STAT-S 670 Mini Project: Life Expectancy

Yash Shah
MS Data Science
Indiana University, Bloomington
Indiana, USA
yashah@iu.edu

Ritwik Budhiraja
MS Data Science
Indiana University, Bloomington
Indiana, USA
rbudhira@iu.edu

Srimanth Agastyaraju
MS Data Science
Indiana University, Bloomington
Indiana, USA
sragas@iu.edu

***Abstract*—This report talks about performing Exploratory Data Analysis on the open-sourced 'gapminder' data using R. The aim is to find trends and meaningful insights by creating data visualizations, transformations, and regression models amongst other methods.**

Keywords— gapminder, life expectancy, GDP per capita, continent, regressions, visualizations, correlations, countries, transformations, effects, R.

I. Introduction

The 'gapminder' data, which is freely available as an R package, was used in this study. This project is an effort to perform data analysis for a researcher to assist him in learning about life expectancy and its relationship to GDP per capita across time—from 1952 to 2007, and across the globe. Asia, the Americas (North and South), Europe, Africa, and Oceania (including Australasia, Melanesia, Micronesia, and Polynesia) are among the continents from which the geographic data was gathered. The major goal of this project is to determine whether a rise in life expectancy since World War II can be explained by large increases in GDP per capita across continents. The researcher has divided this open-ended question into three parts, each having subparts, to make it easier to work with.

II. Questions & Inferences

A. GDP and life expectancy in 2007:

i. How does Life Expectancy vary with GDP per capita in the year 2007?

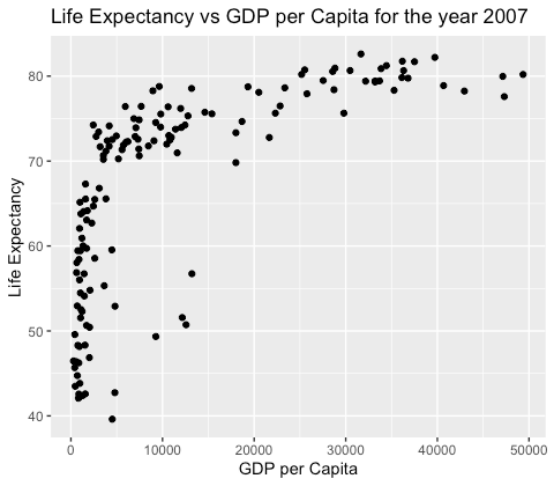

Figure A.1

Figure A.1 shows a scatterplot of life expectancy vs. GDP per Capita in the year 2007. We can see that the plot demonstrates a positive, non-linear relationship between the two variables, which is self-evident. Although a scatterplot depicts correlations or links between two variables, there is no certainty that the graphic depicts a cause-and-effect relationship. When thinking pure logic, GDP per capita increases the life expectancy at birth through increased economic growth and development in a country and thus leads to the prolongation of longevity. But on the contrary, death is aided by a flourishing economy in other ways as well. People begin to spend more time at work, exposing them to workplace risks as well as the stress of overworking. People are driving more, which has resulted in a rise in traffic-related deaths. People are also drinking more, which leads to health issues and accidents. In the short term, or when the economic fluctuations are small, the harmful effects of growth dominate. But in the long term, or when the economic fluctuations are big, those harms are counterbalanced by the positive effects of having more income. We need to go deeper into the data to establish a significant cause-and-effect relationship.

ii. Can the trends be well-described by a simple model such as a linear model, or is a more complicated model required?

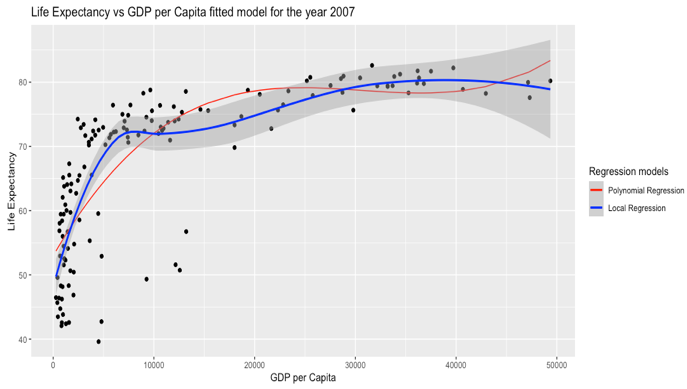

Figure A.2

Since the relationship is not linear, we cannot use a basic model like a linear model to do regression. As shown in *Figure A.2*, we use a polynomial model with degree 3 and a local regression model using the 'loess' method. We considered utilizing degrees higher than 3 while using the polynomial model, but this resulted in the model being overfitted.

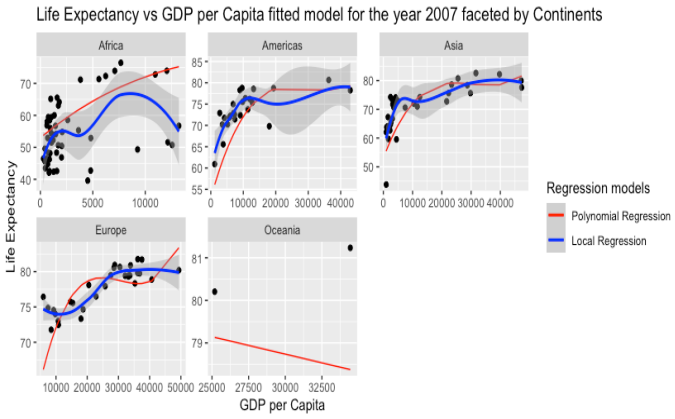

Figure A.3

iii. Is the pattern the same or different for every continent? If some continents are different, which ones?

We only look at Africa, the Americas, Asia, and Europe to see if there are any trends in the plots faceted over different continents. We can exclude Oceania because there is inadequate data on it, making regression model fitting problematic, as illustrated in *Figure A.3*. We also pass *scales = "free"* as an input to the *facet_wrap()* function when plotting these graphs. This frees up the scales, allowing us to see the plots in a more convenient window format.

When we look at the African continent, we can see that the local model does not capture much data until GDP per capita exceeds 2500 and that life expectancy drops to 55 as we approach 10,000 on the x-axis, whereas the polynomial model indicates an increase. This could be a point of disagreement between the two models, or we could argue that the data is insufficient to make a definitive statement about the African continent. On the other side, we do not witness a fall in the Americas, Europe, or Asia, but we do find that it is becoming steadier as GDP per capita rises. The data for Asia and the Americas are clustered near the beginning of the plot, while the data for Europe are clustered near the middle. As the GDP per capita approaches 45000, life expectancy on all three continents tends to reach 80.

iv. Can the difference between continents be simply described by an additive or multiplicative shift, or is it more complicated than that?

When data is a times-series, we recognize it and focus on three primary parameters which are seasonality, trends, and residuals. The link between the x and y variables is easy to understand with the help of a basic linear equation. Multiplicative and additive shifts can be used to define these trends. They provide information on changes in one variable in relation to another. We can claim that the most basic equations derived to see these shifts are as follows:

Additive Shift = Trends + Seasonality + Residuals
Multiplicative Shift = Trends * Seasonality * Residuals

When seasonal variation is generally constant over time, the additive model is useful, but when seasonal variation rises over time, the multiplicative model is useful. For example, if seasonal variations appear to be of similar magnitude throughout time, we can conclude that there is an additive shift. We can claim that there is a multiplicative shift across the continents when seasonal fluctuations rise over time. These are crucial topics to understand when working with time-series data in Machine Learning. Because our situation isn't as straightforward as a single linear equation, we can't be certain whether the changes are simple additive or multiplicative by glancing at the graphs. Since we can witness changes in variances in certain places and stable variations across time while juxtaposing the facet plots, they can be said to represent a mix of both shifts.

B. Life expectancy over time by continent:

i. How has average life expectancy changed over time in each continent?

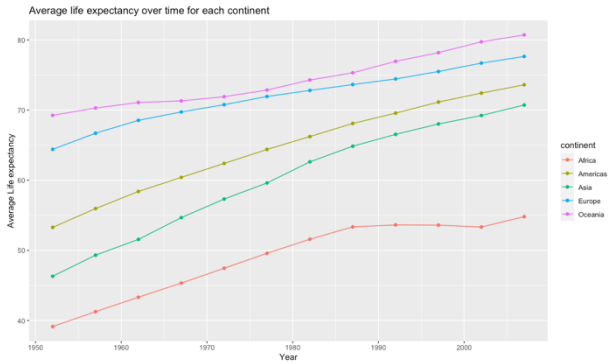

Figure B.1

As shown in *Figure B.1*, Except in Africa, the average life expectancy for each continent has been steadily increasing over time. The average life expectancy in the Americas and Europe increased linearly. From 1987 to 2005, Africa's life expectancy has been stagnant or declining. The HIV/AIDS epidemic, which increased rapidly in the 1980s, may have contributed to the stagnation or reduction in average life expectancy in Africa during this time. Given the range (46 to 66) and slope of the average life expectancy with the time curve, Asia has witnessed the greatest increase in average life expectancy. In the years 1970-1980, the average life expectancy in Oceania decreased.

ii. Have some continents caught up (at least partially) to others? If so, is this just because of some countries in the continent, or is it more general? Have the changes been linear, or has it been faster/slower in some periods for some continents? What might explain periods of faster/slower change?

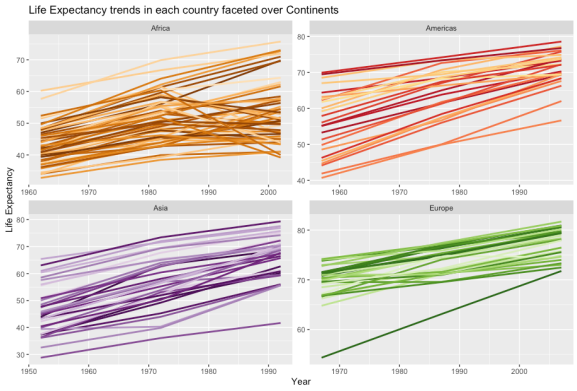

Figure B.2

In terms of average life expectancy, Asia has begun to catch up to the Americas and Europe. We can infer from *Figure B.2* that life expectancy in one of Asia's countries increased dramatically between 1960 and 1970. This could be due to the expansion of healthcare services and general health awareness in developing countries like South Korea and India. When it comes to the African continent, we observe an increase in life expectancy from 1996 to 1980, but as we go ahead of 1980, we see a fall in the life expectancy of many African countries, which is mostly due to the enormous spike in AIDS mortality rates. Overall, because numerous countries show abrupt peaks and dips for unknown causes, visualizing the temporal changes in life expectancy by country is particularly rewarding. This stimulates more comprehensive data aggregation efforts to proactively identify all countries whose data demonstrate certain characteristics. Understanding how each country's trends are changing is part of the future scope.

C. Changes in the relationship between GDP and life expectancy over time:

i. How has the relationship between GDP and life expectancy changed in each continent? 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?

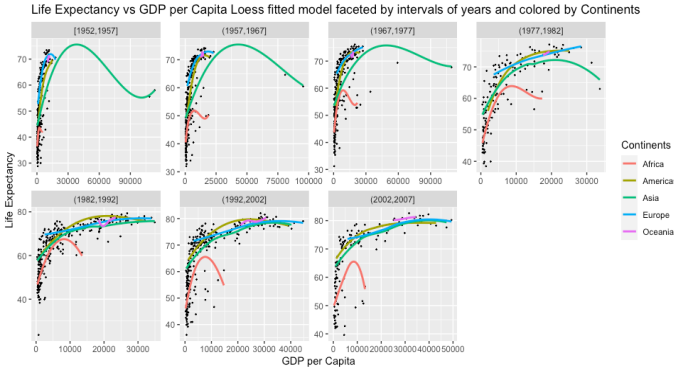

Figure C.1

Figures C.1 shows the link between life expectancy and GDP per capita shown on the axes, with continents separated by colour and faceted through time intervals. While comparing GDP per capita and life expectancy across continents, we can find a general tendency of a quick increase in life expectancy when GDP per capita is between 0 and 10000, but then it increases logarithmically as we move along the X-axis. We can't argue that variations in GDP per capita are fully responsible for changes in life expectancy. Time, among other variables, is particularly important in establishing a cause-and-effect relationship between the two. Technology, industrialization, globalization, and advancement in innumerable sectors came with the progression of time. All these factors, as previously indicated, have an impact on the overall link between life expectancy and GDP per capita. *Figure C.2* shows a correlation matrix that is used to construct relationships between the three variables. We can observe the highest link between GDP per capita and life expectancy is numerically supported by this correlation matrix. We can see that time is a factor that affects life expectancy, and while it is minor, it cannot be overlooked. So, we can advocate the fact that there is a time effect on life expectancy in addition to the GDP effect.

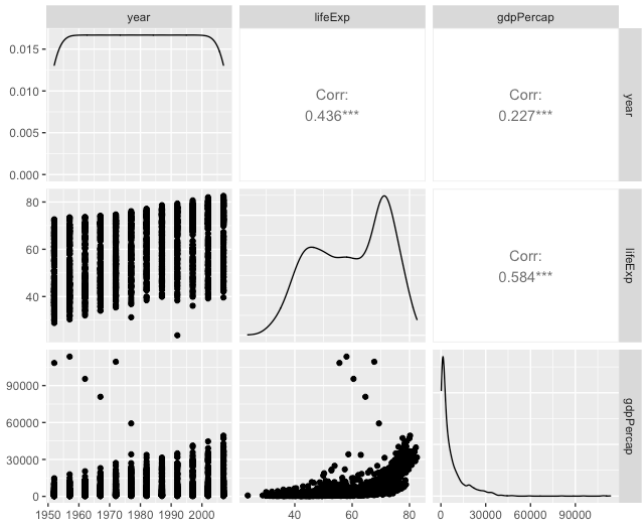

Figure C.2

ii. Has there been "convergence" in the sense that perhaps GDP and/or continent don't matter as much as they used to? Are there exceptions to the general patterns?

Yes, there has been a sense of "convergence" between the factors, indicating that GDP and continent are less important than they once were. One way to think about this is that, regardless of the year, life expectancy has remained fairly consistent for continents such as Asia, the Americas, and Europe, even as GDP per capita has climbed from 10,000 to 60,000. Despite the reduction in life expectancy on the African continent, it has stayed stable (seeing a drop) throughout time and as GDP has increased.

III. Conclusion and Scope

The rate of change in life expectancy in a given location varies with changes in GDP per capita and through time. The questions' abstract character prevents them from yielding specific responses. However, with the use of exploratory data analysis, it is always possible to construct logical arguments;

here is an example. The 'gapminder' data is extensive, and there is plenty of room for more investigation. The main focus of this article was on the impact of GDP per capita, geographic location, and time on life expectancy. We now have a better understanding of how changes in one variable affect the other and how dissimilar the results are. We intend to extensively investigate this data and disclose our findings as part of the project's future scope.

IV. Acknowledgements

We would like to express gratitude to Prof. Julia Fukuyama for being an excellent instructor during this course. We would also like to thank TA Paul Hunt for his constructive feedbacks and guidance.