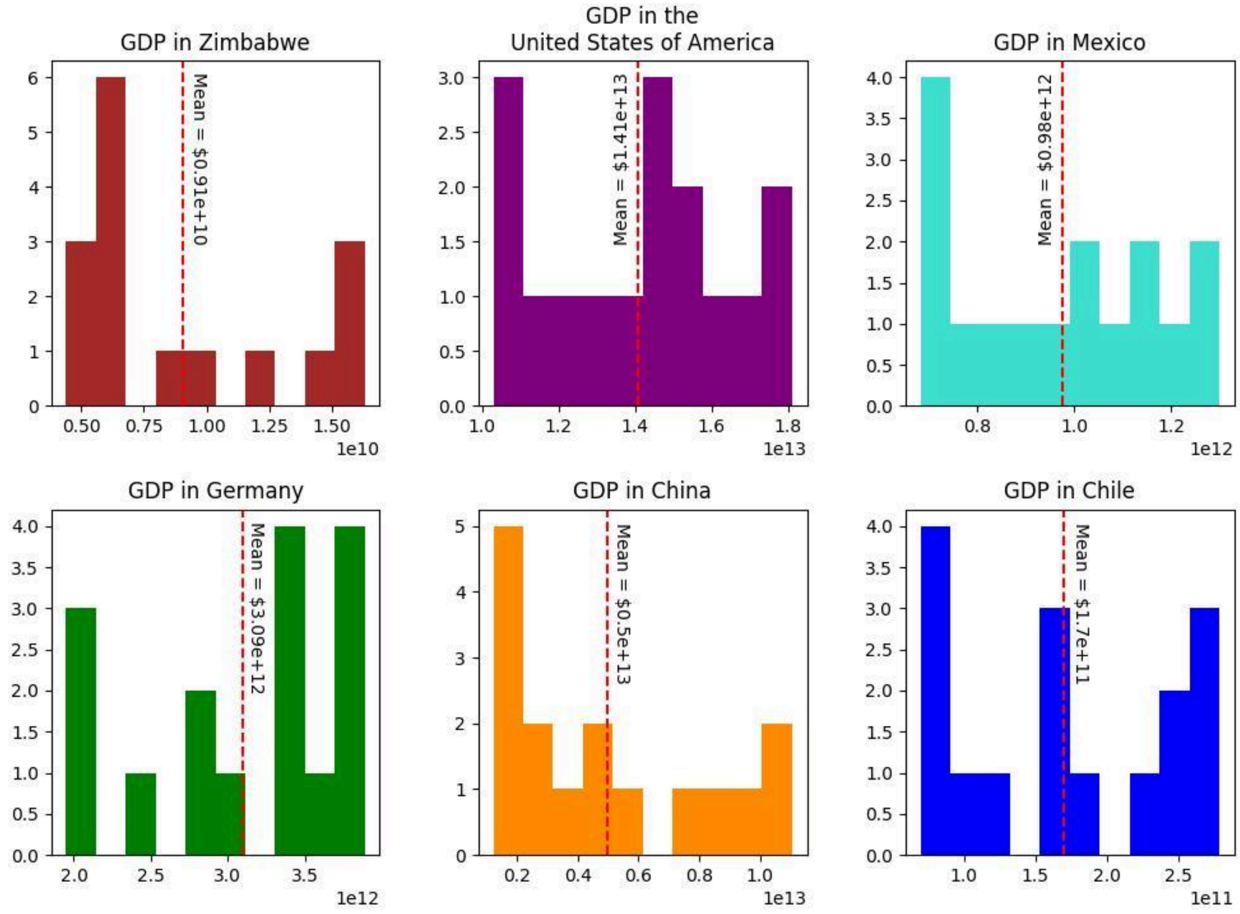


GDP Distribution Plots per Country

Likewise, histograms for the GDP of each country were also plotted for the purpose of visualizing each country's GDP distributions over the 15 year time period from 2000-2015. These distribution plots aim to answer the following question:

- *What is the average GDP in these nations?*

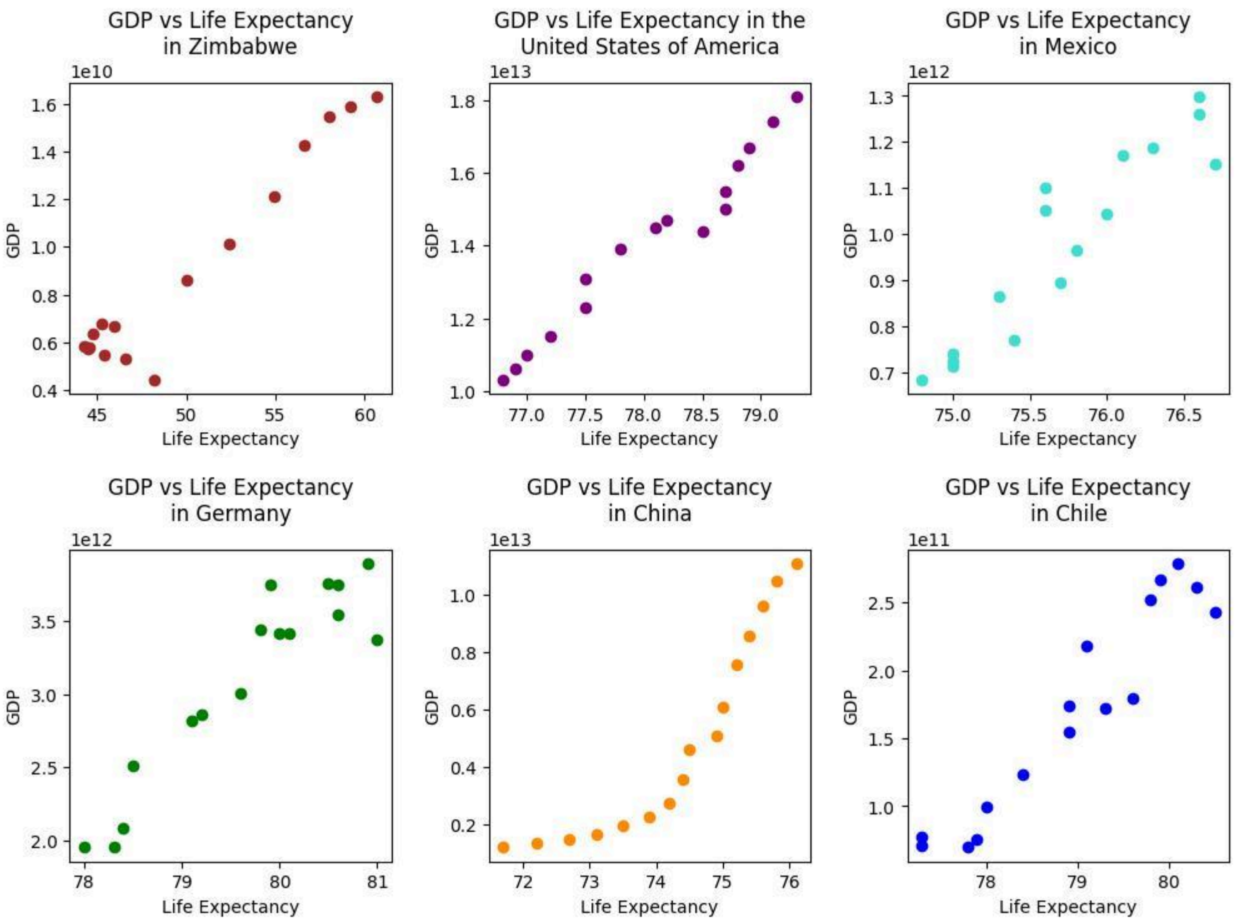
As one can see, these distributions closely resemble the opposite of the life expectancy distributions, with many being rather skewed and only a few being relatively normal.



Scatter Plots per Country

Once the histograms have been plotted and the distributions analyzed for each country, the next step is to plot a scatter plot to picture the relationship between the GDP and life expectancy for each country and to begin an attempt at answering the following question:

- *Is there a correlation between GDP and life expectancy of a country?*



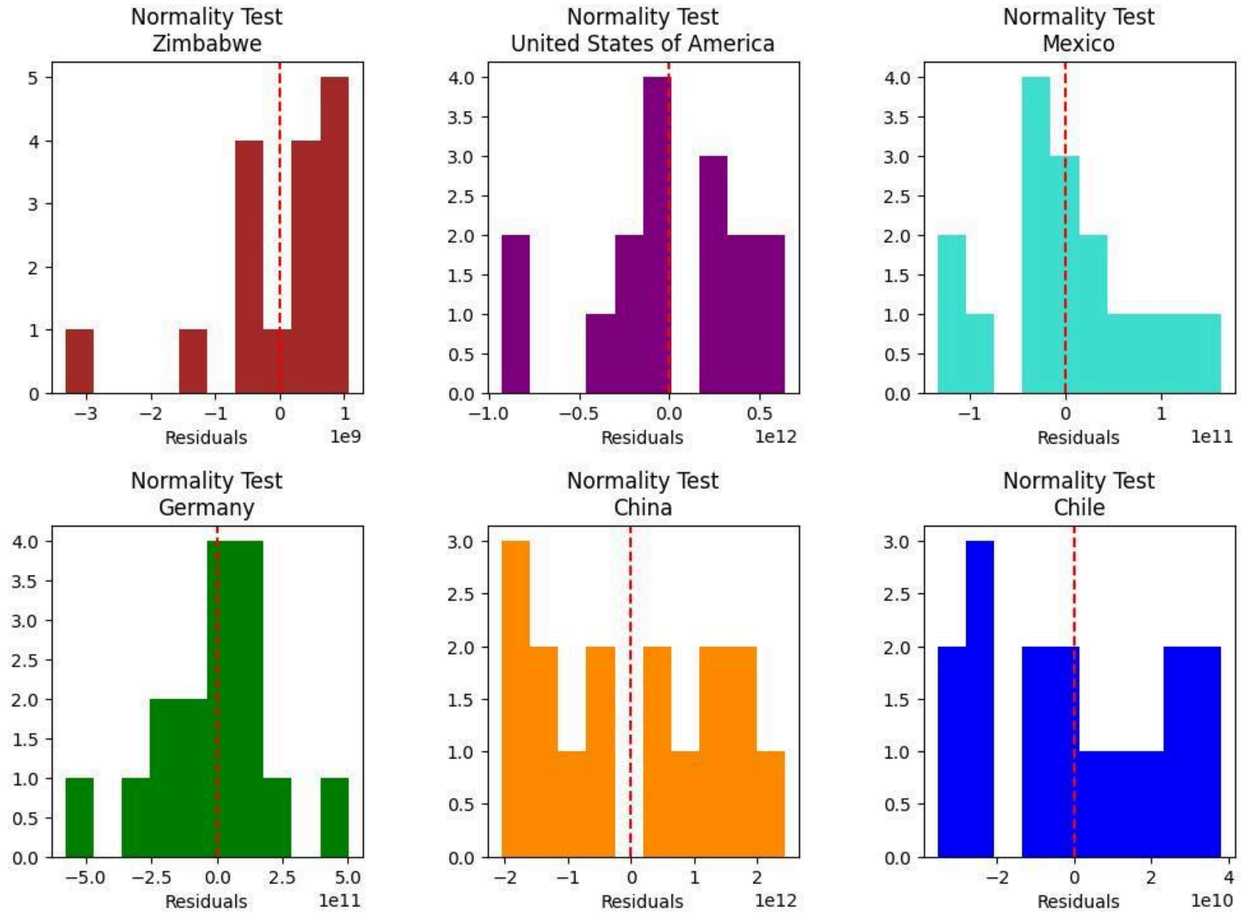
As can be observed from the plots shown above, the GDP of each country shares a strong positive correlation with their respective country's life expectancy. This means that as GDP of a particular country increases, it appears that the average life expectancy for that country also increases. However, there is still a bit left to do in order to confirm this hypothesis, starting with plotting a best-fit line over each plot to better picture how well a linear relationship can match the trends shown.

Normality and Homoscedasticity Tests per Country

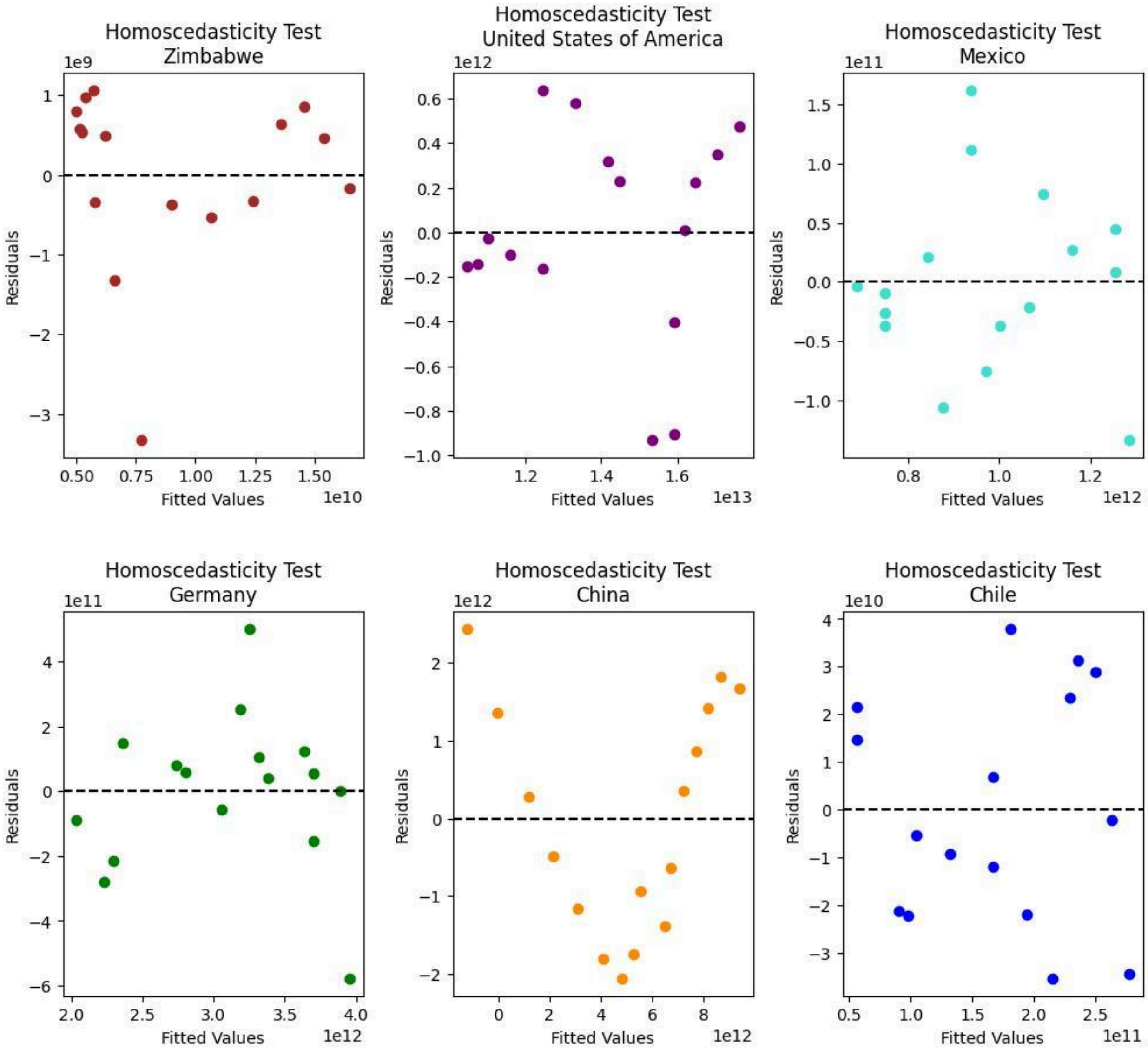
Before linear regression can be conducted on this data, however, normality and homoscedasticity tests must be completed. The normality and homoscedasticity tests can be described as follows:

The normality test checks to see if the normality assumption is met, which states that the residuals of any independent dataset will approach a normal distribution when the dataset is large enough. If this assumption is not met then it means one of two things: either the dataset isn't big enough or the data depends on some third, unknown variable - both of which can lead to biased results.

The homoscedasticity test checks to see if the homoscedasticity assumption is met, which states that the residuals should have equal variation across all values of the predictor, or independent, variable. If this assumption is not met then it means that there is a changing variation in the size of the error term across the independent variable. This can lead to biased results, as linear regression seeks to minimize residuals and gives all observations an equal amount of weight.



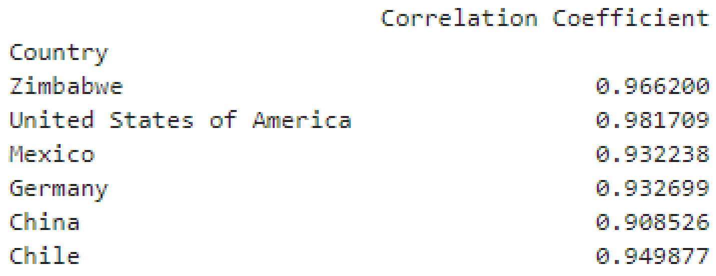
By looking at the graphs above, one can see that most of these plots indicate a normal distribution, but the plot for Zimbabwe appears to be slightly skewed to the left. However, although Zimbabwe appears to be slightly skewed, one can tell that the bulk of the data for this country centers at the graph's origin while only a couple data points are lying far to the left of that point. This seems to hint at the fact that the data suffers from a couple of outliers rather than being "skewed". Therefore, it appears that each plot passes the Normality test, meaning there is sufficient enough data to conduct linear regression on them. Now it's time to check the homoscedasticity assumption.



Based on the plots above, the plot points for each distribution are centered around y=0, meaning that the data for each country meets the homoscedasticity assumption. Thus, each plot has passed both the normality and homoscedasticity tests and it is safe to conduct linear regression.

Correlation Coefficient per Country

Lastly, the correlation coefficient is found for each country to ensure that each plot portrays a strong linear relationship and linear regression will produce meaningful results.

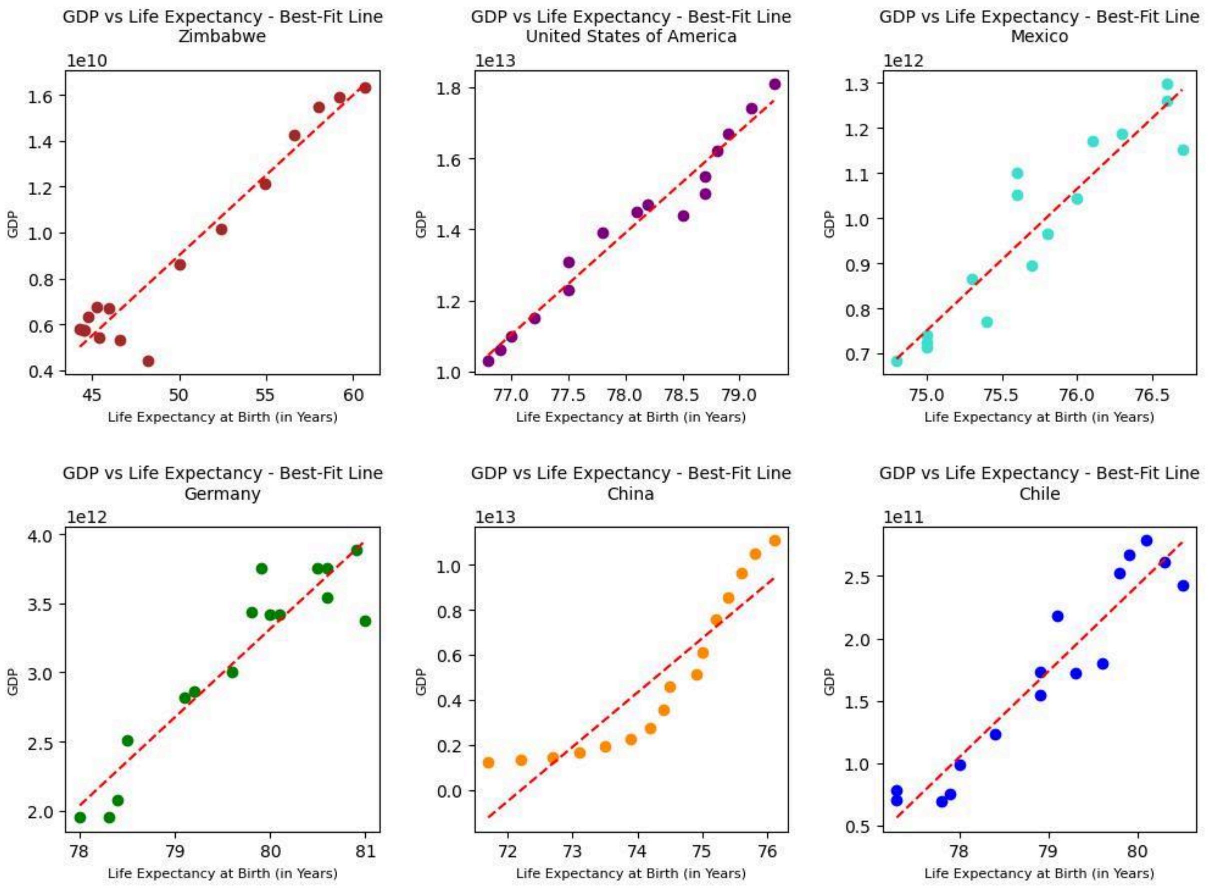


```
                              Correlation Coefficient
Country
Zimbabwe                                     0.966200
United States of America                     0.981709
Mexico                                       0.932238
Germany                                      0.932699
China                                        0.908526
Chile                                        0.949877
```

And, just as predicted, each plot has a correlation coefficient close to 1.0, meaning each portrays a strong linear relationship and linear regression will produce valuable results. More than that, however, this seems to show that there is a strong positive linear relationship between the GDP and life expectancy of each country. To visually demonstrate this, a best-fit line will be produced using linear regression, and plotted over each set of data points.

Best-Fit Line per Country

To portray the positive linear relationship between the GDP and life expectancy of each country, linear regression is conducted to generate a best-fit line, which is then fitted over it's respective plot. This can be seen below.



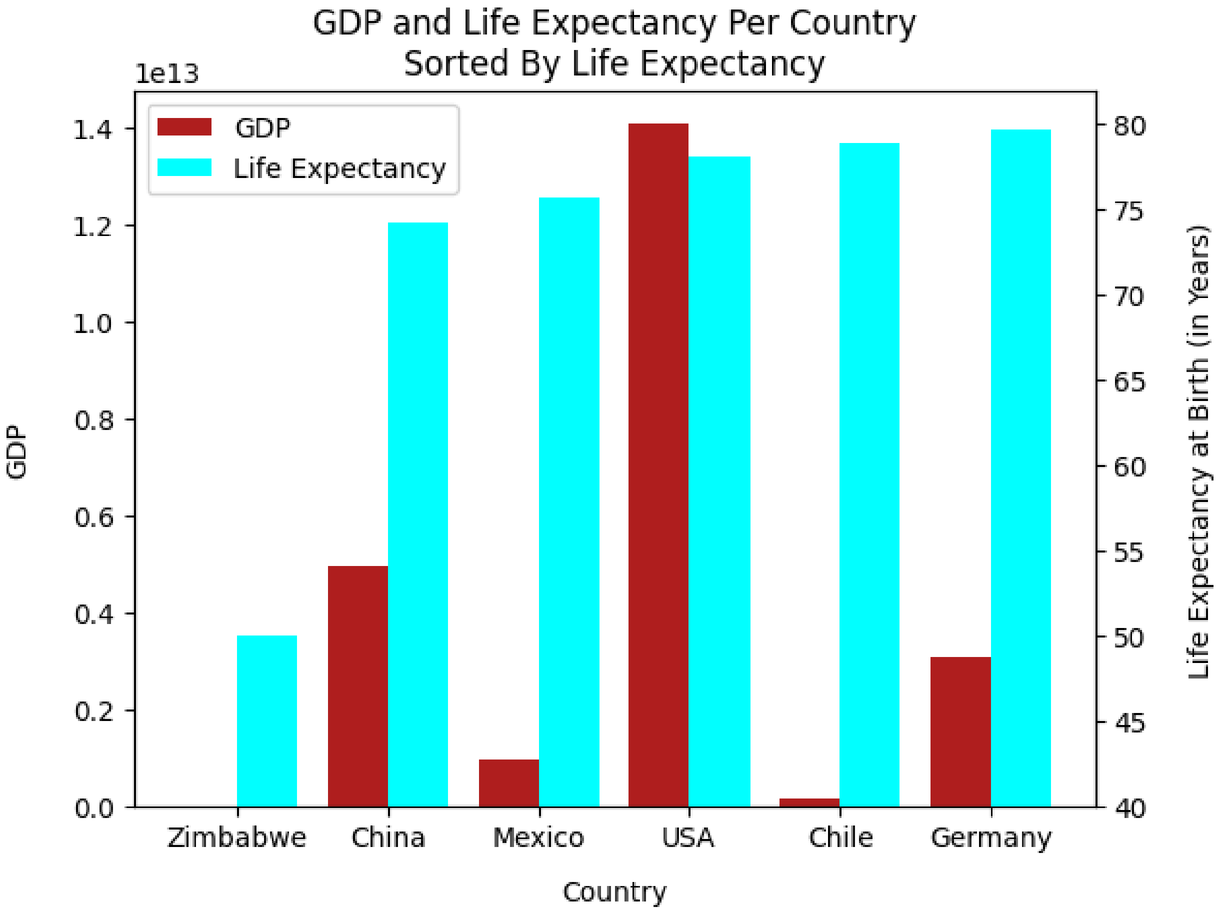
As can be seen from the plots above, each country has a strong linear relationship between it's respective GDP and life expectancy. This means that the life expectancy of each country increases as that country's GDP increases, and so the two are shown to be strongly correlated.

Average GDP and Life Expectancy per Country

Earlier, we were attempting to draw conclusions from the 'GDP vs Life Expectancy at Birth' scatter plot and answer the following question:

- If one country has a higher GDP than another country, is that country also likely to have a longer life expectancy?

But were unable to due to the plot being overcrowded. Thus, a side-by-side bar graph of each of the countries' average GDP and life expectancy is produced below to try and get a better grasp on this information and draw useful conclusions.



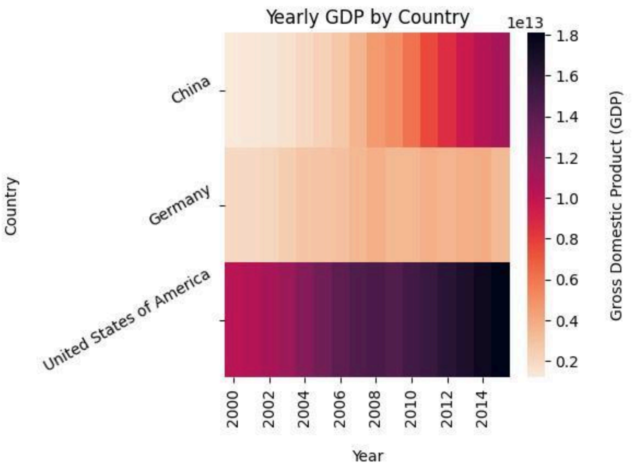
This plot illustrates quite a dissocation between the GDP and life expectancy between countries, as the countries with the highest life expectancy are not necessarily the same countries that have the highest GDP and vice versa. This, in turn, means that just because a country has a higher life expectancy than another country doesn't mean that that country has a higher GDP.

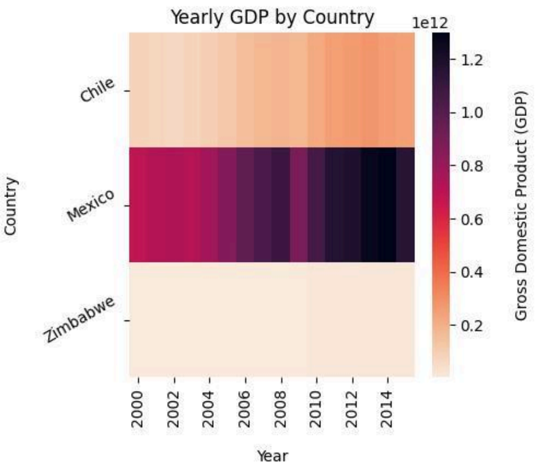
Yearly GDP by Country - Heatmap

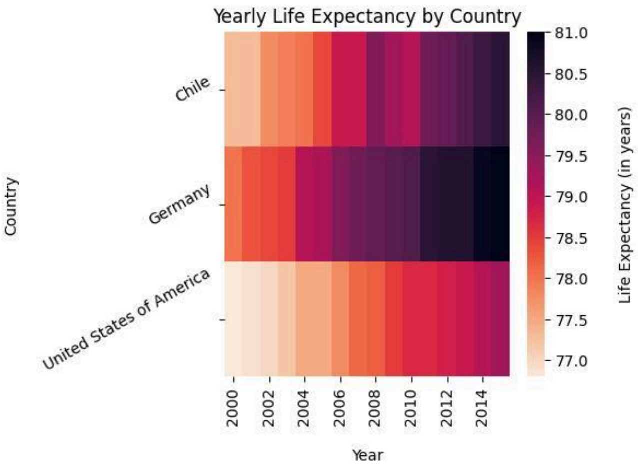
Finally, heatmaps of each country's GDP and life expectancy were generated in an attempt to answer the following questions:

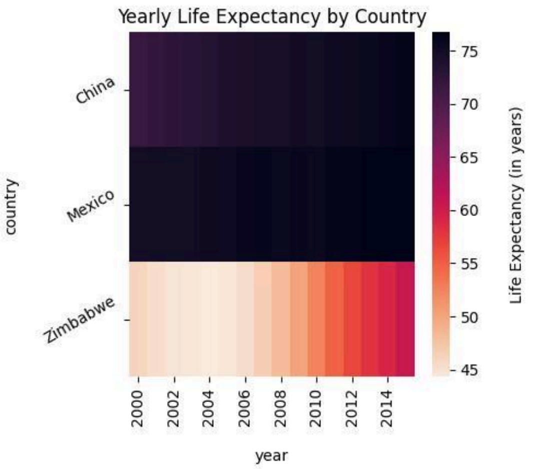
- Has life expectancy increased over time in the six nations?
- Has GDP increased over time in the six nations?

These heatmaps can be seen below.









As one can see, there are, unfortunately, a few of these heatmaps that are hard to make sense of, as they don't seem to show any transition in color. This is mainly because of the imbalance between some of these countries, such as Zimbabwe having a much lower GDP and life expectancy than the other countries. However, although a few may be hard to read, most show a clear positive trend of an increasing GDP and life expectancy over the years. Thus, one can come to the conclusion that the answer to this question is a resounding yes, the GDP and life expectancy of these various countries has increased over time.

Conclusion

The analysis aimed to investigate the relationship between GDP and life expectancy across six nations over a 15-year period. Through data visualization and statistical analysis, several key findings emerged.

Firstly, scatter plots revealed a general positive trend between GDP and life expectancy, suggesting that as GDP increases, so does life expectancy. However, the correlation between GDP and life expectancy varied among countries.

Distribution plots illustrated the diversity in life expectancy and GDP distributions across the six nations. While some countries exhibited relatively normal distributions, others showed skewed distributions, indicating a wide range of life expectancies and GDP levels.

Linear regression analysis confirmed a strong positive linear relationship between GDP and life expectancy for each country. The correlation coefficients approached 1.0, indicating a significant association between the two variables.

Furthermore, a comparison of average GDP and life expectancy across countries revealed a dissociation, as countries with higher life expectancies did not necessarily have higher GDPs, and vice versa.

Heatmaps displayed an overall increasing trend in GDP and life expectancy over time across the six nations. Despite some challenges in interpretation due to imbalances between countries, the majority of the heatmaps indicated positive transitions over the years.

In conclusion, the analysis suggests that while GDP positively influences life expectancy, the relationship is nuanced and varies among nations. The findings highlight the importance of considering multiple factors beyond GDP alone in understanding life expectancy trends. Overall, the study contributes valuable insights into the complex interplay between economic development and public health outcomes.