EXPLORATORY DATA ANALYSIS MINI PROJECT 1

Authors: Vikrant Deshpande, Tanvi Kolhatkar, Saishree Godbole

INTRODUCTION

In this project, we want to understand the relationship between life expectancy and GDP per capita. We've used the gapminder R package containing life expectancy information of 142 countries between 1952 to 2007. To understand any underlying relationships, we will:

1. Check if the relationship can be fitted by a linear model and explain any differences.
2. Analyze the trends of life expectancy over time for individual continents and respective countries.
3. Check if there are any other factors affecting the life expectancy apart from the GDP.

LIFE EXPECTANCY AND GDP TREND IN 2007

We've skipped Oceania from this analysis as there are just 2 countries with great GDP-per-capita values and correspondingly good life-expectancies.

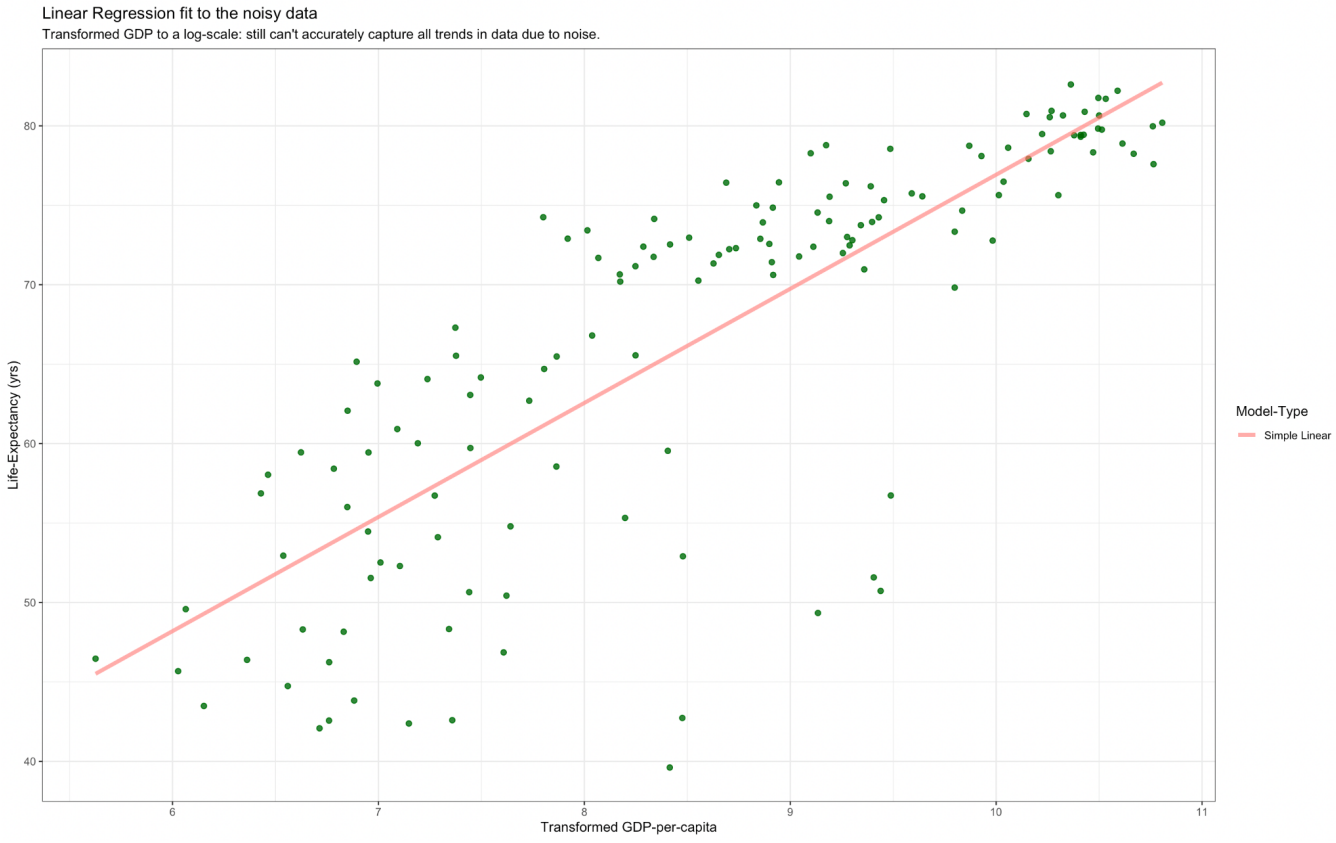

Fig.1

In Fig.1, we fit a linear regression model on the entire data. The linear model performs slightly better after a log transformation of GDP. We use the log-transformed data for further analysis.

** Refer Appendix Fig.A.1. to view the original LifeExp-vs-GDP plot, and residual-plot for this linear model.*

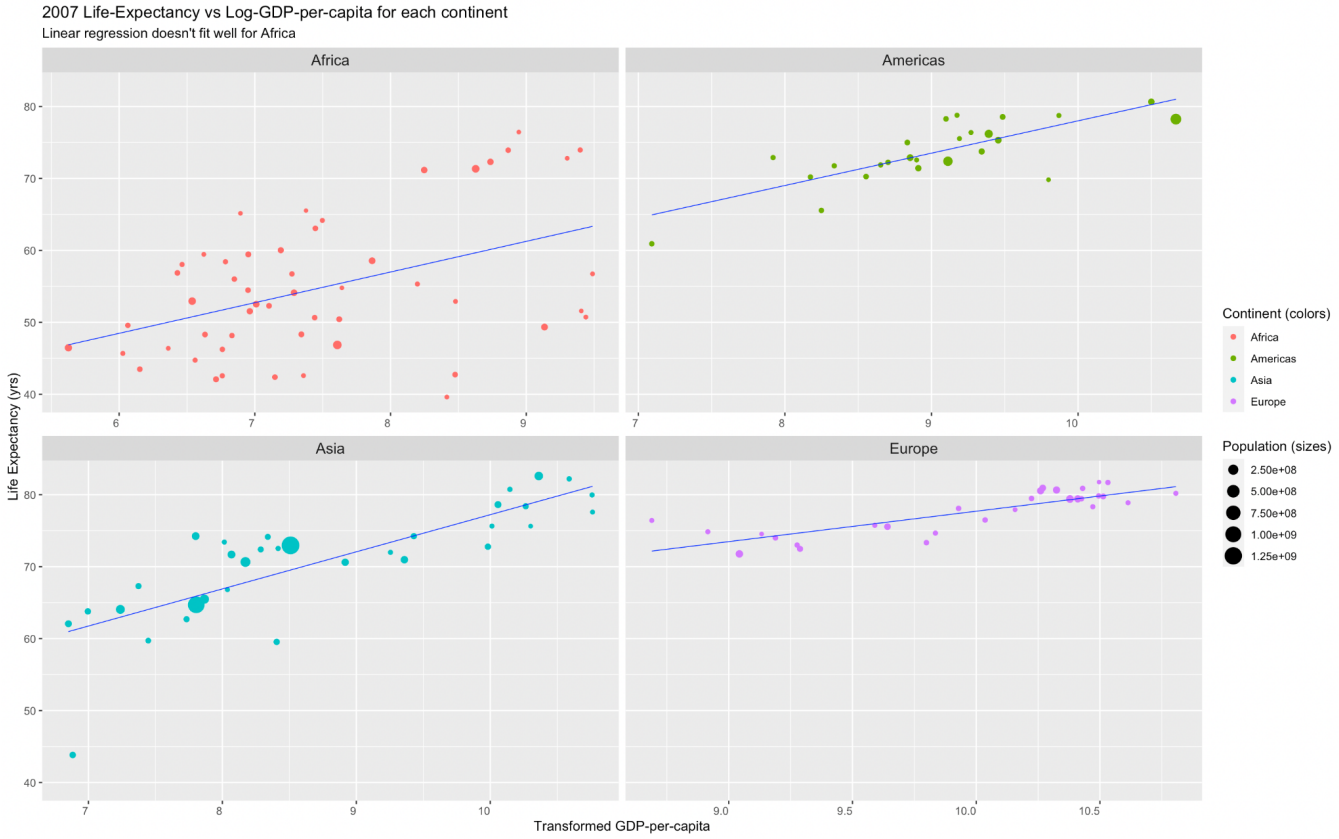

Fig.2

In Fig.2, we try to fit a linear regression model to each continent as opposed to the entire data.
Africa: Most have very low GDP and life expectancies ~40-50 years.
Americas, Asia: There are a few observable outliers for the Americas and Asia. But since we see a fairly linear relationship between GDP per capita and the Life-Expectancy, the predicted Life-Expectancy should increase as GDP-per-capita increases.
Europe: The linear model fits the best for this data. Affordable healthcare could attribute to the lack of a drastic change in life expectancies.

The linear model captures the trends for Asia, the Americas and Europe quite well, but performs poorly on the data for Africa. The differences between America and Europe can be explained by an additive shift.

LIFE EXPECTANCY OVER TIME

We now investigate the average life expectancy for all continents over the years 1950-2010. The average life expectancy of a continent for a specific year is obtained by weighing the countries' average life expectancy with its population.

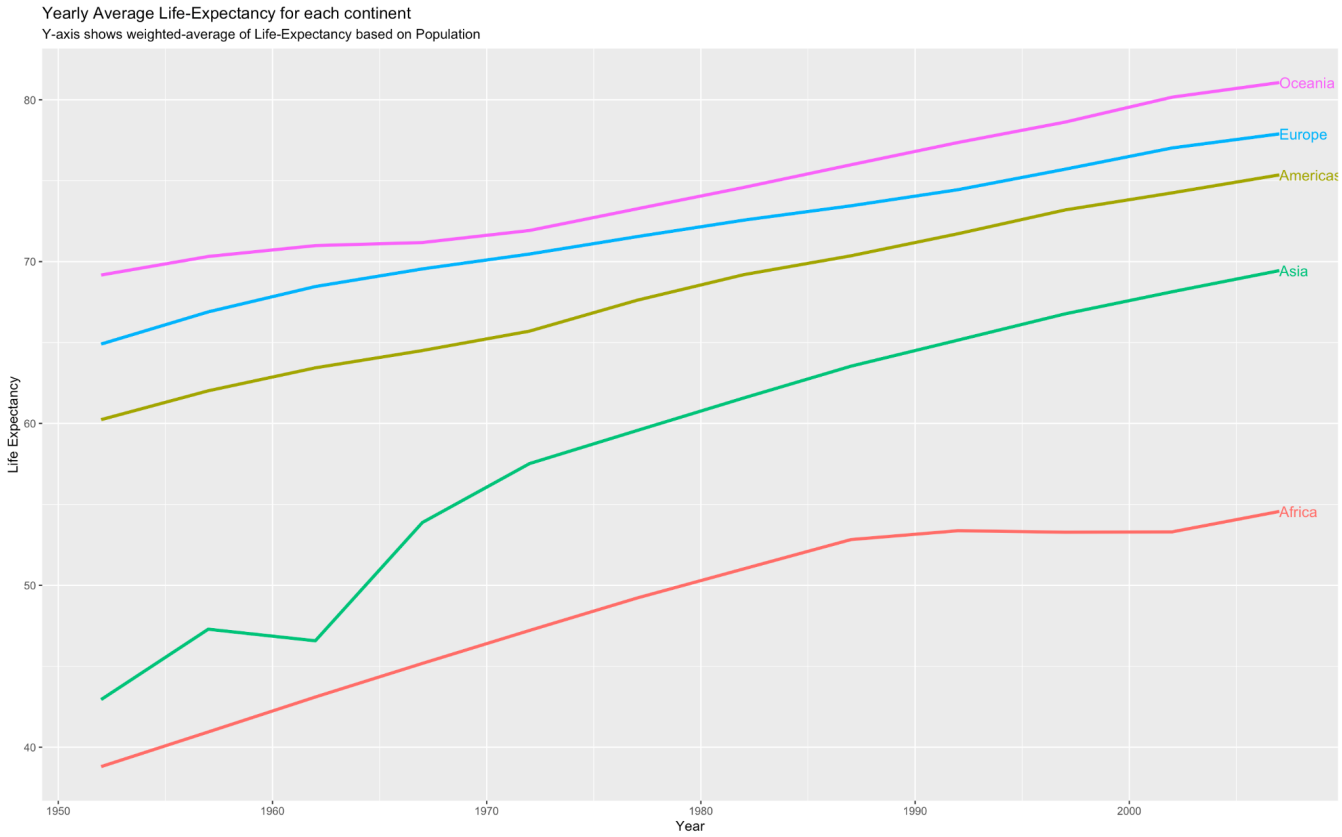

Fig.3

In Fig.3, we see the life expectancies in most continents linearly increasing with time. Continents of Europe and Americas started at high average life expectancies and have had a steady increase over the years. Asia and Africa were the two continents with very low life expectancy in 1950. Over time, notwithstanding a dip around 1962, Asia has caught up with the other continents at a fast rate (slope is slightly higher).
Average life expectancy in Africa has had a steady but relatively small increase from 30-50 to 40-60. It flattened slightly from 1990-2000.

** Refer to Appendix Fig.A.3. to view the breakdown of this plot for each continent at a country level.*

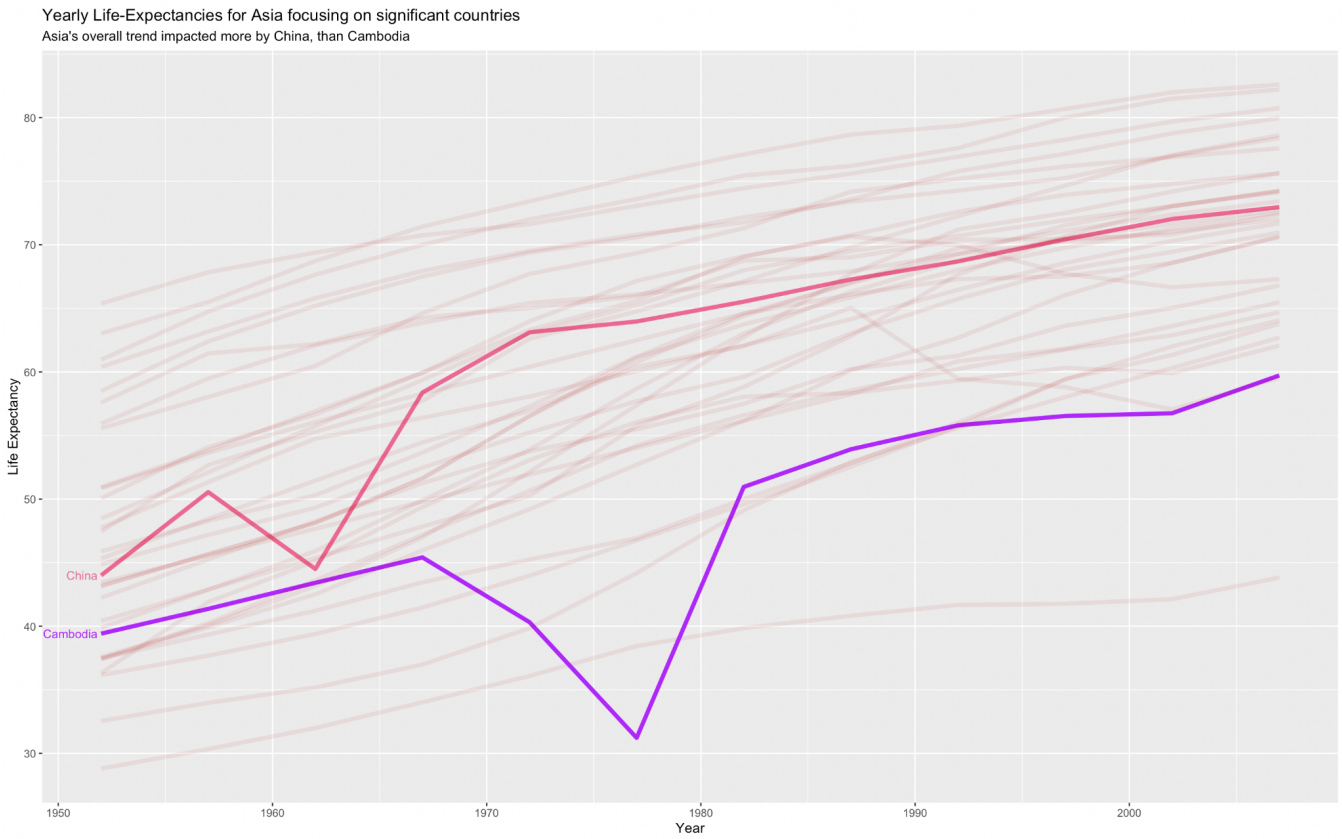

Fig.4

From Fig.4, we see a dip in average life expectancy (around 1962) for China. China has the largest population in Asia and contributes heavily to the weighted average life expectancy of the entire continent. These changes could be attributed to socio-economic changes and natural disasters such as famines occurring around 1962. Apart from this, the majority of Asian countries have had a steady growth in their life expectancy which has contributed to the overall growth of Asia (as in Fig.3). Cambodia on the other hand, although shows a greater dip in life-expectancy in 1975-79 (possibly due to the Cambodia-Vietnam wars), doesn't impact Asia's overall average life-expectancy in Fig.3.

CHANGES IN THE RELATIONSHIP BETWEEN GDP AND LIFE EXPECTANCY OVER TIME

Are the changes in life expectancy entirely explained by changes in GDP, or does time play a role here too?

Fig.5

In Fig.5, for each facet of the year, note that some points lie below the previous year's regression line, and this can be attributed to "regression-to-the-means".
Important Observations:

- Africa in 1952 had a regression line almost parallel to the X-axis: life-expectancy was low there in general, irrespective of GDP. As we move through time till 2007, we see the slope change to a slightly more positive outlook: higher GDP correlates to higher life expectancy.
- Europe and Asia had somewhat parallel regression lines in 1952, with positive slopes. Over time, we see these lines converge towards a fantastic life expectancy of approximately 80 years.
- Similarly, Europe and the Americas seem to have an ideal regression line with a small positive slope that alludes to a better Life-Expectancy for countries with higher GDP per capita.
- As we move from 1952 to 2007, the regression lines for Europe, Asia, and the Americas seemingly get merged into the same line (almost parallel to X-axis) indicative of the idealistic hypothesis. This might be proof that after 2010, such continents with developed nations, will have a regression-line with a small slope converging to a life expectancy of 80.

Thus, we can see that the changes in life expectancy cannot be entirely explained by changes in GDP, they are also dependent on time.

To view the changes over the years in greater detail for each continent, we have faceted the relationship between GDP per capita, life expectancy and time plot by each continent, in Fig.6.

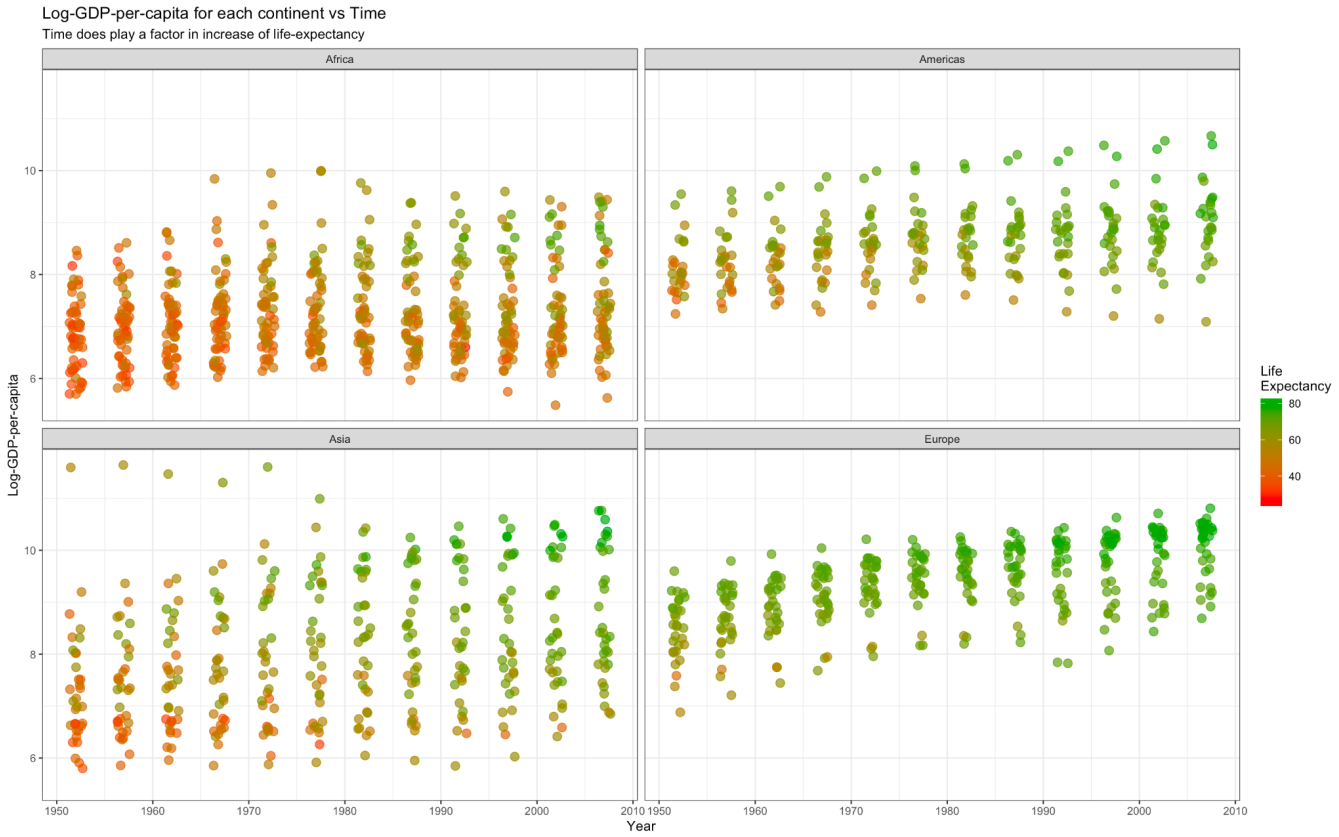

Fig.6

Focusing only on Africa in Fig.6, even though the average GDP of countries stays similar across time, the life expectancy just seems to get better, ie- the number of green circles increases from left-to-right within the same interval of Y-axis.

CONCLUSION

After analyzing the relationship between life expectancy and GDP over the years at a continent level, we can conclude that we can fit a linear model on the data after applying some transformations. Also, we observed that life expectancy in Asia has grown at a significantly fast rate and it has caught up with Americas, Europe and Oceania continents in 2007, whereas life expectancy in Africa has grown at a much slower rate. Individual countries in a continent contribute to an overall increase/decrease of life expectancy over the years. The changes in GDP per capita cannot account for all the changes in life expectancy. Finally, we observed a convergence of life expectancy to 80 years in the developed continents (Asia, Europe, Americas, Oceania).

APPENDIX

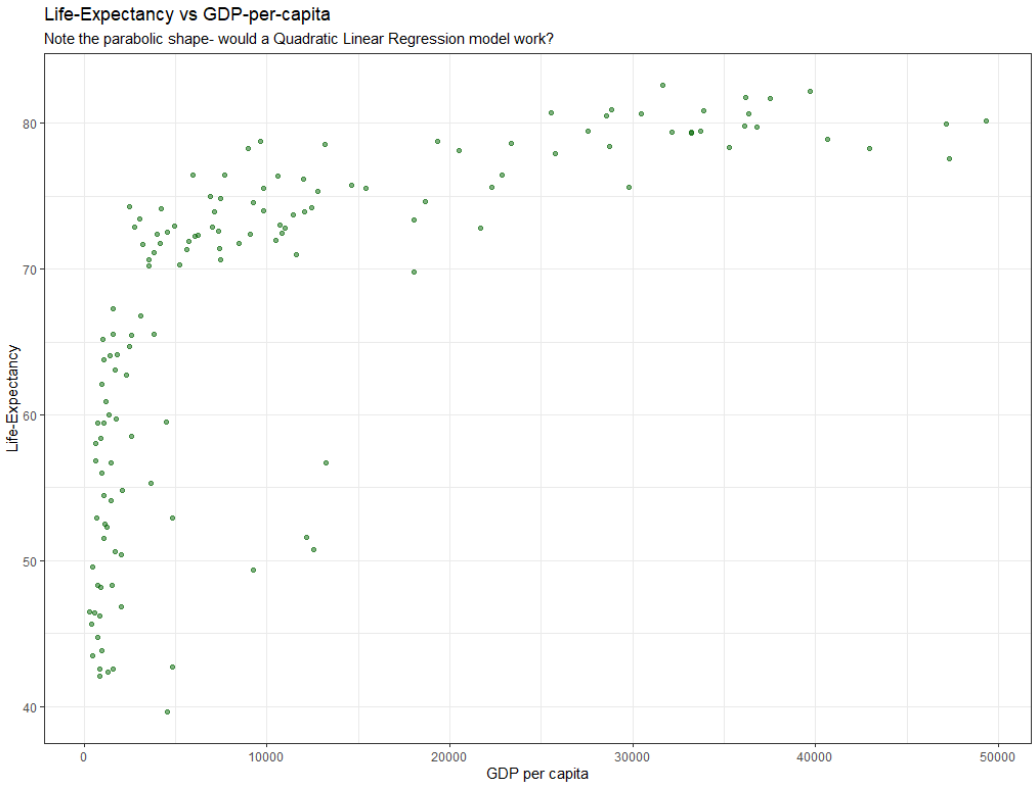

Fig.A.1.

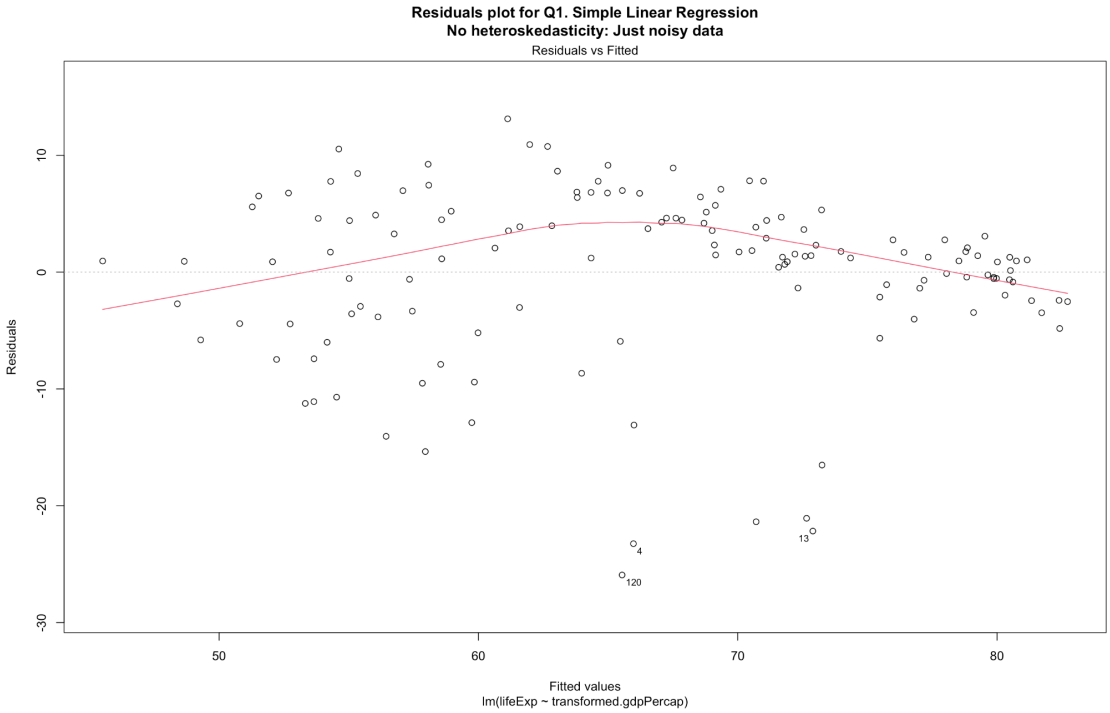

Fig.A.2.

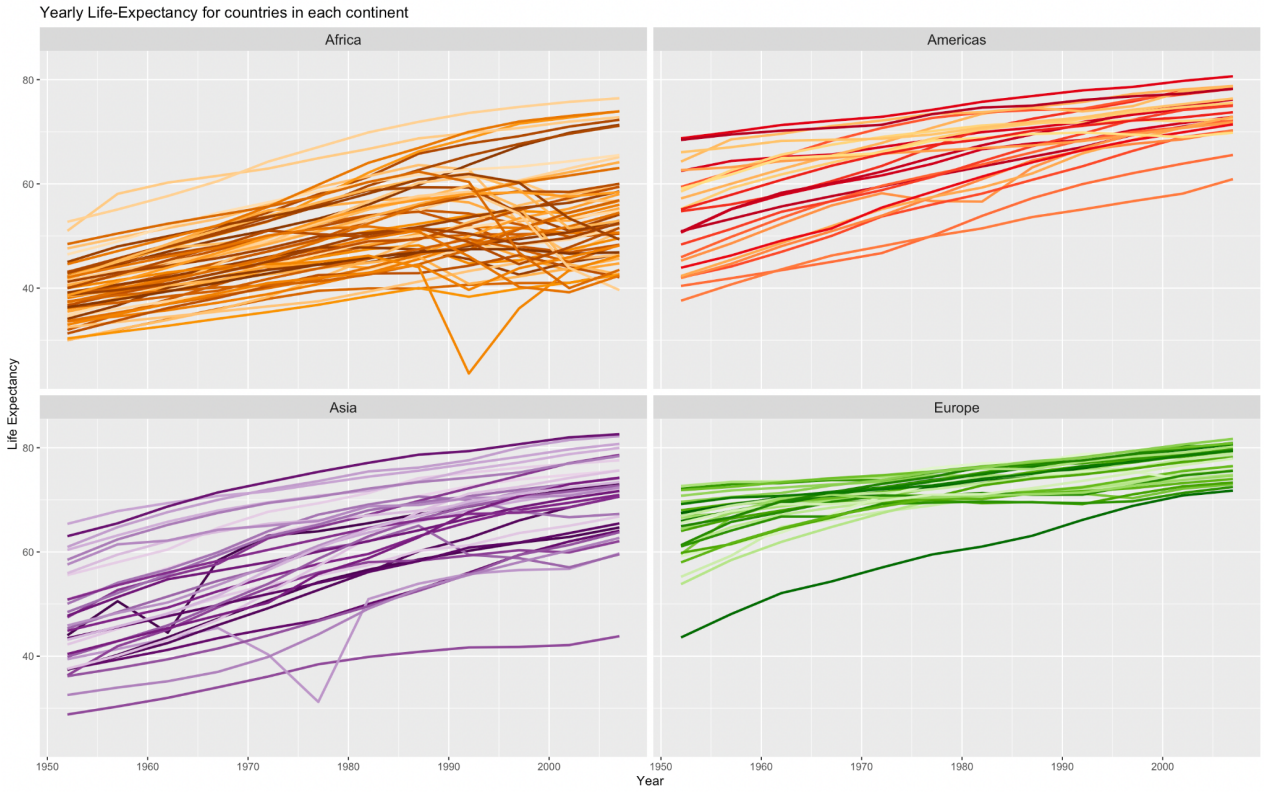
Yearly Life-Expectancy for countries in each continent
Africa
Americas
Asia
Europe
Life Expectancy
Year
80
60
40
1950
1960
1970
1980
1990
2000

Fig.A.3.