Life Expectancy (WHO)

Report

Group members:

* Cyril Avevor
* Cyril Yamoah
* David Afo-Kangah
* David Gebe

Contents

[Introduction 3](#_Toc47817266)

[Context 3](#_Toc47817267)

[Content 4](#_Toc47817268)

[Acknowledgement 5](#_Toc47817269)

[Hypothesis 6](#_Toc47817270)

[Questions 7](#_Toc47817271)

[Processes 8](#_Toc47817272)

[Data cleaning 9](#_Toc47817273)

[Data Exploration and visualization 13](#_Toc47817274)

[Conclusion 18](#_Toc47817275)

# Introduction

## Context

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not considered in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio, and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors, and other health related factors as well. Since the observations in this dataset is based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting to a country which area should be given importance to efficiently improve the life expectancy of its population.

## Content

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. The dataset related to life expectancy among other health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project considered data from years 2000 - 2015 for 193 countries for further analysis. The individual data files have been merged into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B, and GDP. The missing data were from countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult due to data restrictions through bureaucracies and hence, it was decided that we exclude these countries from the final model dataset. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors.

## Acknowledgement

The data was obtained from WHO and United Nations website with the help of Deeksha Russell and Duan Wang.

## Hypothesis

Immunization and Human Development Index increased the life expectancy for people in all countries in the world between 2000 and 2015.Immunization and Human Development Index increased the life expectancy of people in all countries in the world between the years of 2000 and 2015.

## Questions

This project was undertaken to find answers to the following questions:

1. Did the status of a country affect its life expectancy?
2. Did the life expectancy trend between 2000 and 2015 rise/ fall?
3. Were adult mortality and alcohol co-related?
4. Did GDP and income composition of resources affect life expectancy?
5. Did Hepatitis B, Polio, and Diphtheria affect life expectancy?
6. Did countries that spend more on health care have a higher life expectancy than those that do not?
7. Did schooling affect life expectancy?

# Processes

Data from the field is mostly messy and this dataset was no exception. To achieve optimum visualization, it went through the following processes:

* Data Cleaning
* Data exploration visualization

## Data cleaning

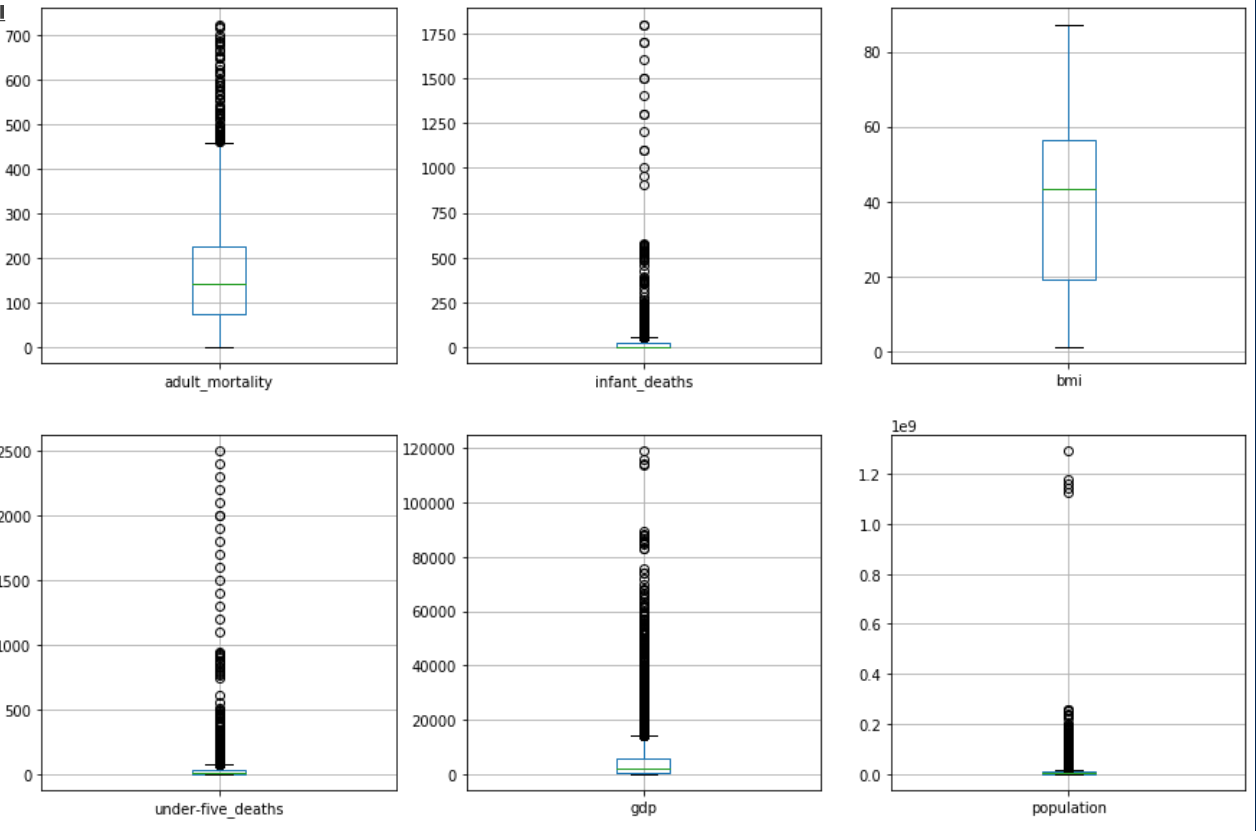
The columns in the table include:

* country - the country in which the indicators are from (i.e. United States of America or Congo)
* year - the calendar year the indicators are from (ranging from 2000 to 2015)
* status - whether a country is 'Developing' or 'Developed' by WHO standards
* life\_expectancy (Ratio) - the life expectancy of people in years for a country and year
* adult\_mortality (Ratio) - the adult mortality rate per 1000.
* infant\_deaths (Ratio) - number of infant deaths per 1000 population; like above, but for infants
* alcohol (Ratio) - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
* percentage\_expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (GDP)
* hepatitis\_b (Ratio) - number of 1-year olds’ with Hepatitis B immunization over all 1-year olds in population
* measles (Ratio) - number of reported Measles cases per 1000 population
* bmi (Interval/Ordinal) - average Body Mass Index (BMI) of a country's total population
* under-five\_deaths (Ratio) - number of people under the age of five deaths per 1000 population
* polio (Ratio) - number of 1-year olds with Polio immunization over the number of all 1-year olds in population
* total\_expenditure (Ratio) - government expenditure on health as a percentage of total government expenditure
* diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1-year olds
* hiv/aids (Ratio) - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
* gdp (Ratio) - Gross Domestic Product per capita
* population (Ratio) - population of a country
* thinness\_1-19\_years (Ratio) - rate of thinness among people aged *10-19* (Note: variable should be renamed to *thinness\_10-19\_years* to represent more accurately the variable)
* thinness\_5-9\_years (Ratio) - rate of thinness among people aged 5-9
* income\_composition\_of\_resources (Rat - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
* schooling (Ratio) - average number of years of schooling of a population

Missing Values:

Detection: Finding the NULL values was done using df.info()

The dataset had various columns. These columns were visually inspected using boxplots to determine irregularities. Below is the boxplot of the uncleaned data.



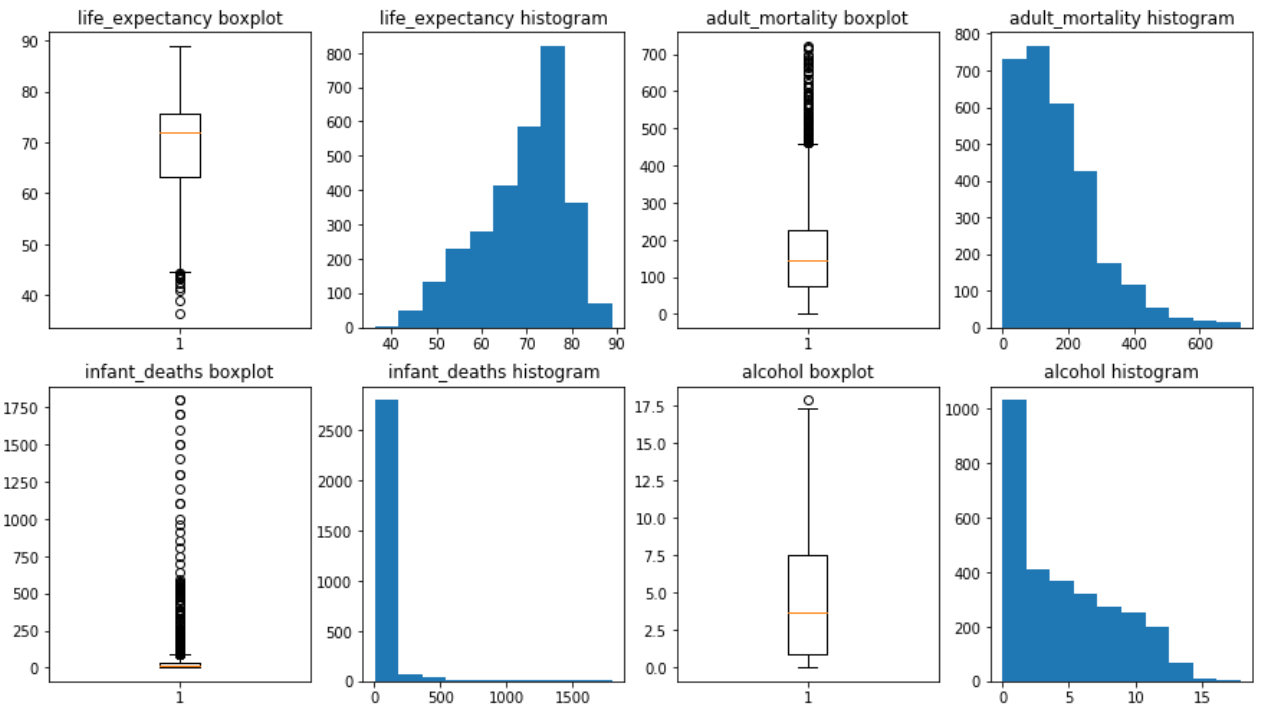
Upon inspection, the irregularities noticed were:

* The least adult mortality was 1
* The least infant deaths were 0
* There were BMIs below 10 and greater than 50
* The least under- five deaths were 0

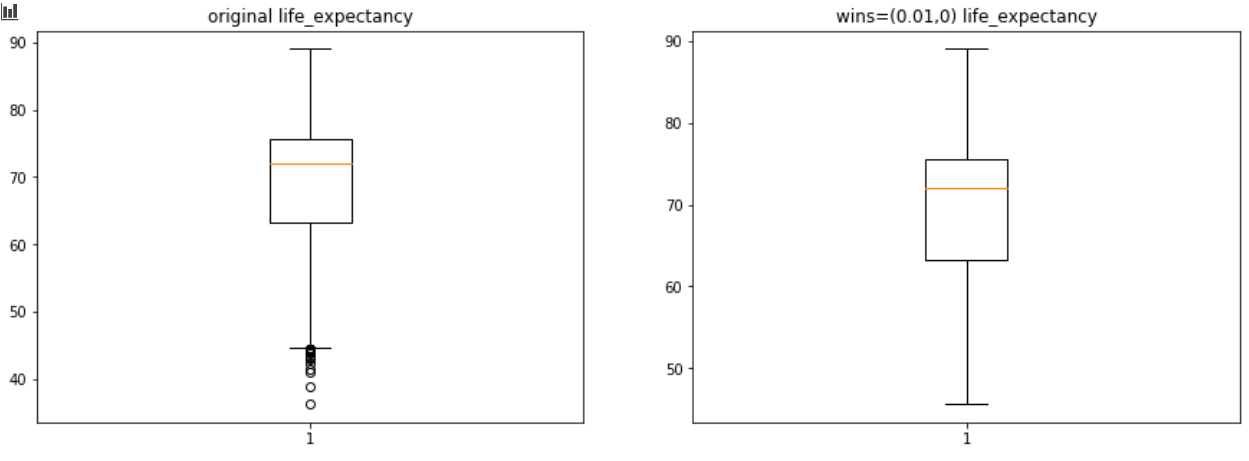
To resolve these and other issues the following were done:

* Replaced all whitespaces (‘ ’) in the column names with underscore (‘\_’).
* All characters in the columns were converted to lowercase. E.g. ‘Country’ to ‘country’.
* Renamed columns with wrong names.
* Infant\_deaths that were 0 were replaced with np.nan
* bmi values less than 0 and greater than 50 are replaced with np.nan
* All null values were replaced with values of the mean of that of that country’s year.

Each column of the dataset is then checked for outliers(values that deviate extremely from the mean). Boxplot and histograms are used to visualize this. Below is a screenshot of a portion of the outlier plot.



The goal was to eliminate the outliers and make the values evenly distributed. Using tukey’s method, the upper and lower bounds to the data was set (winsorize). The upper and lower limits of each column was set manually. The results are shown in the table below.



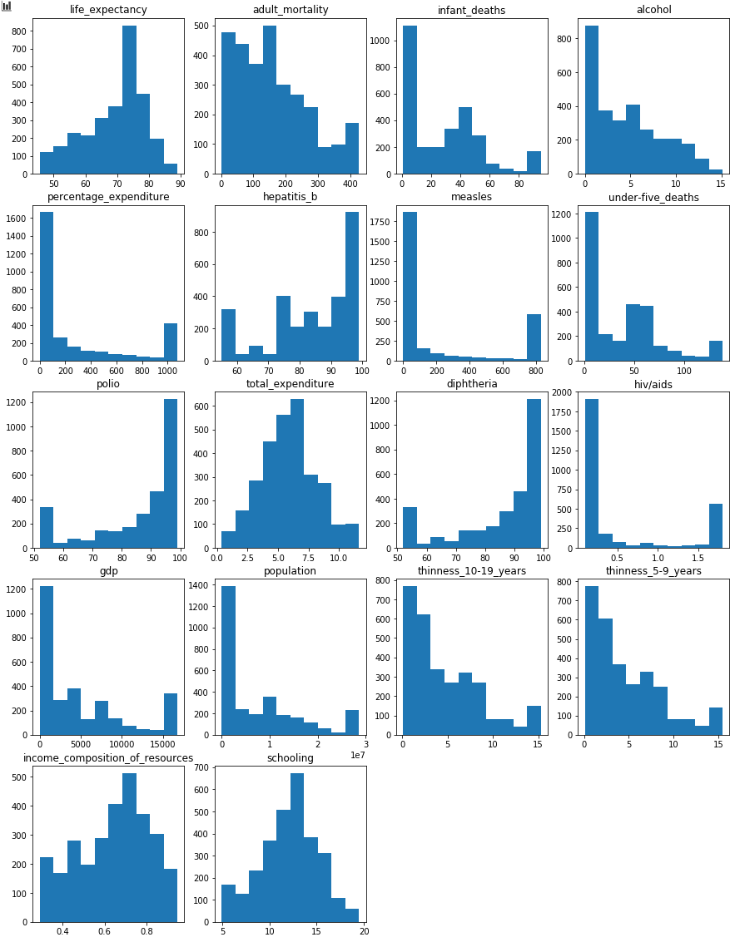
Values winsorized with upper and lower limits are more evenly spread than the original data thus making them ready for exploration and analysis.

## Data Exploration and visualization

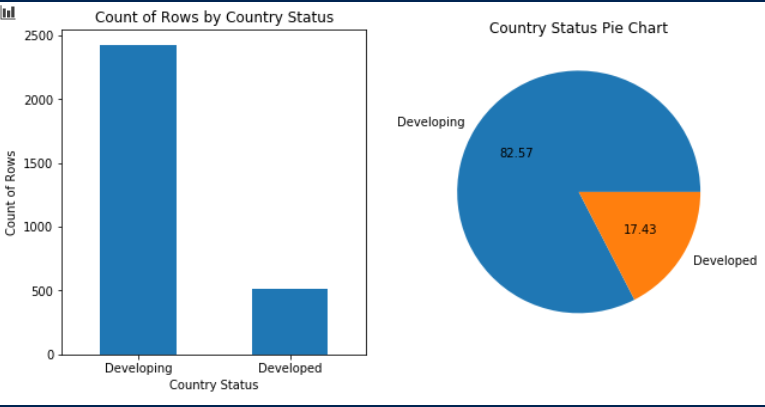
Univariate analysis was used to explore the various variables on their own.

A new data frame which has the winsorized data was thus created. For the continuous data, histograms and bar plots was employed. These categorical data included data such as status of a country.

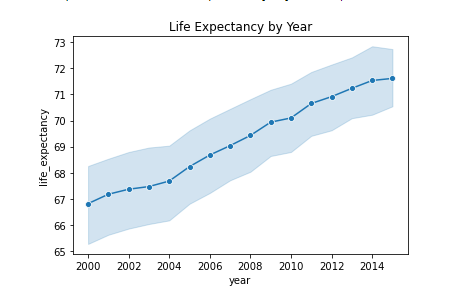
The various histograms generated are shown below.



The categorical column (status) is shown with a bar and pie chart below.

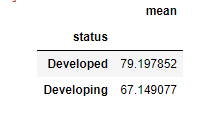


A line plot was used to analyze how life expectancy has changed over the years.



In the plot above, the lifespan of countries has increased over the past 15 years. This plot does not include which countries has had their life expectancy increased over the years.

Calculation shows that developed countries have a much higher average life expectancy than developing countries. Below is the output of the calculated average life expectancy of the status of countries.

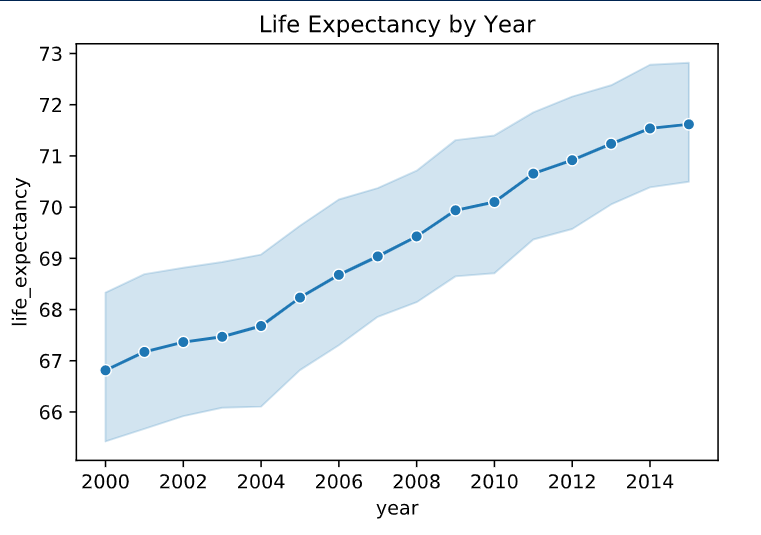


Based on the result of the above test, there appears to be a significant difference in the life expectancy of Developing countries and Developed countries

A study was done on the status of a country and their continuous variable. The result shows developed countries has the following:

* Low adult mortality rate
* Low infant deaths
* High percentage consumption of alcohol
* High percentage expenditure
* High availability of hepatitis B, polio

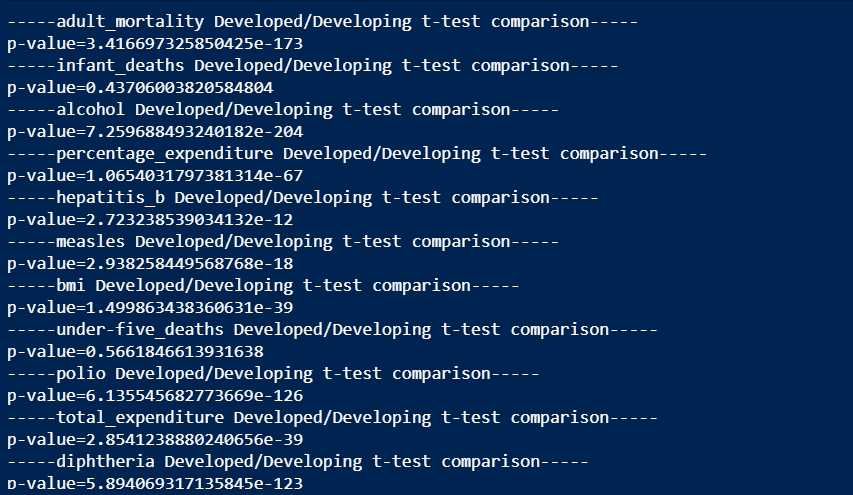
For the life expectancy, the graph is shown below

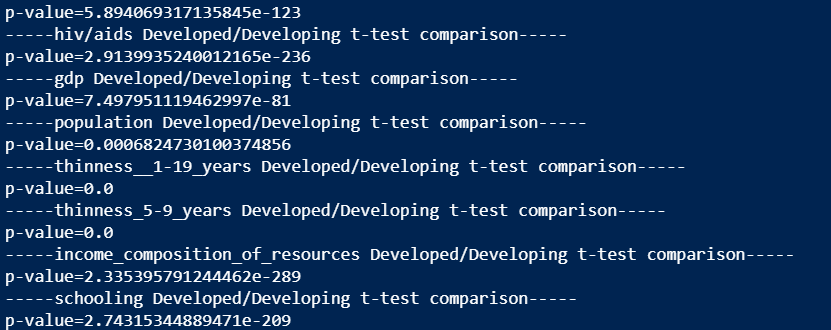


The correlation between the year and life expectancy is checked using the *.corr()* function in winsorize module. From the output, it can be concluded that year is not really correlated to the life expectancy of a country.

For status, the mean of all developed and developing countries are calculated. The mean of the developed countries is **79.19** and that of the developing countries is **67.14**.A t- test using the 2 status with respect to their life expectance to check the significance. A p-value of . A low p- value indicates a high correlation between 2 or more values.

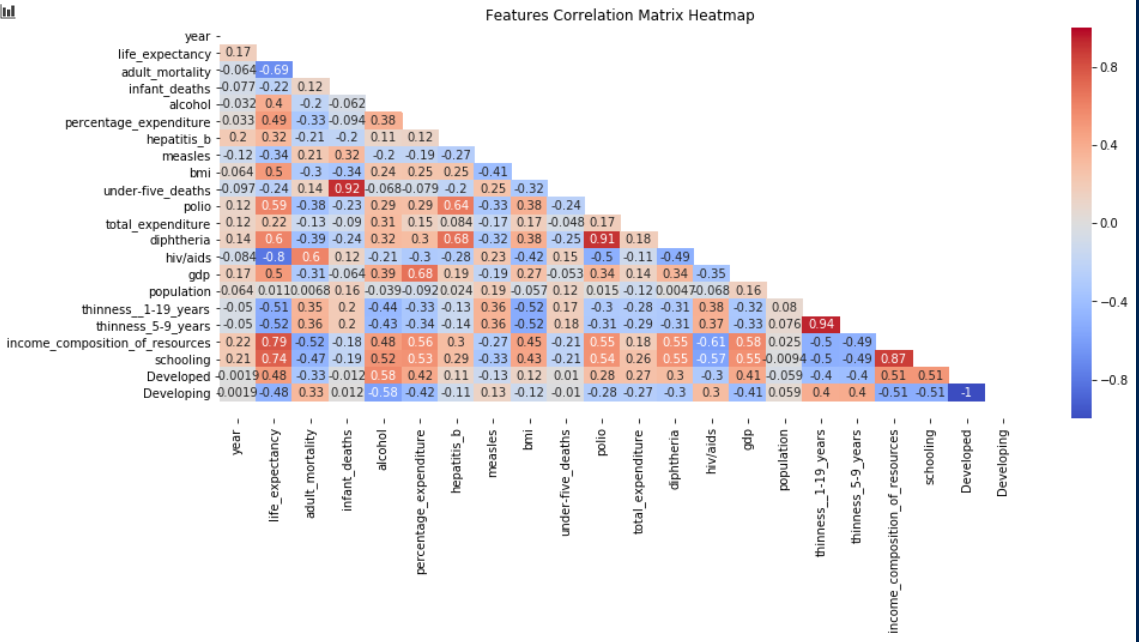
The various means of the remaining columns are calculated with respect to the status for both developed countries and developing countries and then their various p-values are calculated. The various p-values are shown below:





The correlation between the various column are then calculated and used to create a heat map. For developed countries, the heat map show values greater than zero whilst for developing countries, it shows values less than zero. The further away a value is from 0 the more correlated it is to another value from the table. Any correlation factor above 0.7 or below -0.7 was considered correlated anything else was not.

Below is the heat map



From observation, it can be deduced that:

* Under-five deaths are highly correlated to infant\_deaths
* Diphtheria is highly correlated to polio
* Thinness among 10-19-year olds is highly related to thinness among 5-9-year olds.
* The life expectancy was not so dependent on year between 2000 and 2015
* Hiv/aids, income composition of resources, schooling were the correlated factors of life expectancy.
* Vaccinated diseases (polio, diphtheria, and Hepatitis B) were not so correlated to life expectancy.
* Population does not correlate to life expectancy.
* Infant deaths and under- five deaths did not influence the life expectancy as much.
* Percentage expenditure is not significantly correlated with life expectancy.

# Conclusion

## Answers

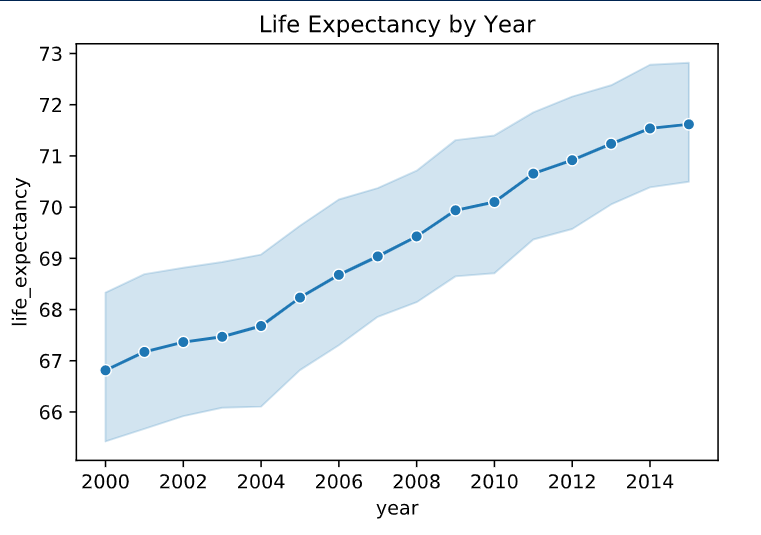
At the beginning of this report some questions were posed and now the answers have been derived.

1. **Does the status of a country affect its life expectancy?**

From the winsorized data, it can be concluded that the status of a country does heavily impact the life expectancy of a country. Low p- value was achieved after the t- test was run to prove this. The life expectancy of people in a developed country was higher than that of those in developing countries.

1. **Did the life expectancy trend between 2000 and 2015 rise/ fall?**

Between the years of 2000 and 2015, the mean life expectancies of all the countries kept rising as demonstrated in the graph below



1. **Were adult mortality and alcohol co-related?**

Looking at the heat map from earlier it can be realized that alcohol had a low correlation to adult mortality which means countries that consumed more alcohol did not have more adults dying.

1. **Did GDP and income composition of resources affect life expectancy?**

From the heat map, it can be noticed that GDP does not have a significant correlation with life expectancy. The opposite can be said of income composition of resources as it had a high correlation with the life expectancy of a country. Therefore, we can say that when the human development index is high the life expectancy of a country is also high.

1. **Did Hepatitis B, Polio Diphtheria affect life expectancy?**

From the heat map, we see that Hepatitis B is the least correlated to the life expectancy followed by polio and then diphtheria (hepatitis b< polio< diphtheria). Diphtheria being the highest still fails to meet our threshold for a column to be considered correlated to the life expectancy.

1. **Did countries that spend more on health care have a higher life expectancy?**

From the heat map, we can see that percentage expenditure is not correlated to life expectancy, but it does not seem far away.

1. **Did schooling affect life expectancy?**

From the heat map, it is evident schooling was highly correlated to life dependency.

## Summary

The hypothesis at the start of this project was ‘Immunization and Human Development Index increased the life expectancy for people in all countries in the world between 2000 and 2015’. This hypothesis was both true and untrue in the sense that immunization (indirectly through vaccinated diseases such as polio, hepatitis b and diphtheria) was not correlated to life expectancy. Human development index (through income composition of resources) was highly correlated to the life expectancy.