STAT S670 Final Project: Life Expectancy Report

Authors: ankale, rdama, sgopalk, sk128

INTRODUCTION: -

The aim of the project is to understand how life expectancy is dependent upon different factors like healthcare expenditure, immunization rates, education, and geographic location. We want to analyze how these variables affects the life expectancy, which will tell us the areas of improvement for countries.

QUESTIONS BEING ADDRESSED: -

- Did ***life expectancy*** grow over the years for developing and developed countries? Is the rate of growth constant among these countries?
- Dependency of ***life expectancy*** with the immunization taken for different diseases. Does taking a vaccine increase life expectancy?
- Does ***education*** play a role in ***life expectancy***?
- Does a country spending more on health care have a higher ***life expectancy***?
- Does the geographic location of the country has any effect on the ***life expectancy***?

STATEMENT OF GOALS:-

The factors affecting ***life expectancy*** have been studied in the past considering demographic variables, income composition, and mortality rates. In the past, the effects of immunization and heath care expenditure were not considered. Some of the past research used data from a single year for all countries. Hence, this gives us an opportunity to resolve both factors while considering data from a period of 2000 to 2014 for all the countries.

The aim of the project is to understand how ***life expectancy*** is dependent upon different factors like economical, immunization rate, diseases, lifestyle, and health care expenditure in a country.

This will help a country determine what areas are most important to improve life expectancy.

We analyzed the four major segments which affect Life expectancy using the following factors:

- Healthcare & Immunization- expenditure on healthcare as a percent of GDP and polio immunization.
- Education- schooling years.
- Economic factors- Developed or Developing countries.
- Geographic location- Continents the country belongs to.

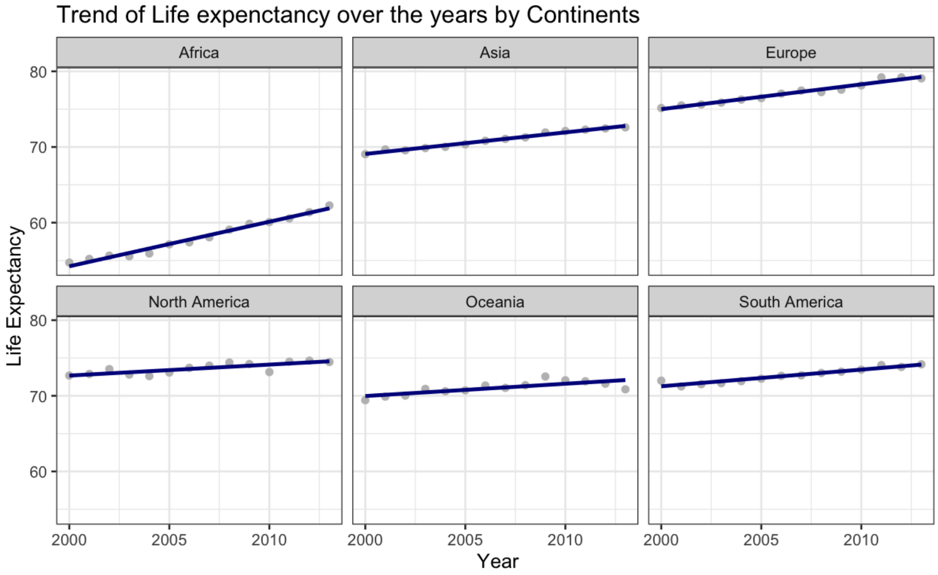
DATA DESCRIPTION: -

The dataset is available at https://data.worldbank.org/indicator/SP.DYN.LE00.IN . The dataset related to life expectancy and health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. In total, there are 8 columns and 3000 rows in the final merged file. An initial look at the data revealed some missing values. Since the data sets come from WHO, we found no obvious errors. Based on the results, most of the missing data were related to population, Hepatitis B, and GDP. The variable descriptions:

- **Life Expectancy:** Measured in Age (Years)
- **Country**
- **Year**: Ranging from 2000-2014
- **Status**: Developed/Developing
- **Polio**: Percent of 1 year-old population vaccinated Polio
- **Diphtheria**: Percent of 1 year-old population vaccinated against Diphtheria
- **Total Expenditure**: General government expenditure on health as a percentage of total government expenditure (%)
- **Schooling**: Number of years of Schooling(years)

EXPLORATORY DATA ANALYSIS: -

Over the span of 15 years, the general trend ***life expectancy*** is in the increasing direction. As the years go by, the ***life expectancy*** increases. The rate of growth of ***life expectancy*** is somewhat constant across all the continents except Africa.



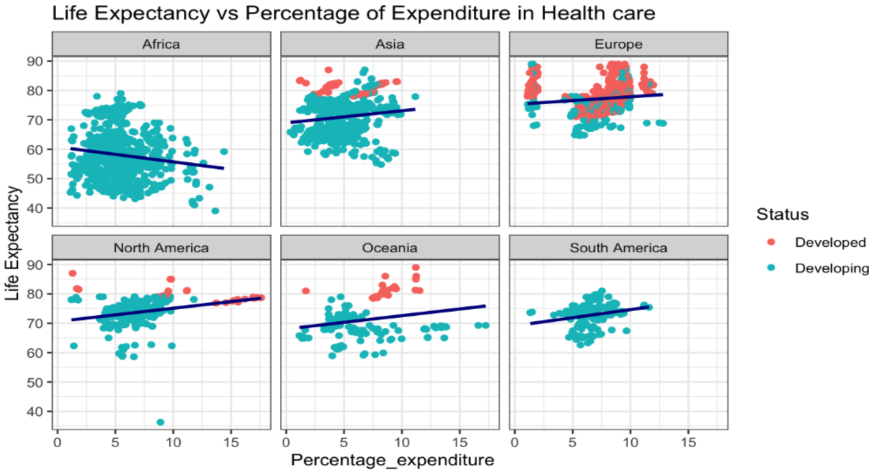
In the continent of Africa, the percentage change in the ***life expectancy*** over the years is comparatively higher. We see that there is almost a 20% increase in the ***life expectancy*** in Africa over the 15 years while there is only a 5-8% increase in life expectancy in the other continents.

We will look into three major factors which might affect ***life expectancy***:

1. Percentage of expenditure
2. Immunization
3. Schooling

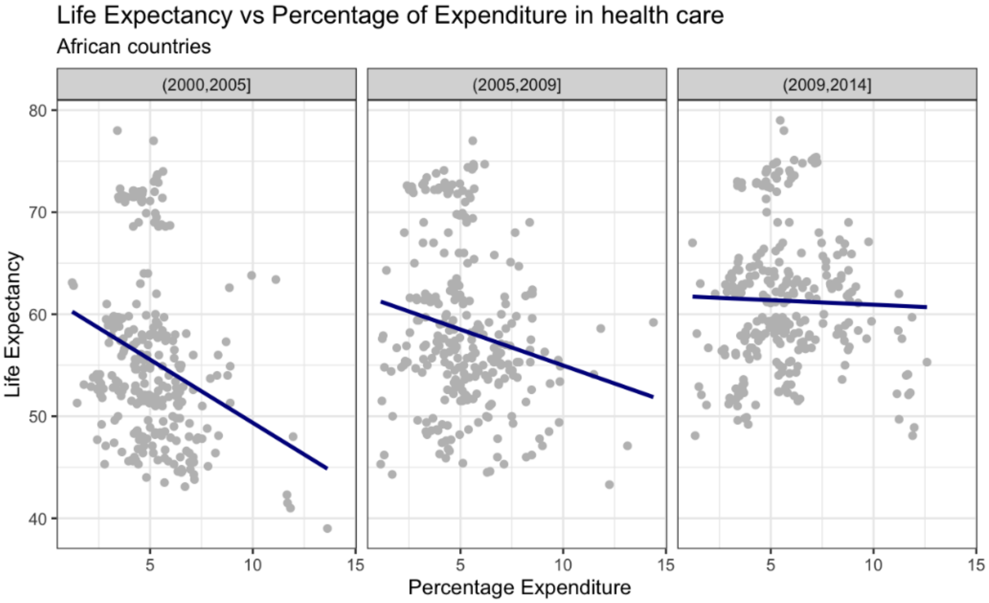
Life expectancy vs Percentage of expenditure:

As the Percentage of expenditure increases, the general trend is that ***life expectancy*** also increases. Usually developed countries spend more on healthcare and thus we can see that more developed continents like Europe and North America have a positive trend to the ***percentage expenditure***. However, in the case of Africa even with the increase in the ***percentage expenditure*** there is a decline in the ***life expectancy***.



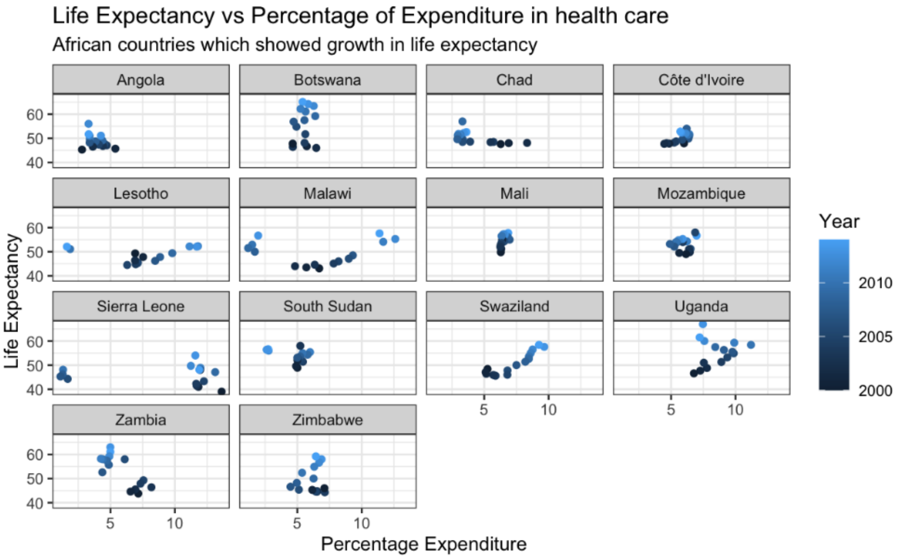
We wanted to dig deeper and understand the reason behind this:

Looking further into Africa, we divided the 15 years African data into buckets of 5 years to see the trend and we observed something interesting. Between the years 2000-2005 ***life expectancy*** was strongly negatively correlated i.e., as the ***percentage expenditure*** increased ***life expectancy*** decreased which was surprising. However, as the years increases this trend goes away.



After some research we noticed that most of the deaths which occurred in the years 2000-2010 were mainly due to the two diseases HIV and tuberculosis.

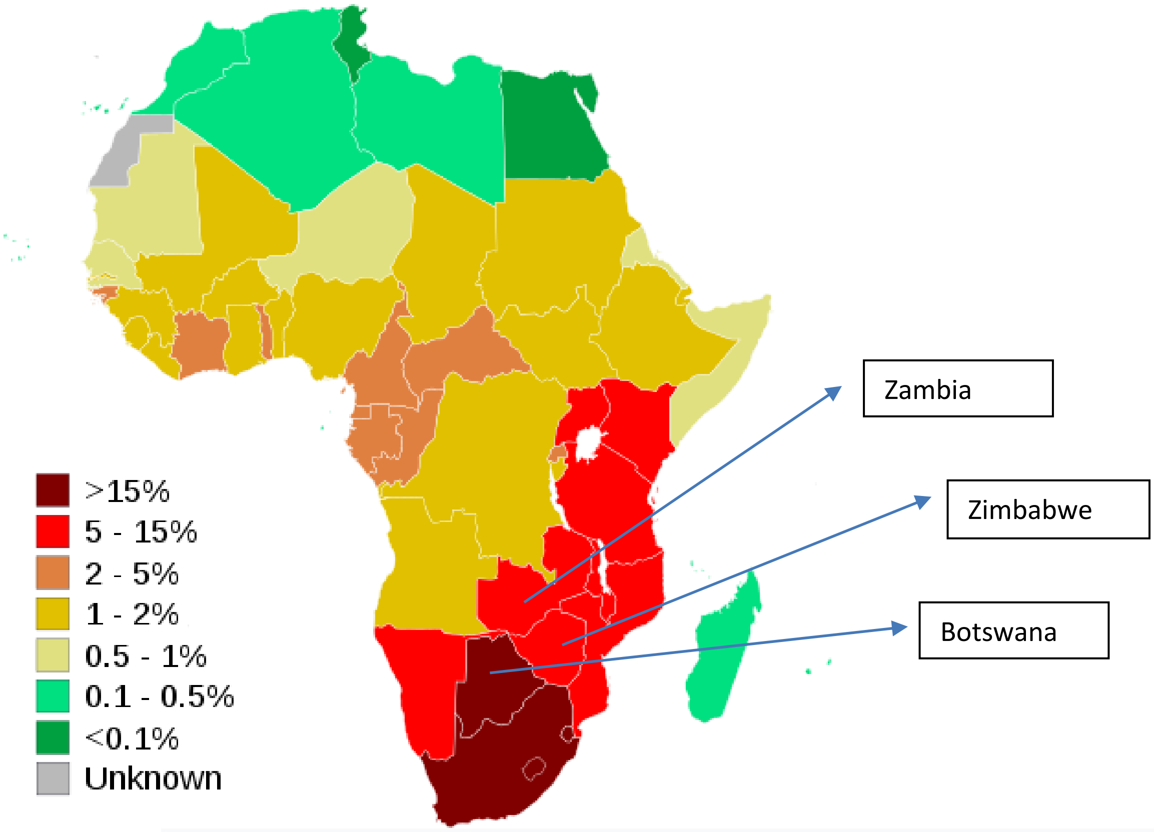
Our hypothesis is that due to the deaths which occurred from HIV there would have been a lot of hospitalizations which would have resulted in the increase in the ***health expenditure*** to those countries affected by it. Hence the ***life expectancy*** kept declining even as ***health expenditures*** increased in these countries most affected by HIV and tuberculosis from 2000-2010.



To further investigate, we took a cohort of those countries from Africa which had ***percentage expenditure*** greater than 7.5 and ***life expectancy*** lesser than 50 between 2000-2005 and saw how it varied across 15 years.

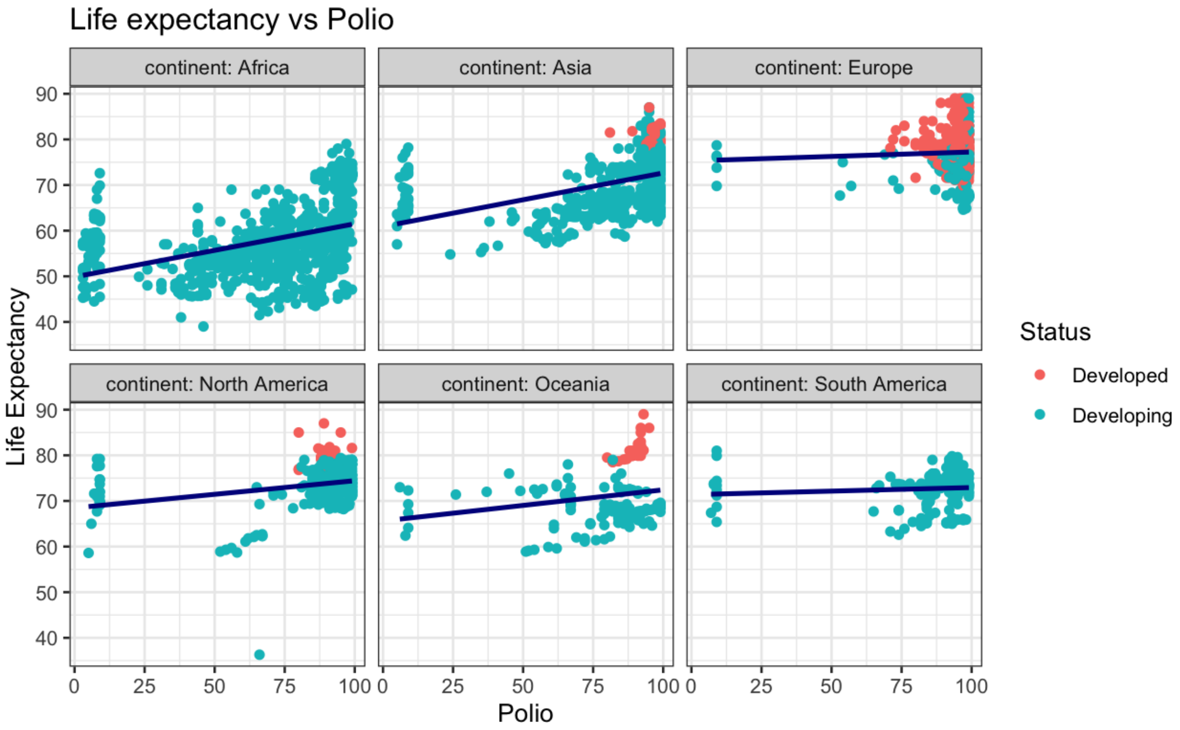
As this graph illustrates, even with the increase in ***percentage expenditure*** over the years, ***life expectancy*** has remained the same in most of the countries like Zimbabwe and Botswana.

We did some research and determined that these are the countries that were most affected by HIV, as you can see on the map below from the World Bank.



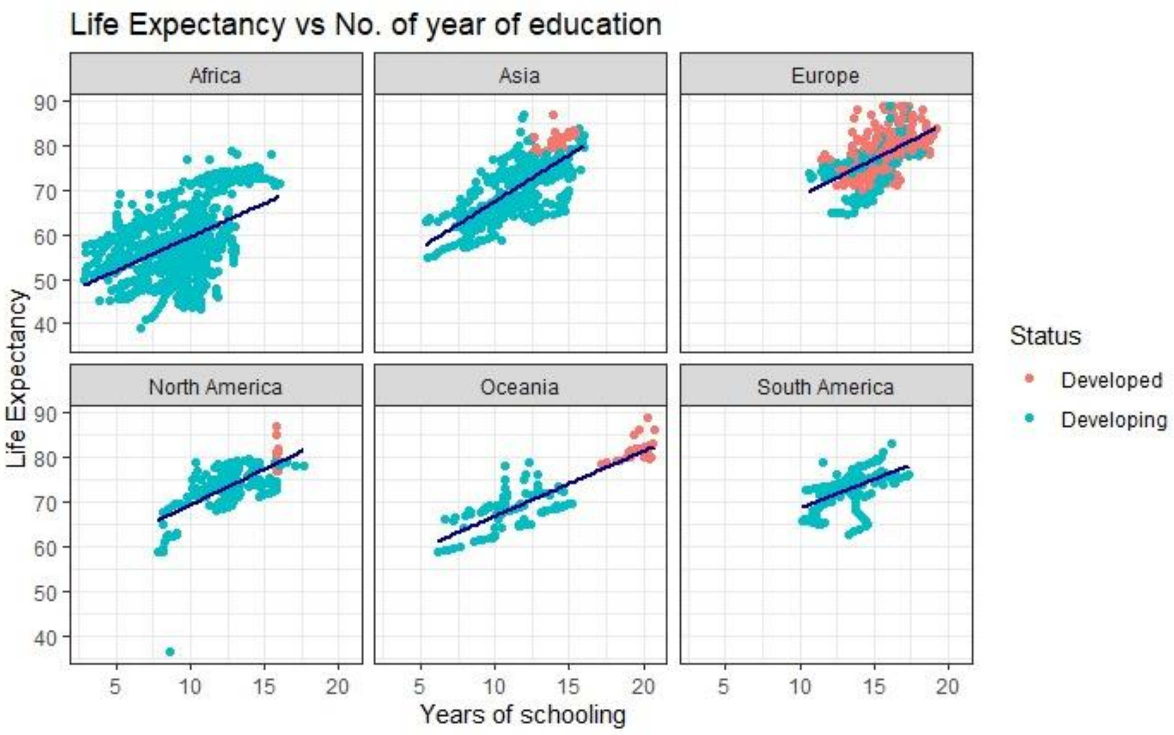
Prevalence of HIV/AIDS in Africa, total (% of population ages 15–49), in 2011 (World Bank)

Life expectancy vs Polio:



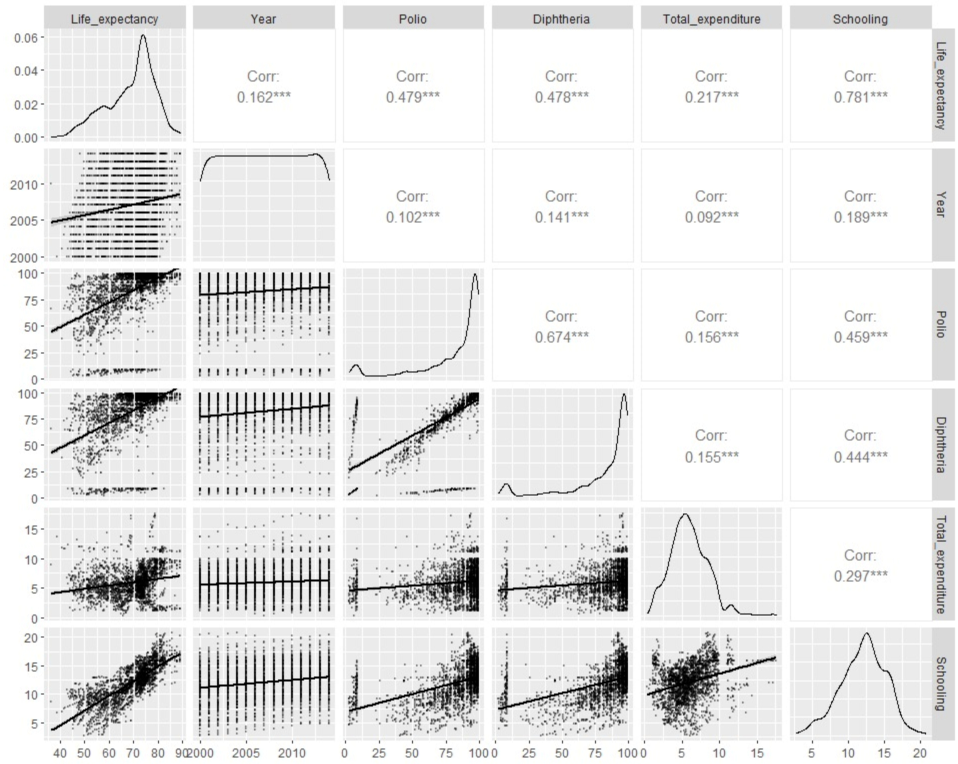
Over the years we can see that, with the increase in the ***Polio vaccinations***, the life expectancy grew. This effect is more evident in the continents of Africa and Asia where we have a higher number of developing countries. In most of the developed countries like those present in Europe and North America we already have a ***Polio vaccination*** percentage of above 75 and hence we don't see much of a trend here.

Life expectancy vs Schooling:



Every continent showed a positive trend with respect to ***schooling*** and as we can see ***schooling*** is highly correlated to ***life expectancy.***

Correlation of the variables:



- ***Life expectancy*** has a high correlation with ***schooling***.

- ***Polio*** and ***diphtheria*** are equally correlated with ***life expectancy***. They both have a high correlation among themselves too. Hence, it's better to only use one of the two columns for our modelling purpose.

We also tried to look into another relationship between the percentage change in the ***life expectancy*** and the percentage change in % expenditure and the other variables as well. But we did not get any useful insights out of it.

MODELING: -

The base model consisted of all the features available in the dataset. This model is a Linear-Model with no interaction between the features. The included features in this model were: ***Population, BMI, Year, Total Expenditure (on healthcare), Continent, Polio, Diphtheria, Status,*** and ***Schooling***. The purpose of this model was to get a picture of how all the features affected Life Expectancy, or whether they affected the Life Expectancy at all. The features with lower values of co-efficients were eliminated from further analysis. As these features have little to no effect on the Life Expectancy. The features eliminated, based on their co-efficients in the model, were ***Population, Year***. Another feature was eliminated i.e., ***BMI*** on the basis of having a large number of null values, which affected the overall predictive power of the model. Population reason ? The reason the Year is not such a good predictor could be that, even though the Life Expectancy is increasing as the years go by, but there are a lot of underlying factors at play. Hence there is much more to the story rather than just Life Expectancy simple increasing by the years. Diphtheria was also removed from the final model as it is highly correlated with Polio (r=0.7).

After eliminating the above-mentioned features, different combinations of models with interaction between different features was tried. Based on the lowest AIC score amongst these models, the model selected was with predictors: ***Polio, Total Expenditure, Schooling, Continent, Continent * Total Expenditure, Continent * Schooling*** … (*where x1 * x2 signifies interaction between features x1 and x2*). The r-squared score for this model was 0.75, which indicates that this model is a good enough model for the purpose of predicting Life Expectancy. The included interactions are essential as they are self-explanatory. ***Continent*** provides us with a generalized sense of how other factors affect Life Expectancy and how they vary geographically. This is extremely important to our analysis as earlier we saw that the ***Total Expenditure*** varied by each ***Continent***, and the effect of ***Total Expenditure*** on ***Life Expenditure*** was different for each ***Continent.*** Hence the interaction term between them bolsters the model predictive power. Apart from this Schooling is a very important feature as it is very highly correlated with ***Life Expectancy***. ***Schooling*** again varies by each ***Continent***; this is mostly based on the fact if the country is Developing or Developed. As we saw for countries with higher Schooling, the Life Expectancy was higher as well, and these were mainly Developed nations, however some of the Developing countries were a part of this bracket as well. This further affirms the strong effect of the years of ***Schooling*** on the overall ***Life Expectancy.*** And ***Schooling*** was different for countries based in different continents, for instance, there was a divide in the years of education if we compare countries in North America to those in Africa.

But to further explore how varied of an effect did ***Schooling***, as well as ***Total Expenditure***, had on ***Life Expenditure*** we decided to generate two models trained with the same structure but only on the subset of data of years 2000 and 2014. This is because, 2000 to 2014 would be a large enough time-period to notice differences between ***Life Expectancy*** and all the factors affecting ***Life Expectancy***. And comparing the results from both these models would help us have a contrastive comparison on how things changed between these years.

After training the models for years 2000 and 2014, the goal was to check the trends for the fitted values for the two models. Dummy data was simulated for both years, by generating combinations of values for the predictors. Using these simulated values, the values were fed to the model to predict the **Life Expectancy**. Using these predicted values, the following plots were used to compare the results of the two models for each of the metrics (i.e., ***Schooling, Total Expenditure*** & ***Polio)***.

a. Schooling:

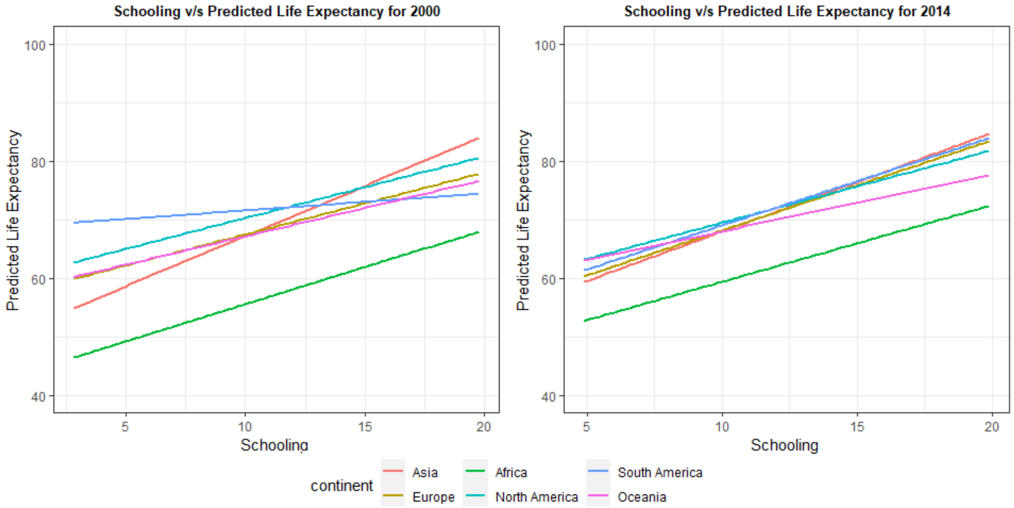


Figure 5.1: Schooling and predicted Life Expectancy

In Figure 5.1, we can observe that in both the years 2000 and 2014, there is a strong positive correlation between Schooling and Life Expectancy, for each continent. But for the year 2000, there is a lot more variation between the continents as compared to year 2014. As for Africa the trend has just shifted upwards from 2000 to 2014. One reason explaining this variation could be that, as for the recent years, the emphasis on education has increased as compared to the earlier years. This impact of more emphasis on education makes the feature ***Schooling*** a much stronger and direct contributor to a better ***Life Expectancy*** value for every country, for the year 2014. This explains the uniformity that can be noticed for every continent for that year. In other words, we can say the for the recent years, Schooling is directly, strongly and positively correlated to Life Expectancy, and this is the case with every continent.

b. Polio

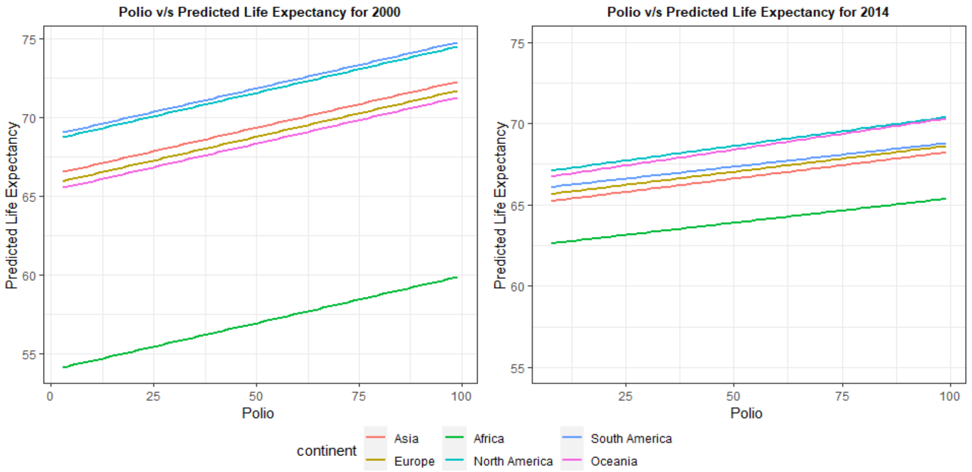


Figure 5.2: Polio and predicted Life Expectancy

In Figure 5.2, we can see that the ***Polio*** vaccination rate has always had a positive correlation with life expectancy. Also, the overall ***Life Expectancy*** for Africa has gone up by 10 years. The life expectancy of Asia also increased by nearly 5 years. We can say that immunizations always have had a strong effect on ***Life Expectancy***.

c. Total Expenditure:-

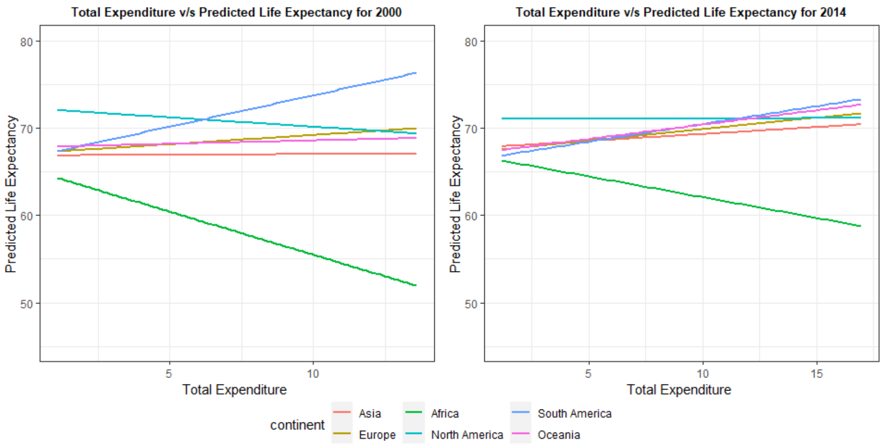


Figure 5.3: Total Expenditure and predicted Life Expectancy

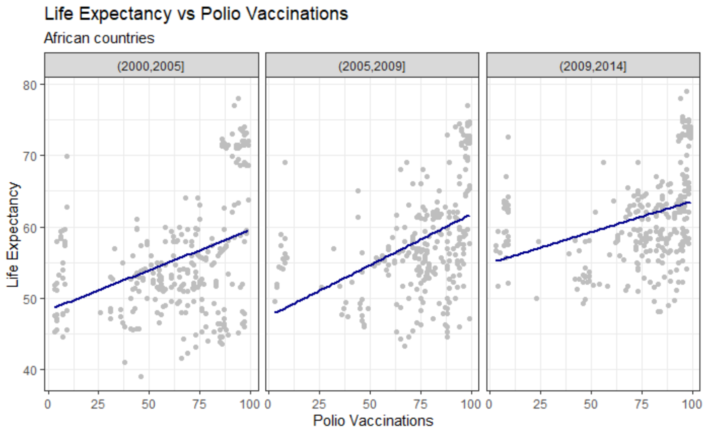
In Figure 5.3, for the year 2000, North America has a slight downwards ***Total Expenditure*** trend, as opposed to other continents, except Africa. However, the trend for North America is quite flat. As for Asia, the trend goes from flat to slightly increasing from the years 2000 to 2014. Africa, as we saw before, is a special case here. The trend is downwards for both the years 2000 and 2014. But it can be observed that for the year 2000, Africa has a much steeper trend as compared to the year 2014, signifying the increase in Life Expectancy in Africa, amongst the two years 2000 and 2014. However, the trend is still interestingly downwards. This can be attributed to our hypothesis earlier in the EDA section, where even with the higher amount of healthcare expenditure, the prevalence of HIV/AIDS affects Life Expectancy inversely.

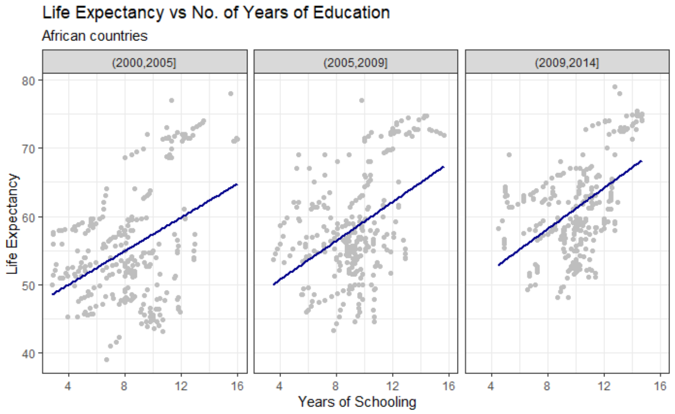
CONCLUSION, FUTURE WORK & LIMITATION:-

- As the years increased, the ***life expectancy*** for all countries increased, but the growth is substantially high in the African countries.
- ***Healthcare expenditure*** plays an important role in positively affecting the ***Life Expectancy*** in most countries, except the African countries.
- Factors like ***Polio Immunization*** rates and Schooling affect ***Life expectancy*** in a positive way.
- ***Life expectancy*** is also dependent on the geographical location of the country i.e., ***the continent***.
- Whether the country is a developing country or a developed country also plays an important role in determining ***life expectancy*** (as life expectancy is high for all the developed countries).
- HIV There are other factors which affected the ***life expectancy*** in many African countries over the years and one major factor we could see was the ***HIV disease.*** Accumulating more and accurate data with regards to HIV, GDP, alcohol consumption, BMI. We can consider air quality index as well.

APPENDIX:-

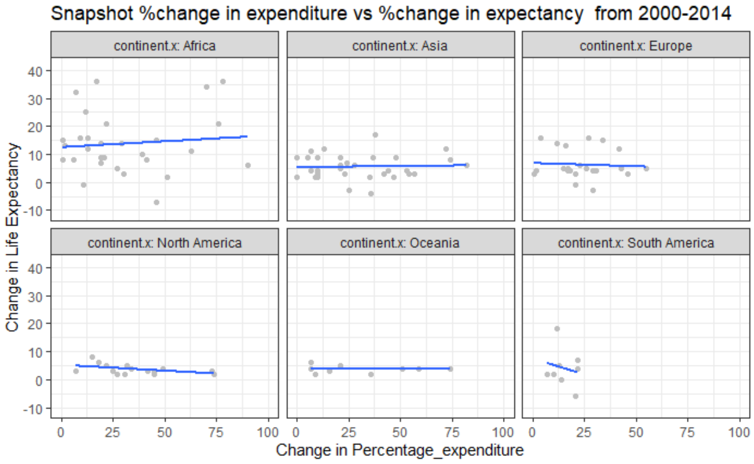
LIFE EXPECTANCY VS POLIO & SCHOOLING FOR AFRICAN DATA:-





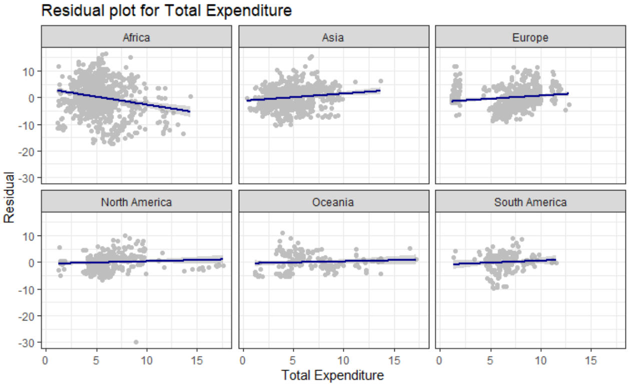
we divided the 15 years African data into buckets of 5 years to see the trend. The above plots, show the trends of polio with life expectancy and schooling with life expectancy. We can see that there is positive correlation in every bucket, so, there isn't much change among the buckets.

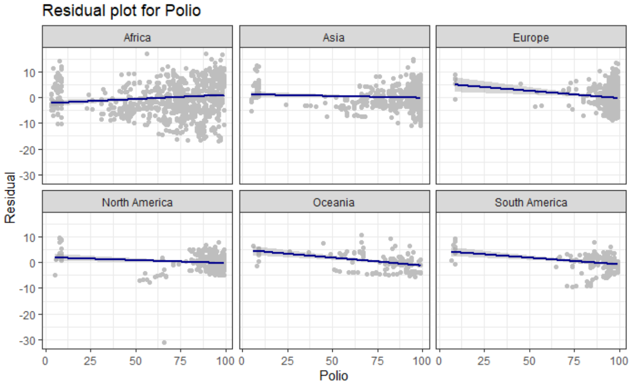
PERCENTAGE CHANGE IN EXPENDITURE VS PERCENTAGE CHANGE IN EXPECTANCY:-

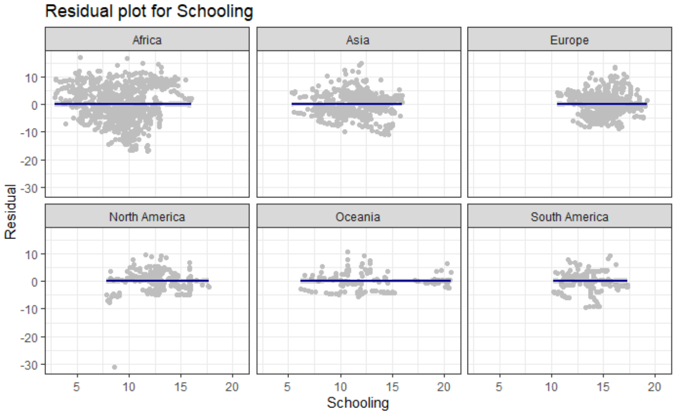


The above plot shows the trend between percent change in expenditure v/s percent change in expectancy. There is not any interesting trend between change in expenditure v/s percent change in expectancy.

RESIDUAL PLOTS OF THE MODEL:-

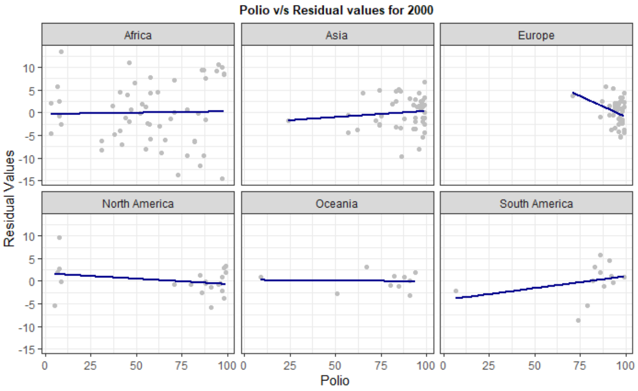


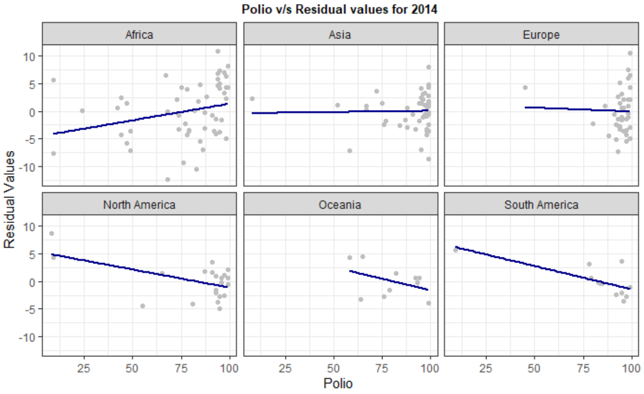


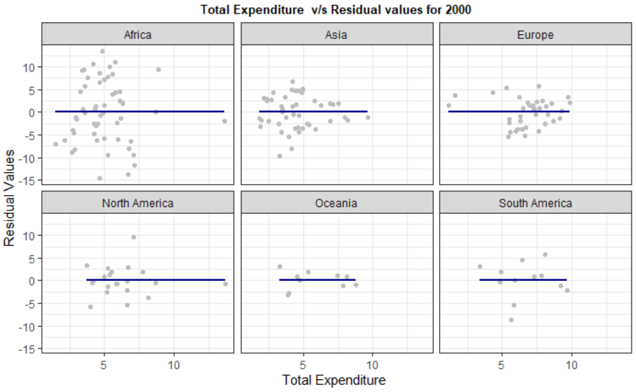


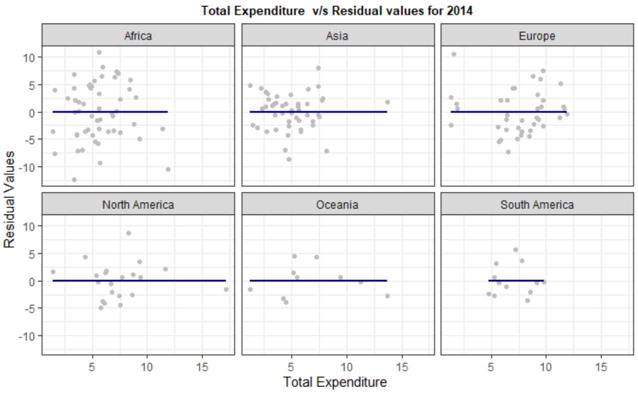
The above plots are the residual plots of the model that we used. We can see that, for ***Polio*** and ***Schooling*** the line is nearly straight, but for ***total expenditure***, we can see a downward line for ***African continent.***

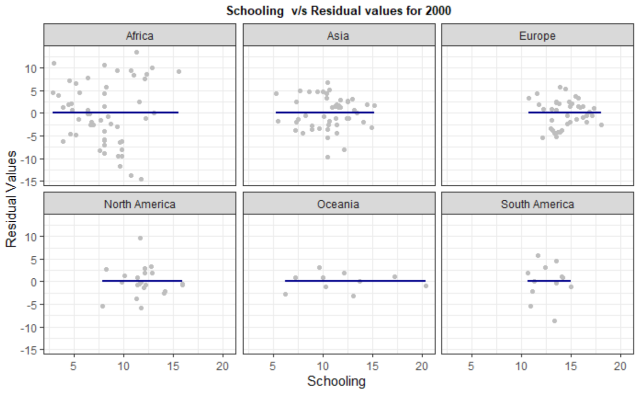
COMPARISION FOR RESIDUAL PLOTS OF 2000 MODEL and 2014 MODEL:-

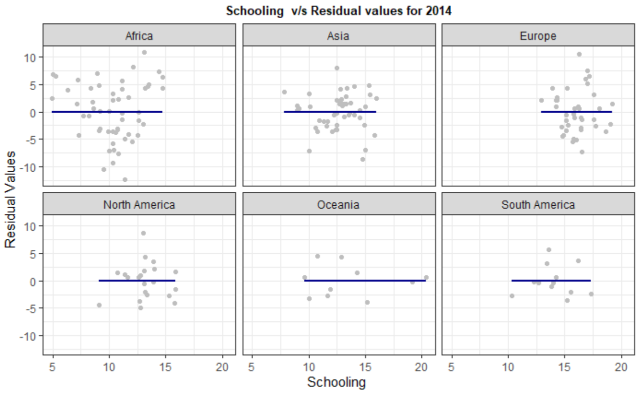












From the above plots we can see that there is randomness in the residual plots. We can observe this for both the years 2000 and 2014 , for both the features, schooling and Expenditure, but Polio is following a trend and it does not show that kind of randomness.