EXPLORATORY DATA ANALYSIS MINI PROJECT 1

Authors: Vikrant Deshpande, Tanvi Kolhatkar, Saishree Godbole

INTRODUCTION

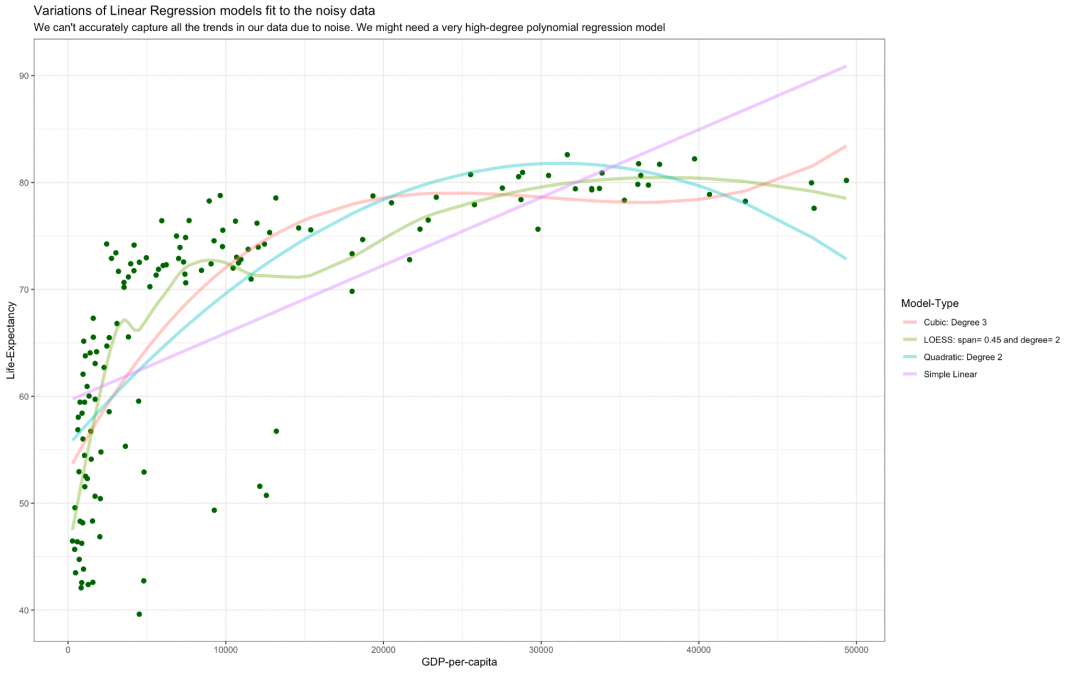
In this project, we want to understand the relationship between life expectancy and GDP per capita. We have used the gapminder R package having life expectancy information of 142 countries between 1952 to 2007. We do the following to understand the type of relationship between life expectancy and GDP:

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

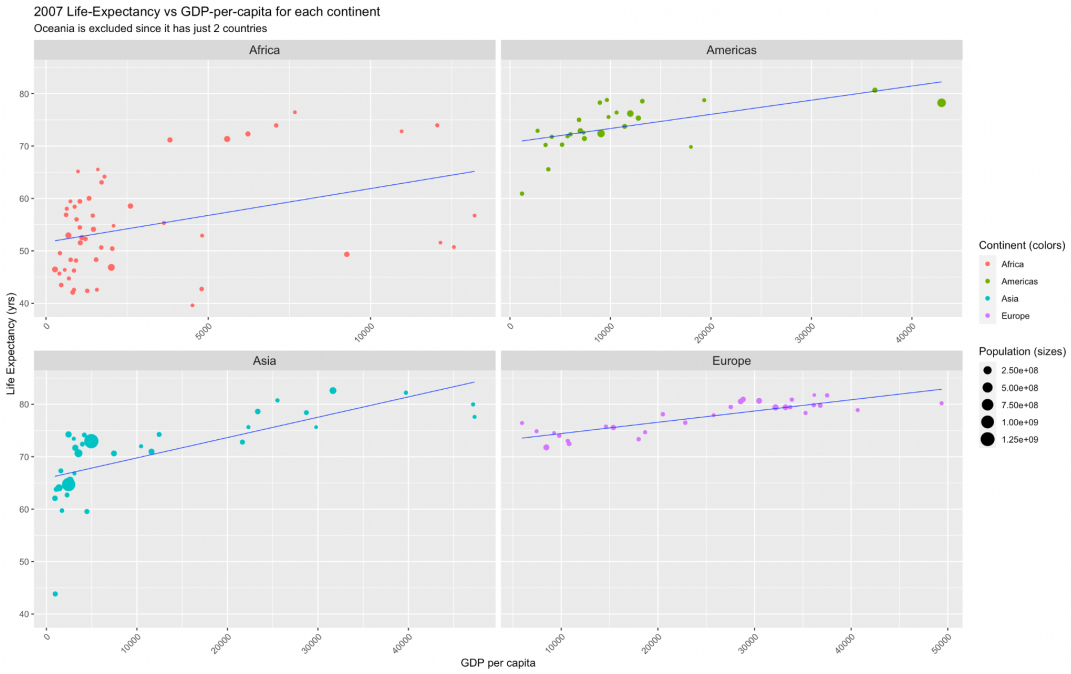
We have skipped Oceania from this analysis as there are just 2 countries with great GDP-per-capita values and correspondingly good Life-expectancies.

How does life expectancy vary with GDP per capita in 2007? Can the trends be well-described by a simple model such as a linear model, or is a more complicated model required?

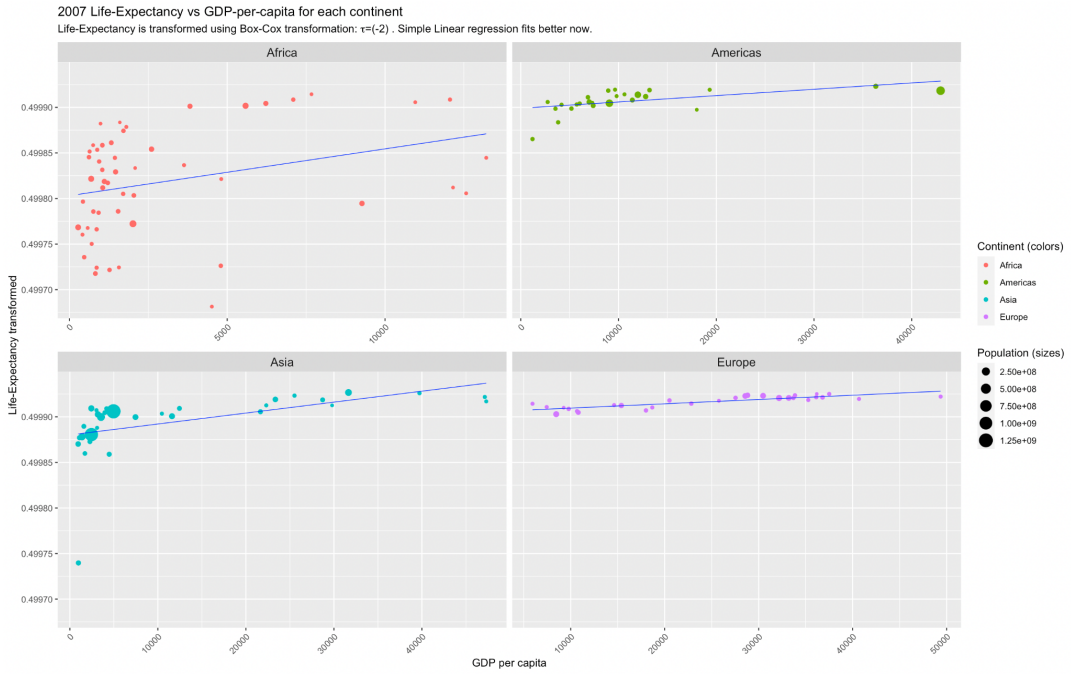


In the above plot, we fit four different models i.e. Linear Regression, Quadratic, Cubic and Loess for the entire data. The above graph shows that linear regression performs poorest in modeling the entire data. Other models such as quadratic and cubic models do slightly better but even a complex model such as Loess does not capture the data trends well.

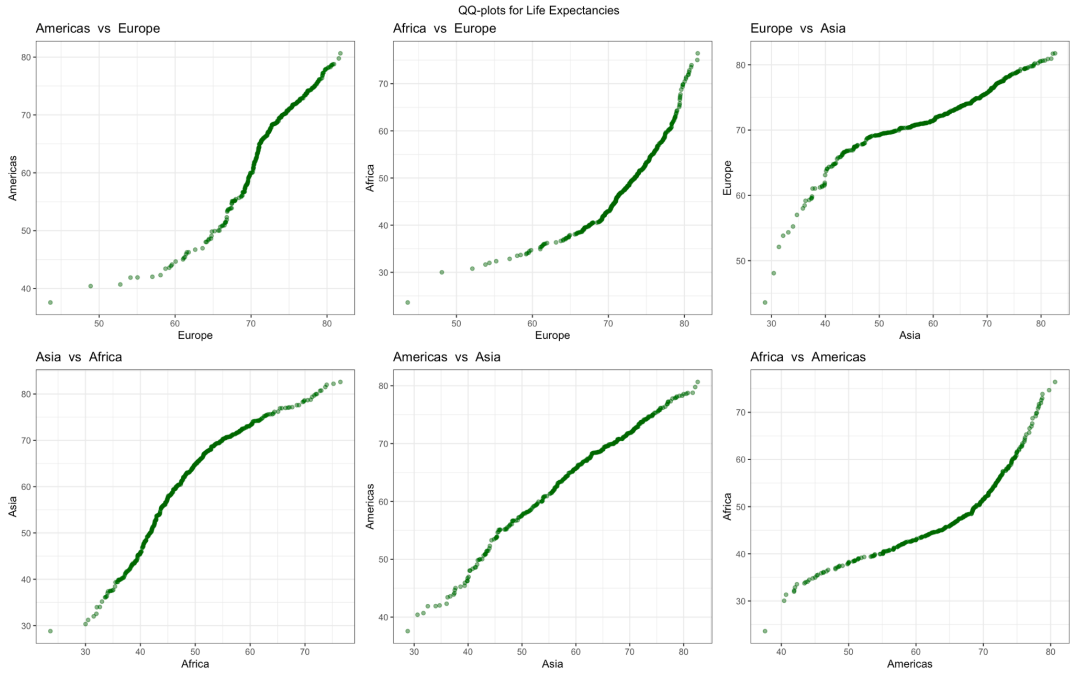
How are some continents different from others in terms of life expectancy vs GDP?



In the above plot, 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. Since there is no clear observable linear relationship between GDP and Life-Expectancy, a simple linear model here doesn't make sense.
Americas, Asia: We see a fairly linear relationship between GDP per capita and the Life-Expectancy. There are a few observable outliers for the Americas and Asia. If we fit a simple linear regression line, the predicted Life-Expectancy increases as GDP-per-capita increases.
Europe: Pretty uninteresting linear relationship between GDP and Life-Expectancy. Affordable healthcare could attribute to the lack of a drastic change in life expectancies.



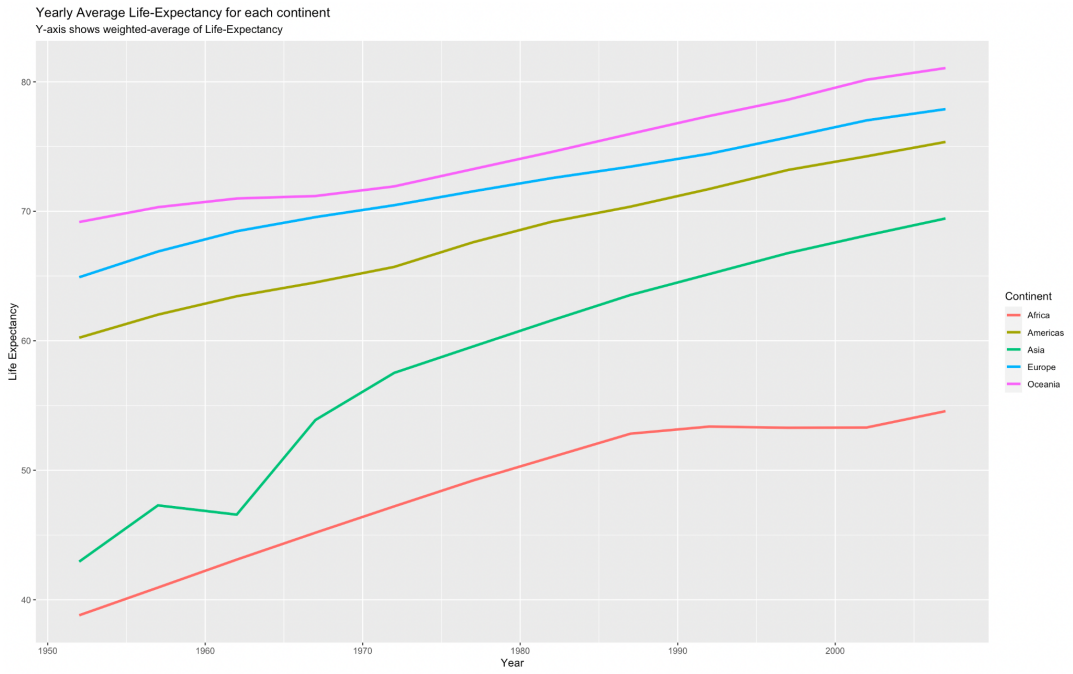
Africa: There is still no observable linear relationship between the features. We shouldn't model the distribution for Africa using a Linear-model.
Americas, Asia, and Europe: We see a better linear relationship between the transformed Life-Expectancy and the GDP-per-capita. Now we see fewer observable outliers for Americas and Asia if we fit a linear regression model.



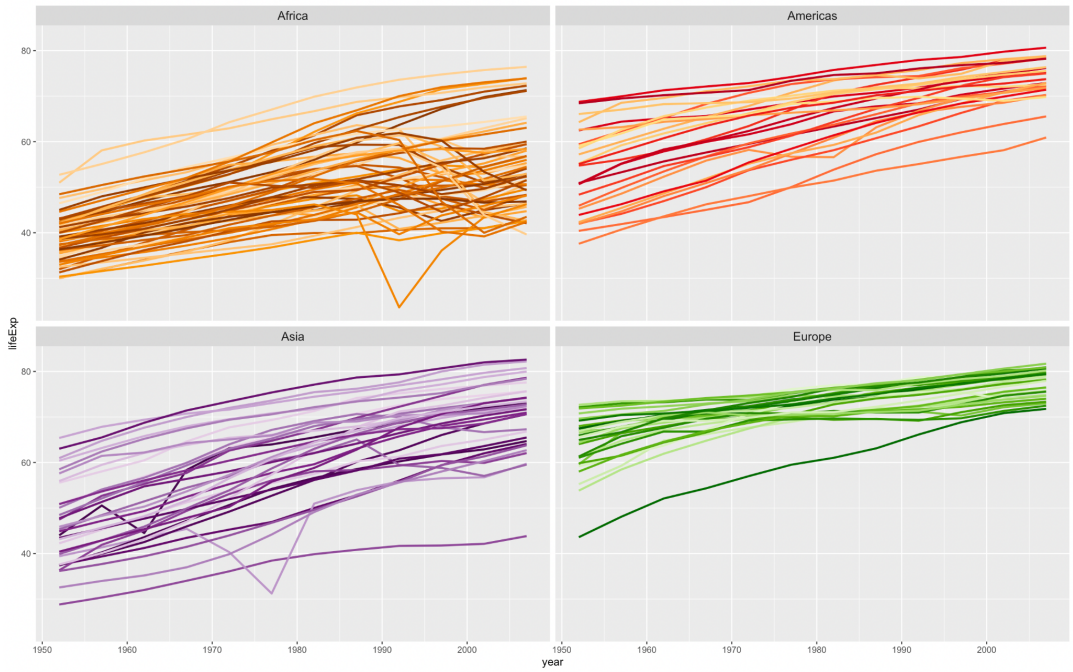
Can differences between continents be simply described by an additive or multiplicative shift, or is it more complicated than that?
The QQ-Plot of Life-Expectancies for Americas vs Asia is following a fairly straight line barring one/two outlier points. Essentially distribution of life expectancy in Asia and the Americas can be explained with some additive or multiplicative shift. The remaining QQ plots show quite complex relationships (the distributions are quite different) which cannot be merely described by simple additive or multiplicative shifts.

LIFE EXPECTANCY OVER TIME

How has average life expectancy changed over time in each continent? Have some continents caught up (at least partially) to others?
To answer this question we have plotted the average life expectancy for all continents over the years 1950-2010.

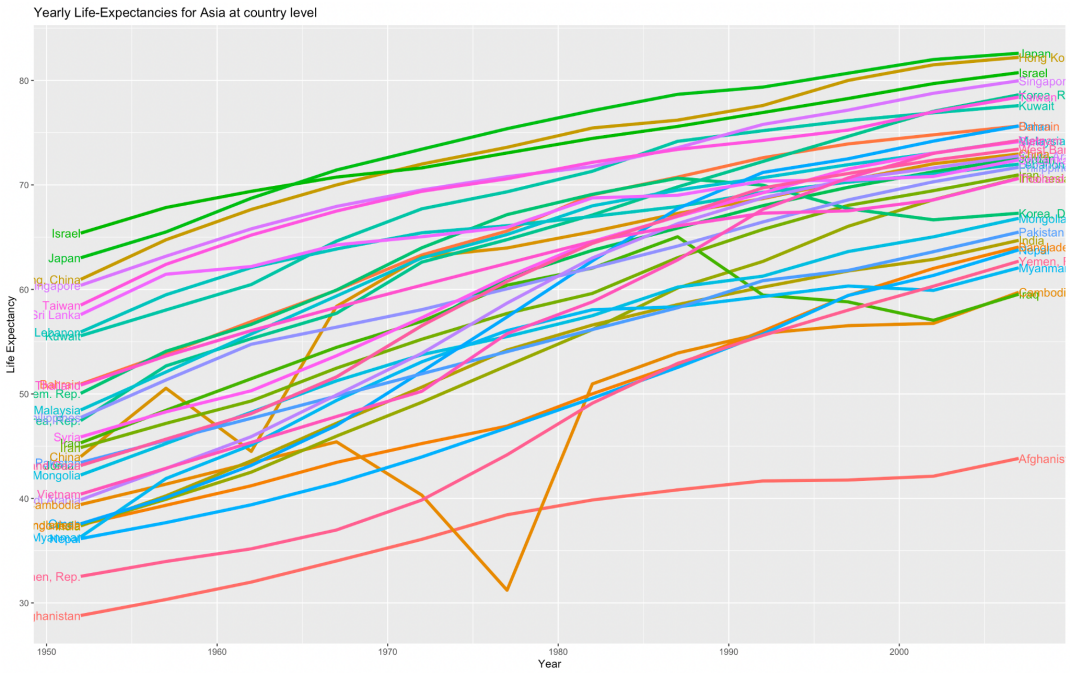


Is this just because of some countries in the continent, or is it more general? What might explain periods of faster/slower change?



On looking at the plot, we can see the life expectancies in most continents have been 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.

Have the changes been linear, or have they been faster/slower in some periods for some continents?
Average life expectancy in Africa has had a steady but relatively small increase from 30-50 to 40-60.



In Asia, the dip in average life expectancy (around 1962) can be attributed to a dip in 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 the countries have had a steady growth in their life expectancy which has contributed to the overall growth of Asia.

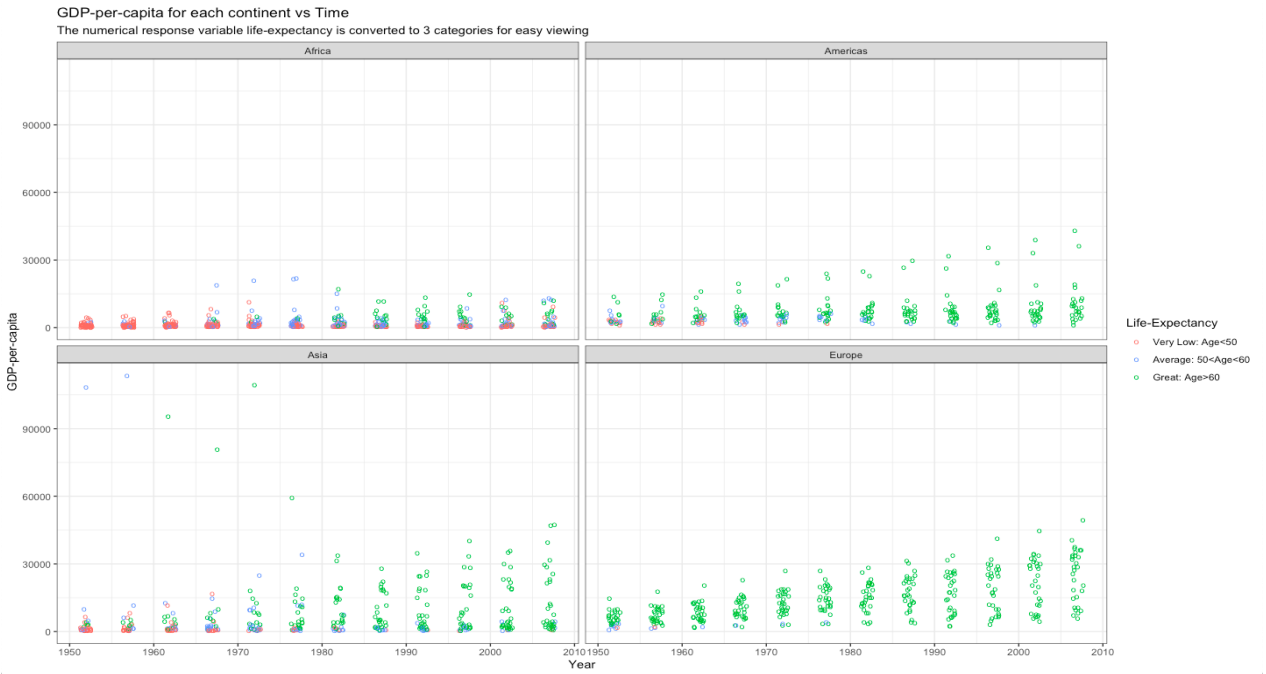
CHANGES IN THE RELATIONSHIP BETWEEN GDP AND LIFE EXPECTANCY OVER TIME

How has the relationship between GDP and life expectancy changed in each continent? Has there been "convergence" in the sense that perhaps GDP and/or continent doesn't matter as much as they used to? Are there exceptions to the general patterns?

Each continent has a regression line with a positive slope indicating that for higher GDP-per-capita, Life-Expectancy is higher on average. For each facet of the year, note that some points lie below the linear 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 just in general low there, irrespective of GDP. As we move through time till 2007, we see the slope change to a more positive outlook.
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.

Americas and Europe 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 this 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.

Can changes in life expectancy be entirely explained by changes in GDP per capita? Does it look like there's a time effect on life expectancy in addition to a GDP effect?

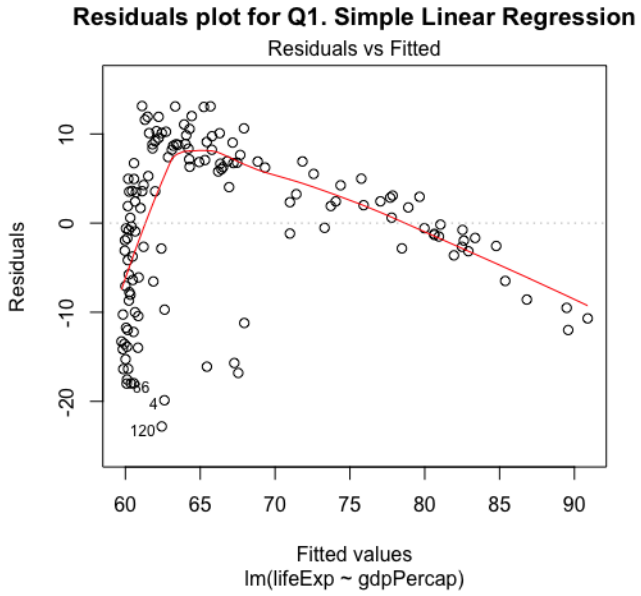


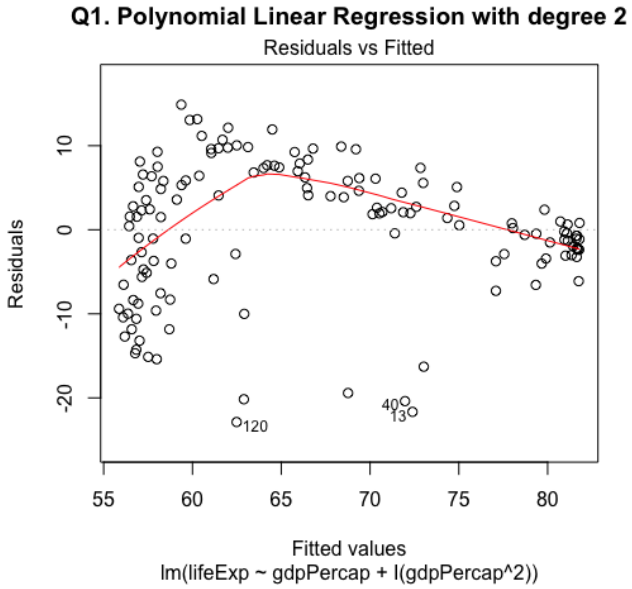
Changes in life expectancy cannot be entirely explained by changes in GDP. Changes in life expectancy are also dependent on time. Within a continent like Africa, even though the average GDP of countries stays similar, the life expectancy just seems to get better.

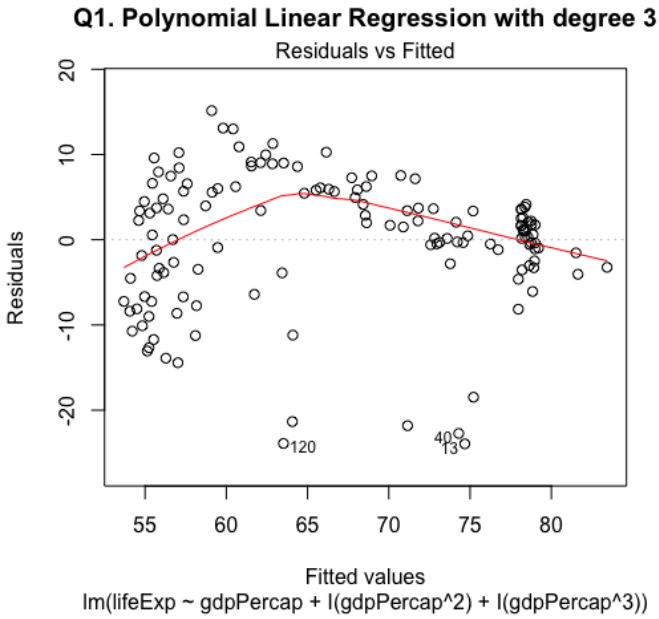
CONCLUSION

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






Q1. Polynomial Linear Regression with degree 3
Residuals vs Fitted
Residuals
20
10
0
-10
-20
120
40
13
55
60
65
70
75
80
Fitted values
lm(lifeExp ~ gdpPercap + I(gdpPercap^2) + I(gdpPercap^3))