Class project: Analysis of GDP/GDP-Per-Capita and its relation to Life Expectancy

Introduction

Understanding the relationship between a nation's economic output and the health of its population is crucial for policymakers, economists, and public health experts. This project examines how Gross Domestic Product (GDP) correlates with life expectancy across 189 countries, building upon the well-known Preston Curve, which suggests that life expectancy rises with income but at a diminishing rate. While prior research has established this general trend, my analysis delves deeper by investigating regional variations, identifying outliers, and quantifying the strength of this relationship at different income levels and regions of the world.

The data for this study comes from four primary sources:

-The World Bank's GDP records (2019 nominal values): https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

-The World Bank's life expectancy estimates: https://data.worldbank.org/indicator/SP.DYN.LE00.IN

-The World Bank's population records: https://data.worldbank.org/indicator/SP.POP.TOTL

-Geographic boundaries were sourced from Natural Earth (embedded in code) to enable regional comparisons.

After cleaning the data (removing countries with missing values and log-transforming GDP to account for its skewed distribution) I conducted exploratory visualizations to determine correlation and to model how this correlation changes as GDP and GDP-Per-Capita increases. My goal is not only to confirm the GDP-life expectancy link but also to highlight nations that deviate from expectations, as well as shedding light on regional disparities, prompting further investigation into the policies and social factors that drive these inequalities.

Data wrangling: Merging GDP, Life Expectancy, and Geographic Data while removing Blank Values

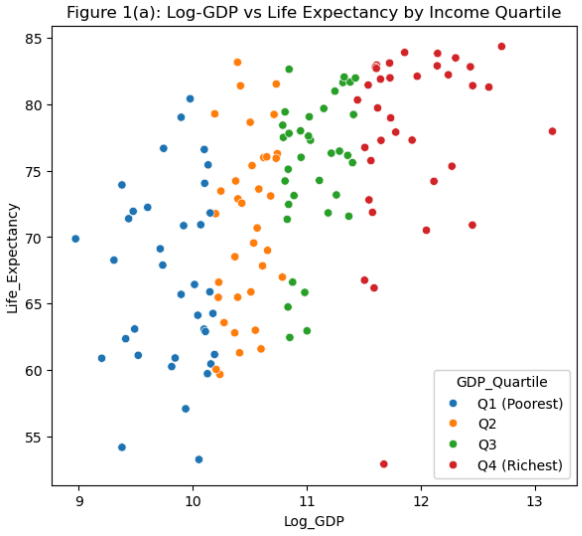
- Merging GDP, life expectancy, and geographic data
- Removing blank values
- Scaling and slicing GDP Data so that it is better visually depicted in visualizations

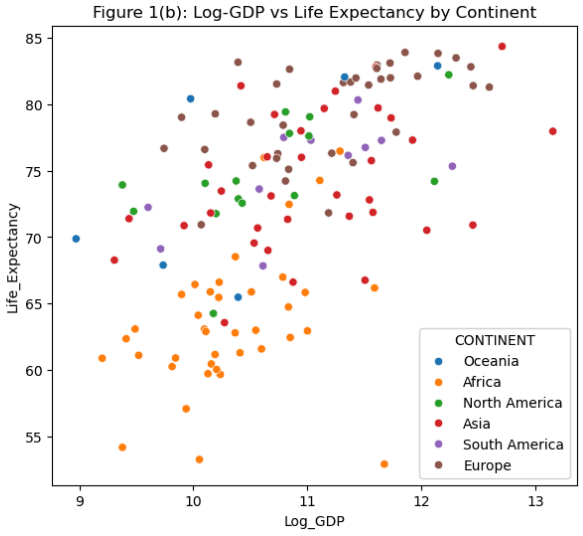
In [1]:
```
import geopandas as gpd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
```

In [2]:
```
# Load and clean GDP data (World Bank)
gdp = pd.read_csv('gdp_data.csv', skiprows=4)[['Country Name','2019']].renam

# Load and clean life expectancy data (WHO)
life = pd.read_csv('life_expectancy.csv', skiprows=4)[['Country Name','2019'

# Merge datasets and remove missing values
merged = gdp.merge(life, on='Country Name').dropna()

world = gpd.read_file(
    "https://naciscdn.org/naturalearth/110m/cultural/ne_110m_admin_0_countri
)

analysis_data = world.merge(merged, left_on='NAME', right_on='Country Name')

# Create analysis variables
analysis_data['Log_GDP'] = np.log10(analysis_data['GDP'])
analysis_data['GDP_Quartile'] = pd.qcut(analysis_data['Log_GDP'], 4, labels=
```

Visualize the data: Showcasing Global Disparities

In [3]:
```
# Plot scatterplots showcasing relationship between GDP and Life Expectancy
fig, ax = plt.subplots(1,2, figsize=(15,6))
sns.scatterplot(x='Log_GDP', y='Life_Expectancy', hue='GDP_Quartile', data=a
sns.scatterplot(x='Log_GDP', y='Life_Expectancy', hue='CONTINENT', data=anal
ax[0].set_title('Figure 1(a): Log-GDP vs Life Expectancy by Income Quartile'
ax[1].set_title('Figure 1(b): Log-GDP vs Life Expectancy by Continent')
plt.show()
```



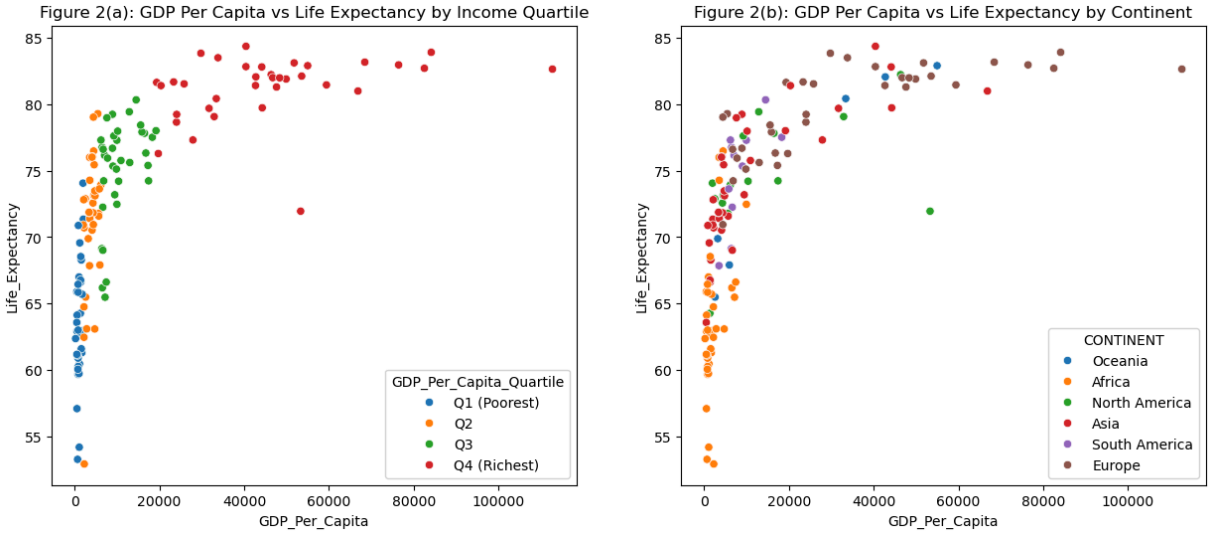


The scatter plot of GDP (log-transformed) against life expectancy demonstrates a clear positive correlation, particularly pronounced among low- and middle-income countries. Figure 1(a) indicates that as Log_GDP increases, life expectancy rises steeply by approximately 15 years on average, highlighting how economic development in poorer nations delivers dramatic improvements in population health.

However, this trend flattens significantly among high-GDP countries, where additional wealth yields diminishing returns. For instance, while the United States has the highest GDP globally, its life expectancy (around 79 years) lags behind nations like Japan (around 85 years) and Germany (around 81 years), suggesting that factors like healthcare access and income inequality may outweigh pure economic output at higher development levels. The regional clustering in Figure 1(b) further underscores these disparities: For instance, African nations predominantly occupy the plot's lower-left quadrant, whereas European and North American countries dominate the upper-right.

However there are striking outliers, such as Nigeria (the red point on the bottom right of Figure 1(a) and the yellow point on the bottom right of Figure 1(b)), that surpass expectations for its GDP, challenging assumptions.

In [4]:
```
population = pd.read_csv('population_data.csv', skiprows=4)[['Country Name',
new_analysis_data = analysis_data.merge(population, left_on='Country Name',
new_analysis_data['GDP_Per_Capita'] = new_analysis_data['GDP'] / new_analysi
new_analysis_data['GDP_Per_Capita_Quartile'] = pd.qcut(new_analysis_data['GD

fig, ax = plt.subplots(1,2, figsize=(15,6))
sns.scatterplot(x='GDP_Per_Capita', y='Life_Expectancy', hue='GDP_Per_Capita
sns.scatterplot(x='GDP_Per_Capita', y='Life_Expectancy', hue='CONTINENT', da
ax[0].set_title('Figure 2(a): GDP Per Capita vs Life Expectancy by Income Qu
ax[1].set_title('Figure 2(b): GDP Per Capita vs Life Expectancy by Continent

plt.show()
```



Total GDP can mask critical disparities. For instance, a nation with high aggregate wealth may still have widespread poverty if resources are concentrated. GDP per capita better

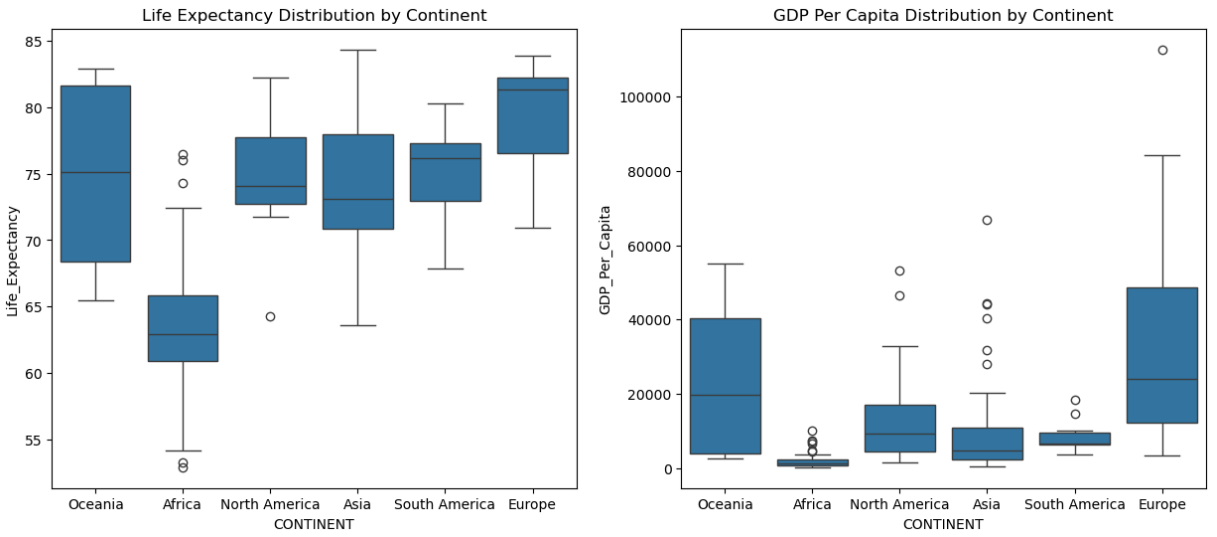
reflects individual economic capacity, therby removing the added population variable, making it a more effective metric for life-expectancy analysis.

Figure 2(a)'s depiction of the relationship between GDP per capita and life expectancy reveals a striking pattern: at lower income levels, even small increases in per capita wealth correlate with dramatic improvements in lifespan, as access to basic necessities like clean water, nutrition, and primary healthcare transforms survival rates. However, this curve flattens significantly at higher income tiers—wealthy nations cluster near a biological ceiling where additional economic gains yield minimal longevity returns. This plateau suggests that while poverty reduction remains crucial for improving health in developing nations, high-income countries must focus on equitable healthcare access and preventive medicine rather than relying on GDP growth alone.

Looking on a geographic scale, Figure 2(b) clearly demonstrates persistent global disparities, with African nations overwhelmingly clustered at the lower end of the curve, where low GDP per capita coincides with life expectancies far below what is likely the global average. This reflects systemic challenges like underfunded healthcare infrastructure, infectious disease burdens, and resource limitations that continue to constrain health outcomes across the continent. In stark contrast, European countries dominate the upper right of the plot, their high incomes and robust social systems delivering consistently strong longevity figures.

To understand where these issues are likely rooted, we would need to look at the various regions (continents) more closely as opposed to at a global scale.

In [5]:
```
fig, ax = plt.subplots(1,2, figsize=(15,6))
sns.boxplot(x='CONTINENT', y='Life_Expectancy', data=new_analysis_data, ax=a
ax[0].set_title('Life Expectancy Distribution by Continent')
sns.boxplot(x='CONTINENT', y='GDP_Per_Capita', data=new_analysis_data, ax=ax
ax[1].set_title('GDP Per Capita Distribution by Continent')
plt.show()
```



To understand why this discrepancy exists, we must look into each continent or region separately.

In Figure 3(b), African nations' narrow spread in GDP per capita—clustered at the lower end of the spectrum—belies a vast range in life expectancies in Figure 3(a), suggesting that factors beyond raw economic output, such as healthcare infrastructure, disease burden, and governance quality, play outsized roles in determining health outcomes. However, certain African countries achieve life expectancies approaching global averages despite modest GDP per capita as seen with Africa's many outliers in Figure 3(a), likely through targeted public health investments, while others lag severely due to conflicts or underfunded systems.

In contrast, Europe's uniformly high life expectancies in Figure 3(a), despite a wide dispersion in GDP per capita (Figure 3(b)), suggest governmental intervention and underscore how robust social safety nets and universal healthcare systems can decouple health outcomes from pure economic metrics.

This contrast highlights a critical insight: while wealth may set the baseline for health potential, its realization depends on how effectively governments translate economic resources into equitable care—a lesson embodied by Europe's consistency and Africa's disparities.

Analyses: Piecewise Regression Reveals GDP's Diminishing Health Returns

From the above analyses, it can be deduced that the GDP of a country is heavily correlated with their respective life expectancy, but to what extent do they correlate to each other?

In [6]:
```
# Split data into low/high GDP groups
threshold = 1e11

low_gdp = analysis_data[analysis_data['GDP'] < threshold]
high_gdp = analysis_data[analysis_data['GDP'] >= threshold]

# Fit models
models = {
    'Low GDP': LinearRegression().fit(low_gdp[['Log_GDP']], low_gdp['Life_Ex
    'High GDP': LinearRegression().fit(high_gdp[['Log_GDP']], high_gdp['Life
}

print(f"Low GDP slope: {models['Low GDP'].coef_[0]:.1f} years per log-GDP-ur
print(f"High GDP slope: {models['High GDP'].coef_[0]:.1f} years per log-GDP-
```

```
Low GDP slope: 5.3 years per log-GDP-unit
High GDP slope: 2.9 years per log-GDP-unit
```

The piece-wise regression analysis above suggests that countries that have lower GDPs would likely benefit greater in terms of life expectancy with each successive increase in GDP.

The steep slope for low-GDP countries (5.3 years gained per log-GDP increase) reveals a powerful opportunity: strategic healthcare investments in developing nations can create a virtuous cycle where improved health drives economic growth, which in turn fuels further health gains. When low-GDP prioritize vaccinations, maternal care, and infectious disease control, it is likely they will reap immediate life expectancy boosts while simultaneously building a healthier workforce—one that misses fewer workdays, attracts foreign investment, and achieves higher productivity. For low-income nations, healthcare doesn't seem to just be a social expense, it is likely one of the most efficient catalysts for both economic development and population longevity.

In [7]:
```
# Split data into low/high GDP groups
analysis_data['low_gdp'] = analysis_data['GDP'] < threshold

# Add summary statistics
gdp_groups = analysis_data.groupby('low_gdp').agg(
    Country_Count=('Country Name', 'count'),
    Avg_Life_Expectancy=('Life_Expectancy', 'mean'),
    Median_GDP=('GDP', 'median')
).round(1)

print("\n=== Group Characteristics ===")
print(gdp_groups)

# Show example countries with their data
sample_countries = analysis_data[['Country Name', 'GDP', 'Life_Expectancy',
print("\n=== Low GDP Country Data ===")
print(sample_countries.sort_values('GDP', ascending=True).head(10))
```

```
=== Group Characteristics ===
         Country_Count  Avg_Life_Expectancy    Median_GDP
low_gdp
False               54                 77.6  3.944905e+11
True                87                 69.7  1.879944e+10

=== Low GDP Country Data ===
      Country Name           GDP  Life_Expectancy  low_gdp
73         Vanuatu  9.365263e+08         69.87700     True
52   Guinea-Bissau  1.588639e+09         60.88200     True
16     Timor-Leste  2.032550e+09         68.26800     True
31          Belize  2.388300e+09         73.93100     True
18         Lesotho  2.390702e+09         54.17300     True
61         Burundi  2.576519e+09         62.35100     True
80          Bhutan  2.735684e+09         71.39100     True
15       Greenland  2.997310e+09         71.94439     True
133       Djibouti  3.088854e+09         63.08500     True
53         Liberia  3.319596e+09         61.10400     True
```

This table of low-GDP countries with lower-than-average life expectancies serves as both a diagnostic tool and a priority list for global health efforts. By quantifying the exact gap between these nations and their higher-performing economic peers (like the 5.3-year/2.9-year life expectancy gain per GDP unit identified earlier), we are able to transform abstract disparities into concrete intervention targets. The clustering of low-gdp nations from all over the world in this table highlights that while this issue is more prevalent in certain regions, it is still an issue that is felt (even in pockets) all across the world. This information can be used to strategically allocate health funding (e.g., malaria prevention vs. broad infrastructure) to those low-gdp nations that require it. This can dramatically alter both a region's economic and health trajectories. For policymakers, this isn't just a ranking; it's a roadmap showing where targeted health investments could deliver the highest marginal returns in both economic and human terms.

Conclusions

The findings demonstrate that GDP growth has its greatest impact on life expectancy in low-income countries, where each incremental increase in wealth delivers substantial health improvements—a clear indication that escaping poverty is the most powerful health intervention. However, as nations develop economically, the health returns diminish significantly, revealing that factors like healthcare quality, education, and social equality become increasingly important. The outliers—countries that either overperform or underperform their GDP predictions—highlight how policy choices can bend these economic constraints, proving that targeted investments in public health and equitable access to care can amplify or undermine the benefits of economic growth.

Moving forward, this research could be expanded in several meaningful directions. A longitudinal study could track how specific health policies, such as universal vaccination programs or primary care expansion, have altered the GDP-life expectancy relationship over time in different regions. Additionally, incorporating inequality metrics would provide a more nuanced understanding of how wealth distribution—not just national averages—shapes population health outcomes. Finally, comparative cost-effectiveness analyses could identify which health interventions deliver the greatest longevity gains per dollar spent in low-GDP settings, helping policymakers prioritize limited resources. Together, these insights underscore that while economic growth lays the foundation for health, its full potential is only realized through deliberate, equitable policy choices.

Reflection

This project successfully identified key global trends in the GDP-life expectancy relationship through a combination of visual and statistical analysis. The scatter plots and boxplots effectively illustrated both the overall correlation and regional disparities, while the piecewise regression provided quantifiable evidence of diminishing returns.

The geographic mapping added critical context, revealing clusters of underperformance in Africa and overperformance in Europe that might otherwise be overlooked in purely numerical analysis.

However, several challenges arose during execution. Merging the World Bank and WHO datasets required careful handling of country name discrepancies (e.g., “Côte d’Ivoire” vs. "Ivory Coast"), and some countries were dropped due to missing data, potentially biasing the results. Additionally, while I initially planned to include healthcare expenditure data, inconsistent reporting across countries forced us to abandon this variable. These limitations underscore the importance of data quality in drawing robust conclusions.

The project took approximately 10 hours to complete, including data collection, cleaning, analysis, and visualization. If given more time, I would explore machine learning techniques to predict life expectancy using multiple socioeconomic indicators or conduct a time-series analysis to track changes over the past 50 years. Despite its limitations, this study offers a clear foundation for understanding how economic and health outcomes intersect—and where policymakers might focus to bridge the gaps.