Factors Influencing Life Expectancy Using the Gapminder Dataset

Understanding factors that influence life expectancy is important for addressing global disparities. For this project, I used the 2007 Gapminder dataset, which contains global data on life expectancy, GDP per capita, and population.

I chose life expectancy as the response variable. Life expectancy, while a single number, reflects the overall socioeconomic health of a population. I selected GDP per capita (numeric), population size (numeric), and continent (categorical) to determine if they are good predictors of life expectancy.

Descriptive Statistics

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Life Expectancy	39.61	57.16	71.94	67.01	76.41	82.60
GDP per Capita	277.6	1,624.8	6,124.4	11,680.1	18,008.8	49,357.2
Population	199,600	4,508,000	10,520,000	44,021,220	31,210,000	1,319,000,000

Table 1. Summary Statistics from Gapminder Dataset

1. As shown in Table 1, Life expectancy ranges from 40 to 83 years. The median life expectancy is 72 years, while the mean is 67 years. The slightly lower mean is reflective of countries with very low life expectancy.

2. Table 1 shows that GDP per capita ranges from $278 to $49,357. The median is $6,124, while the mean is $11,680. A few wealthy nations inflate the mean global average GDP per capita, masking the true extent of global economic inequality.

3. According to summary statistics from Table 1, population ranges from 200,000 to 1.3 billion. The median population is 10.5 million, while the mean is 44 million. The disparity between the mean and median population is due to population dense countries like China and India.

Continent	Mean Life Expectancy (2007)
Africa	54.8 years
Americas	73.6 years
Asia	70.7 years
Europe	77.6 years
Oceania	80.7 years

Table 2. Average Life Expectancy by Continent from Gapminder Dataset

As shown in Table 2, there is a difference in life expectancy across continents, with Africa's life expectancy lagging behind all other continents. This suggests that further analysis is required to determine whether the differences in life expectancy between continents is statistically significant.

Exploratory Plots

I visualized the relationship between GDP per capita and life expectancy across continents using a scatter plot. As seen in Figure 1, there is a clear positive relationship between GDP per capita and life expectancy.

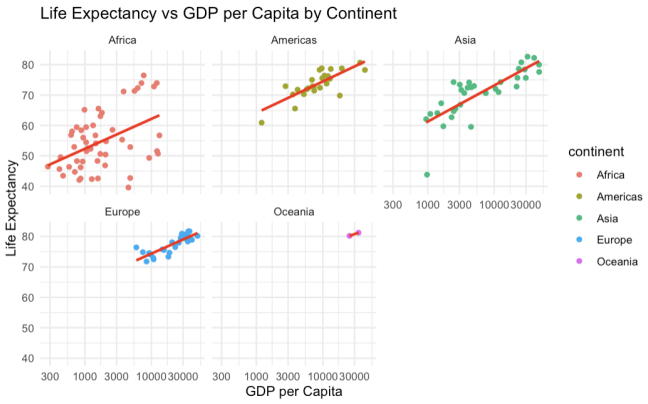

Figure 1 Life Expectancy vs GDP Per Capita by Continent from Gapminder Dataset

continent <fctr>	correlation <dbl>
Africa	0.3847152
Americas	0.5909661
Asia	0.6893590
Europe	0.8499711
Oceania	1.0000000

Table 3. Pearson's Correlation Between GDP per Capita and Life Expectancy by Continent

Pearson correlation analysis in Table 3 above shows that GDP per capita is positively correlated with life expectancy within each continent, ranging from moderate in Africa (r = 0.38) to strong in Europe (r = 0.85). This pattern is also visible in Figure 1, where the points for Africa are more scattered around the regression line, while the points for Europe are tightly clustered along the line, reflecting the stronger correlation. Oceania showed a perfect correlation (r = 1.0) due to Gapminder's small sample of just Australia and New Zealand, two highly developed nations with both high GDP and high life expectancy. This further reinforces GDP as a meaningful indicator of life expectancy.

The moderate positive correlation in Africa (r = 0.38) suggests that GDP per capita alone is not a good predictor of life expectancy in Africa. Factors like political stability and access to foreign aid help explain why life expectancy can vary widely among African countries with similar income levels. For example, Sudan and Ethiopia may have similar GDP per capita to other African countries, but their life expectancy may be much lower due to the effects of civil war, ongoing conflict, and border disputes rooted in the colonial era.

Multiple Linear Regression

```
Call:
lm(formula = lifeExp ~ gdpPercap + pop + continent, data = gapminder_recent)

Residuals:
     Min       1Q   Median       3Q      Max
-22.8199  -2.8905   0.1574   2.9046  20.0585

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.371e+01  9.356e-01  57.413  < 2e-16 ***
gdpPercap         3.479e-04  5.717e-05   6.086 1.13e-08 ***
pop               9.586e-10  3.926e-09   0.244  0.80747
continentAmericas 1.603e+01  1.671e+00   9.592  < 2e-16 ***
continentAsia     1.256e+01  1.621e+00   7.751 1.97e-12 ***
continentEurope   1.520e+01  1.966e+00   7.730 2.20e-12 ***
continentOceania  1.662e+01  4.993e+00   3.329  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.597 on 135 degrees of freedom
Multiple R-squared:  0.7141,    Adjusted R-squared:  0.7014
F-statistic:  56.2 on 6 and 135 DF,  p-value: < 2.2e-16
```

Figure 2 Multiple Linear Regression Model

The multiple linear regression model in Figure 2 broadly explains how GDP per capita, population size, and continent together help explain the differences in life expectancy across countries. The model explains 71% of the variation in life expectancy (Adjusted $R^2 = 0.7014$).

With a p-value less than 0.0001, there is sufficient evidence that the GDP per capita affects life expectancy. Even though the Pearson correlation coefficient between GDP per capita and life expectancy is only moderate in Africa (r = 0.38, Table 3), the regression model in Figure 2 shows that GDP per capita remains a statistically meaningful predictor of life expectancy.

With a p-value of 0.807, there is not sufficient evidence that population significantly affects life expectancy. This is because large countries can be rich or poor, and small countries can be rich

or poor. Life expectancy is a consequence of factors like healthcare access and infrastructure which are not determined simply by the number of people, rather the quality of governance. Removing population as a factor from the model would not make a significant difference.

The y-intercept of the model in Figure 2, 53.71 is the estimated life expectancy in Africa when GDP per capita and population are zero. While this is unrealistic, it is a meaningful point to extrapolate the impact of GDP per capita and continent on life expectancy. Furthermore, for every additional dollar of GDP per capita, life expectancy increases by ~0.00035 years (~0.13 days), holding other factors constant.

Analyzing the continent coefficients from the model:

- On average, countries in the Americas have ~16 years higher life expectancy than Africa, controlling for GDP and population ($p < 0.05$).
- On average, countries in Asia have ~ 12.6 years higher life expectancy than Africa, controlling for GDP and population ($p < 0.05$).
- On average, countries in Europe have ~ 15.20 years higher life expectancy than Africa, controlling for GDP and population ($p < 0.05$).
- On average, countries in Europe have ~ 16.60 years higher life expectancy than Africa, controlling for GDP and population ($p < 0.05$).

The multiple linear regression model in Figure 2 compares each continent to Africa, the reference group. Further analysis is worthwhile to explore the differences between the other continents themselves, not just their differences relative to Africa.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lifeExp ~ continent, data = gapminder_recent)

$continent
                     diff        lwr       upr     p adj
Americas-Africa  18.802082  13.827042 23.777121 0.0000000
Asia-Africa      15.922446  11.372842 20.472051 0.0000000
Europe-Africa    22.842562  18.155858 27.529265 0.0000000
Oceania-Africa   25.913462  11.183451 40.643472 0.0000305
Asia-Americas    -2.879635  -8.299767  2.540497 0.5844901
Europe-Americas   4.040480  -1.495233  9.576193 0.2630707
Oceania-Americas  7.111380  -7.910342 22.133102 0.6862848
Europe-Asia       6.920115   1.763372 12.076858 0.0027416
Oceania-Asia      9.991015  -4.895221 24.877251 0.3464970
Oceania-Europe    3.070900 -11.857807 17.999607 0.9793776
```

Figure 3 Tukey's Honestly Significant Difference (HSD) Between Continents

Reinforcing the results from the multiple linear regression model in Figure 2, the Tukey HSD test shows that life expectancy in all continents is statistically different from Africa ($p < 0.05$). Additionally, the test reveals a statistically significant difference in life expectancy between Europe and Asia. Europe includes many wealthy nations like Norway, Germany, France, and Switzerland, driving life expectancy above 75 years. Similarly, Asia is home to wealthy nations

like Japan and Singapore with very high life expectancy. However, Asia includes large, developing countries like India and Bangladesh where life expectancy is significantly lower.

Conclusion

Analysis of the 2007 Gapminder dataset explored factors influencing life expectancy across countries, using GDP per capita, population size, and continent as predictors. The analysis combined descriptive statistics, Pearson correlation analysis, multiple linear regression, and Tukey's HSD to determine meaningful predictors of life expectancy.

Results were largely in line with expectations. As anticipated, GDP per capita was a statistically significant predictor of life expectancy, although its strength varied by continent — moderate in Africa and strong in Europe. Continent also proved to be a powerful predictor, with Africa lagging behind other continents and Europe and Asia showing meaningful differences between each other. In contrast, population size was not a significant predictor, which was somewhat expected, as life expectancy is shaped more by how scarce resources are distributed and how accessible they are across the social ladder. While population size plays a role, government policies play a much larger role.

While the difference between Europe's and Asia's life expectancy was significant in 2007, this gap may have changed in recent years, as many developing countries in Asia have made substantial progress since the time this data was collected.

Further research using more recent data and incorporating additional predictors such as healthcare spending and education levels would be valuable. Since the current multiple linear regression model in Figure 2 explains 71% of the variation in life expectancy, integrating these additional factors could help account for the remaining 29% of unexplained variation.