STAT – S670: Exploratory Data Analysis

Mini Project 1: LIFE EXPECTANCY

Authors: Saranjeet Singh Saluja, Sricharraan Ramaswamy, Athulya Anand

1. Introduction:

In this exploratory data analysis project, we have used the Gapminder Dataset for analysis. Three specific sets of questions have been answered through a series of graphs and related inferences. These graphs function in the following way:

- Identifying the relationship between Life Expectancy and GDP Per Capita for a specific year
- Identify changes in Life Expectancy over time and continent
- Identify changes in the relationship between GDP and life expectancy over time and continent.

We draw the following inferences from the above:

- Plotting the Life Expectancy vs Log GDP per capita showed a linear trend (increase in Log GDP Per Capita leads to increase in Life Expectancy) with quite a spread
- Plotting changes in Life Expectancy over time and continent depicts the growth of life expectancy of Asia.
- Plotting the relationship between GDP and life expectancy over time and continent depicts the variation and changes in trend of GDP in each continent and the effect of time & GDP on life expectancy.

2.1. Exploring Relationship between GDP Per Capita and Life Expectancy in 2007:

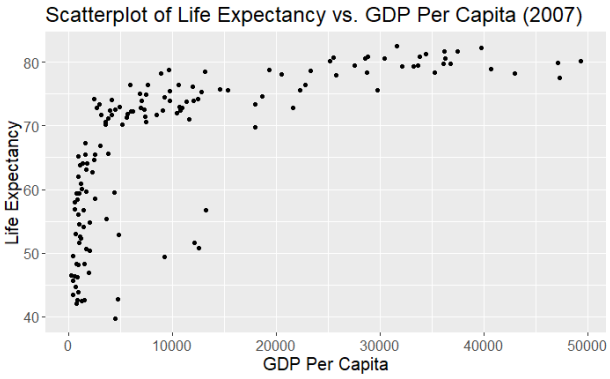

Fig 1. Relation between Life Expectancy and GDP

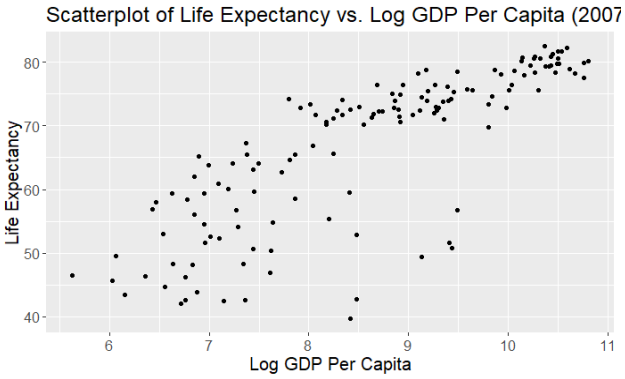

Fig 2. Relation between Log Life Expectancy and GDP

Fig.1 shows the relationship between GDP per capita and Life Expectancy (2007). There is a non-linear relationship between the variables in the form of an inverse-L shape or logarithmic. There are outliers, but the general pattern appears logarithmic. The majority of the values of Life expectancy are between 40 - 70 (some are around 85). Thus, the transformation of such a variable does not make sense. However, the GDP values range from around 300-50000. Applying a logarithmic transformation might help.

Fig.2 shows the application of log transformation to GDP per capita variable. Based on Fig.2, Life expectancy can be determined by a simple linear model. However, it most probably won't be the perfect model.

2.2. Identifying continents with similar patterns:

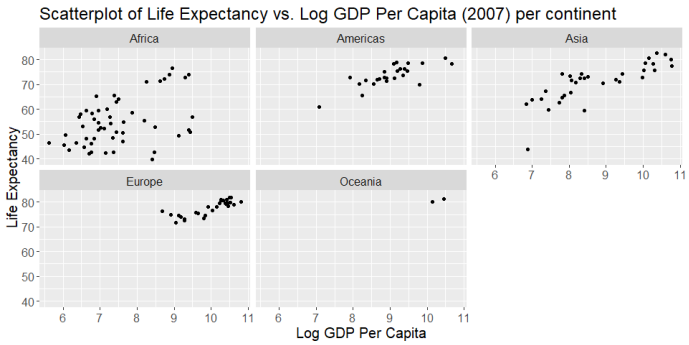

Fig 3. Relation between Life Expectancy and Log GDP for each continent

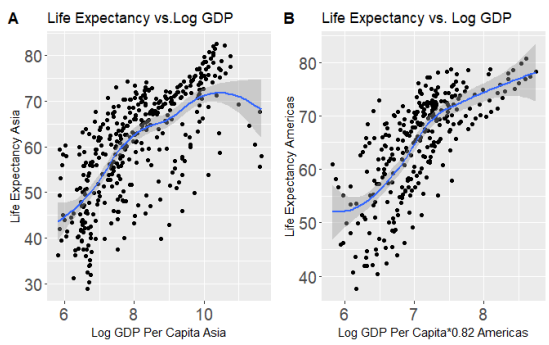

Fig 4. Deciphering similarities between Asia and America

The African continent has majorly a low GDP per capita (barring some outliers) but their Life Expectancy is mostly between 41-70. The life expectancy of the Americas, Asia, and Europe appear to have a linear relationship with GDP per capita. The trend is not the same for every continent except Asia and the Americas as they look similar. However, the multiplicative shift is not apparent, due to fewer data points.

Therefore, we will see the similarity by taking into consideration the entire dataset. The difference in the Asian and Americas continents can be explained by a multiplicative shift. As seen in Fig.4, the GDP Per Capita(log-transformed) of the Americas continent is multiplied by 0.82 for every value to show a similar trend of Life Expectancy vs GDP for the Asian and Americas continents (observe the 6-8 range on the x-axis).

3.1. Exploring Life Expectancy over time and continent:

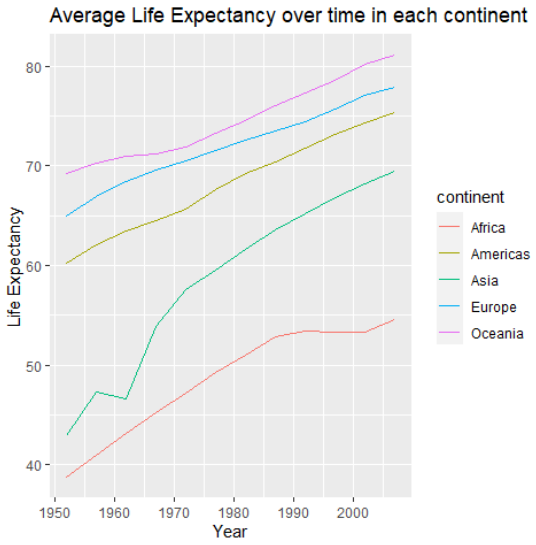

Fig 5. Relation between Average (Weighted) Life Expectancy and Year

3.2. Average life expectancy changed over time in each continent:

Africa: The continent of Africa is following a linear trend till 1987 and after that, the growth of life expectancy has been constant till 2002 and a slight growth between 2002 and 2007.

Asia: Initially, the plot is disruptive, and it is very tough to conclude it, but the growth has been exponential till 1972 and from there Asia is trying to play the catch up compared to its counterparts.

Americas: The growth of life expectancy has been linear throughout the timeline except for the year 1972 when there is a small dip in the growth.

Europe: The life expectancy of Europe has been linear throughout the timeline.

Oceania: Oceania has been in the forefront throughout the timeline and has a linear growth expect between 1962 - 1967 where the growth has been stagnant.

3.3. Life Expectancy over the years in different countries of Asia:

Fig 6. Life Expectancy over the years in different countries of Asia.

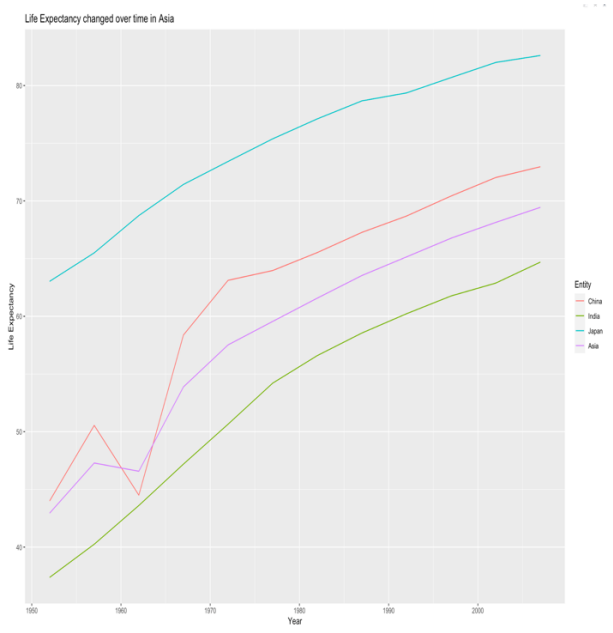

Fig 7. Life Expentancy over the years in Asia's top countries

From Fig. 5, we can see that Asia is trying to catch up with other continents barring Africa. So, after taking a closer look at the countries that contribute to the average life expectancy of the continent in Fig. 6, we can see that the life expectancy of the countries like Japan, China, and India play a major role in the growth of the life expectancy of the continent since the population of these countries are quite large and in weighted average population plays a major role in calculating the average life expectancy of the continent. In Fig. 6, the number of countries being more it is difficult to interpret substaintially. Thus, on further analysis on the countries which contribute more to the life expectancy in Fig. 7 we can clearly see that the average life expectancy of Asia is almost similar to China as in 1957, the dip in life Expectancy in China affects the dip in life expectancy of the whole continent.

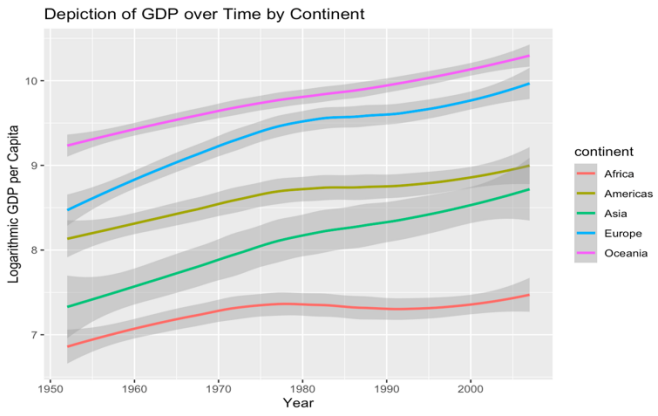

Fig 8. GDP per Capita over the years

3.4. Factors deciding the rate of growth of Life Expectancy in Asia and Africa:

In Fig. 5, we can see the faster growth of life expectancy in the African continent whereas after 1987 the growth became stagnant whereas, from 1962 the continent of Asia has had a faster growth rate of life expectancy till 2007.

On comparing the Fig. 5 and Fig. 8, one can clearly see that the rate of growth of GDP per Capita was more linear for the continent of Africa till the 1980s and it became stagnant afterwards similar to the life expectancy in Fig. 5 whereas for the continent of Asia the growth of GDP per capita is linear throughout the years and because of which the rate of growth of life expectancy has been more for Asia and less for Africa because of the stagnant growth in GDP per Capita.

4.1. Relation between GDP and Life Expectancy in each continent between 1952-2007:

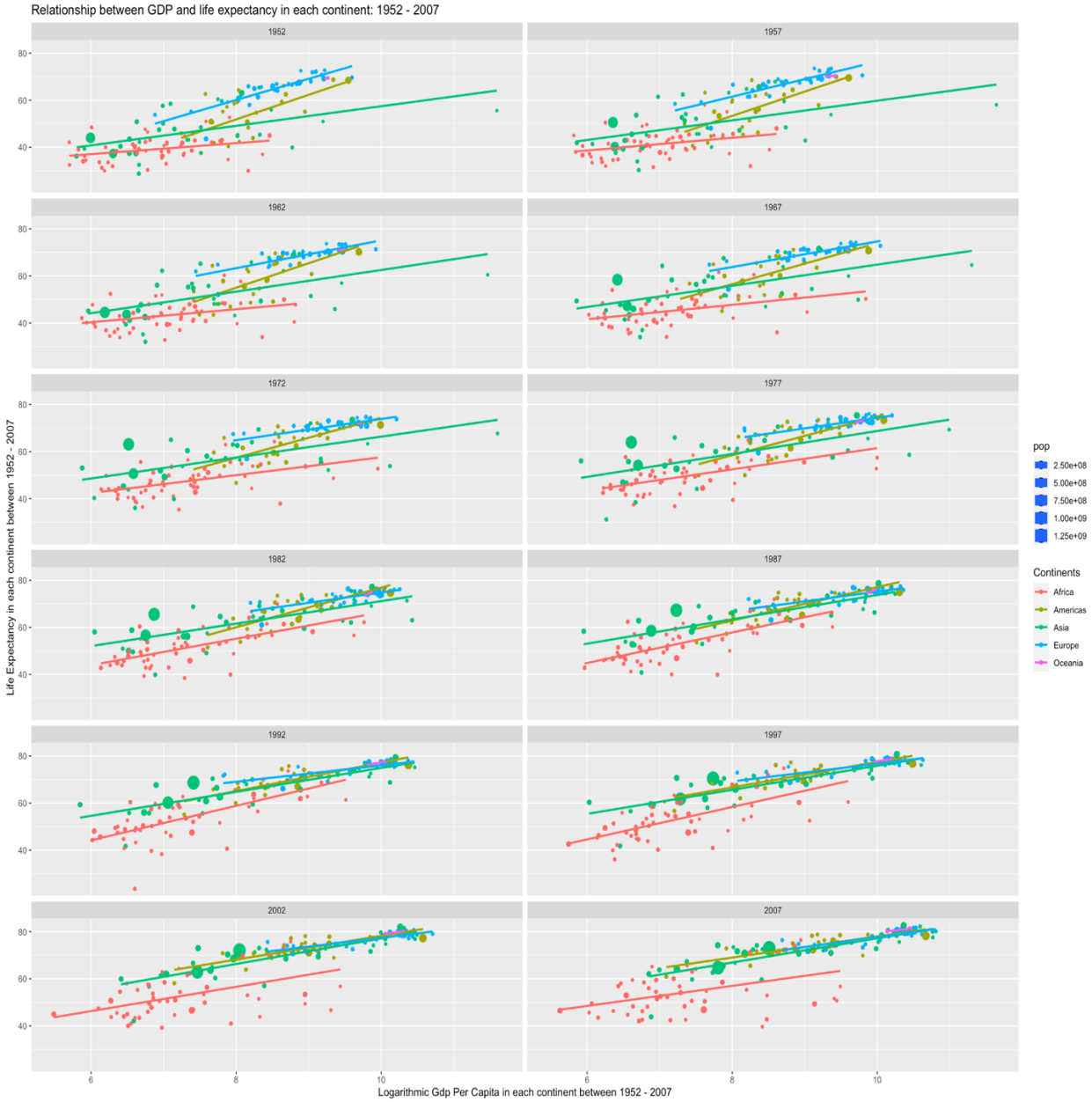

Fig 9. Year-wise Relationship between Life Expectancy and GDP per capita per country in the continents

4.2. Effect of the relationship between GDP per capita and life expectancy in each continent:

Fig 8: From 1952 – to 1977, it can be observed that African nations, along with many Americas and Asian nations, have a low GDP and low Life Expectancy. However, the distinction between the continents becomes clearer from 1977 onwards as the Asian and Americas nations have an increased GDP per capita and potentially a higher Life Expectancy.

In a nutshell, we can conclude the relationship between the GDPs per capita and Life Expectancy is very different in each continent. However, the data points depicting Americas and Asia nations fall on approximately the same positions (especially in 2007), thereby indicating a similar linear relationship as explained by the multiplicative shift earlier. Thus, we can infer from the plot that changes in life expectancy cannot be entirely explained by changes in GDP per capita. This is solely because there are multiple factors like the population size, per capita income, longevity, etc which affect this relationship. Noticeably, Asia has the largest population size as compared to other continents which could be one of the factors affecting its life Expectancy.

Additionally, it can be observed that the data points representing life expectancy in the African countries is closer to the regression line in the year 1952. However, in 2007 although its life expectancy has increased slightly, there is an increase in variability and the datapoints are more spread out from the regression line. Thus, even though the GDP per capita has increased, the life expectancy is still comparatively low. This is not the case in rest of the continents, wherein the data points are approximately closer to their regression lines between 1952-2007 throughout.

4.3. Relationship between convergence and GDP per capita with respect to each continent

Taking the regression lines into consideration, we can observe that between 1952 and 1977, the lines representing each continent follows a different trend indicating different GDP growth in each continent. We can also notice that as the years have passed, precisely after 1982, the regression lines representing each continent namely Africa, Asia, Americas, Europe and Oceania converges towards the same point. However, in years 2002 and 2007 it is evident that the regression line representing Africa diverges away from the other countries, thereby representing a slow GDP growth. Overall, we can infer that the growth in GDP per capita is independent of the growth in each of the continents.

4.4. Time effect on life expectancy in addition to GDP per capita effect:

In conclusion, this graph tells us that over time the GDP per capita has increased for almost all continents and that time has a major role to play in altering the Life Expectancy and the GDP Per Capita of each continent. Moreover, from fig 8 and 9 we can clearly tell that although each continent's GDP per capita follows a very different trend, it confirms that there is a time effect, playing a pivotal role on the life expectancy in addition to GDP Per Capita effect.

5.1. Modeling:

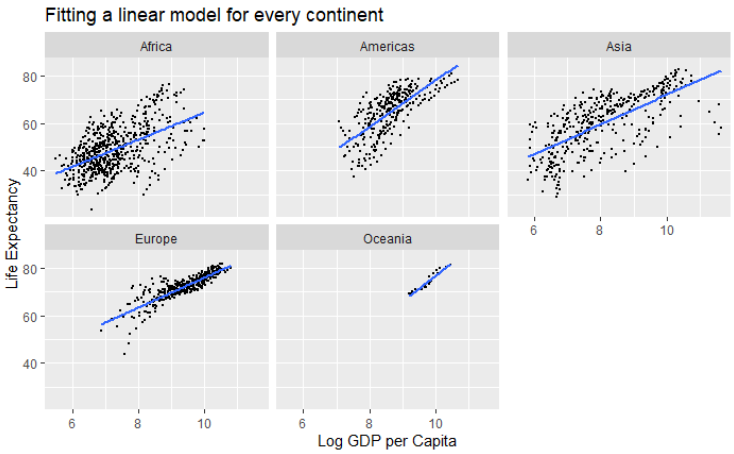

Fig 10. Fitting a simple linear regression model for every continent

On simply fitting a simple linear regression model, we can see that for the continents Americas, Europe and Oceania, the model fits quite well with the single predictor Log GDP per Capita. However, it does not fit as well with Asia and Africa, and a more complicated model such as LOESS might be required if the error tolerance is to be kept too low.

Interpretation of the model:

Consider fitting the simple linear regression model for Europe:

Coefficients:

	Estimate	Std. Error	t-value	p-value
Intercept	12.9651	1.9171	6.763	5.52e-11
gdpPerCap	6.3074	0.2045	30.838	< 2e-16

Fig 11. Summary of the simple linear regression model (Europe)

As we can observe, the coefficient is 6.3074, which will be divided by 100 [1]. We get the value of the new coefficient as 0.063074. Now this implies that, for every 1% increase in the independent variable, our dependent variable increases by about 0.06.

Similarly, if we carry out fitting a simple linear regression model for every continent, the GDP per cap variable is statistically significant for every continent (p-value close to 0).

Conclusion:

As per our analysis, GDP per capita plays an important role in determining Life Expectancy. Even though it can be argued that GDP per capita does not entirely define Life Expectancy, it can be said that majorly does affect Life Expectancy. However, both the dependent and independent variables are dependent on time, i.e., time is confounding the model estimates. Unfortunately, this issue cannot be resolved by including time as a dependent variable since one of the assumptions of multiple linear regression is independent predictors.

Reference/s:

[1] https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/