Visual Analytics for Health Data

Statistical Analysis on factors influencing life expectancy across the globe.

Sai Pramod Kudapa

112686280

Ghanendra Piplani

113070787

Video Link

INTRODUCTION

Life expectancy is defined as the average time an organism is supposed to survive, based on various factors such as the year of its birth, current age and health conditions, the environment it lives in among others. Through this project we try to find the factors which affect life expectancy the most, either positively or negatively, using which a country can decide the correct way to spend it's resources (both man and money).

This report is segregated into 3 parts, first we discuss the objectives of our project, followed by a discussion on the approach used wherein we talk about the dataset, coding techniques and visualizations used to represent the data, finally we discuss in detail the observations we found after visualizing the data.

Objectives

We formed the following hypothesis:

1. Higher literacy rate translates into a higher life expectancy.
2. Higher expenditure on healthcare leads to a higher life expectancy.
3. Higher adult mortality leads to lower life expectancy.
4. Higher alcohol consumption leads to health hazards and consequently a lower life expectancy.
5. High population density leads to a lower life expectancy in a nation.

We aim to find the factors which highly influence the life expectancy of a nation so that the leaders of a nation can decide where to invest the manpower and money in order to increase life expectancy.

Approach

1. **Data:** We carefully picked a dataset from kaggle which would help us in answering our hypothesis. Our primary dataset consists of features like:
    a. **Country:** Name of the country.
    b. **Year:** Values of years ranging between 2000 and 2015.
    c. **Status:** This tells whether a country is developed or developing.
    d. **Life expectancy:** This is the average life expectancy across the world.
    e. **Adult Mortality:** Tells the adult mortality of the country.

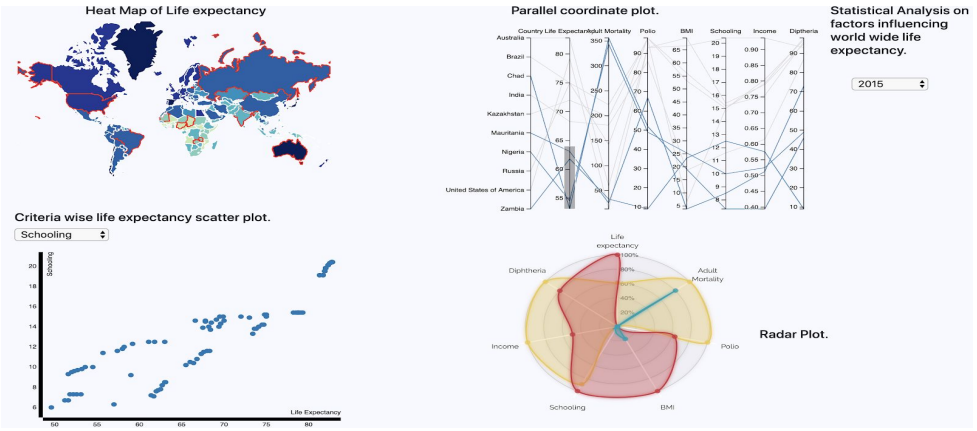
f. **Infant Mortality:** Tells the infant mortality rate of the country.
g. **Hepatitis B:** The rate of hepatitis B vaccination rate.
h. **Polio:** The rate of Polio vaccination in the country.
i. **Diphtheria:** The rate of Diphtheria vaccination in the country.
j. **Percentage of GDP expenditure on health:** Tells the expenditure of a country on health services as a percentage of its GDP.
k. **Alcohol:** The net alcohol sold in the country.

The dataset has values from 194 countries and data from 2000 to 2015 which makes up a total of 2938 rows. To begin with we subset the data by removing the years from 2000 to 2007 as we wanted to visualize the data from 2008 to 2015. Next we cleaned the data by replacing null values by taking the mean of the values. We merged our dataset with the dataset of countries with details of the country code and latitude, longitude. Our secondary dataset consists of a world map to plot a heatmap of life expectancy distribution across the globe.

2. Coding Techniques:

a. We used Python-3 and the Flask webapp framework to create our application's ***back-end***. The main code exists in the file app.py which has the "__main__" function, it has routes defined which help in navigation across the application on the webpage.

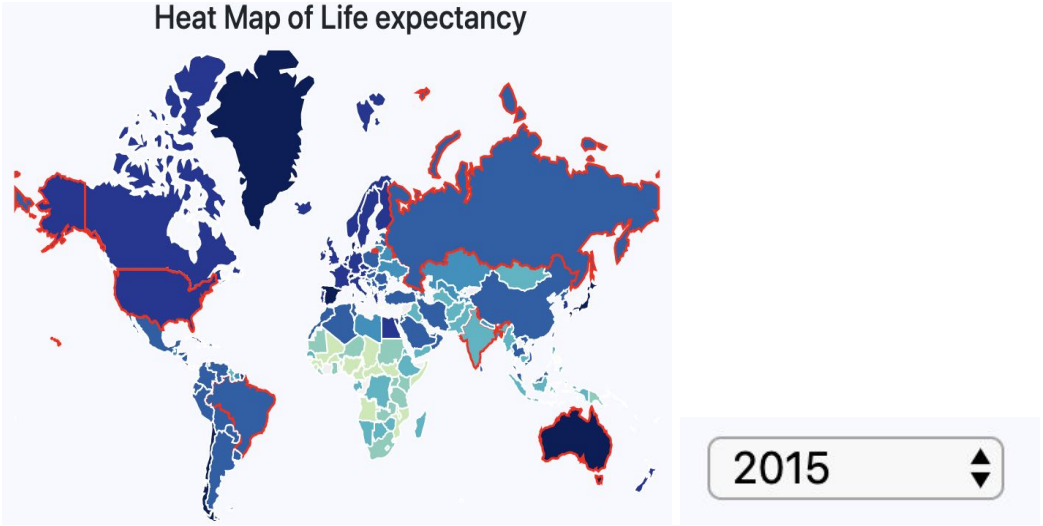
b. A *Single Page Application* architecture is followed. To achieve this we used D3 Version3 and HTML/CSS to create the ***front-end***. AJAX (Asynchronous JavaScript and XML) queries are used to communicate between the back-end and front-end.



Visualizations:

1. **World Map**

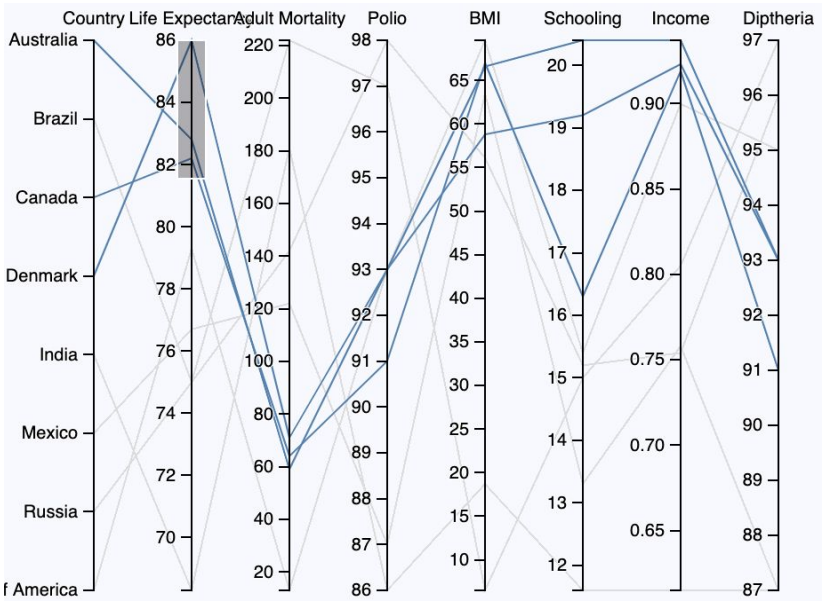
In order to show the geographical distribution of life expectancy in years across the globe, we use a map to plot the country wise life expectancy values. We utilise this map for the purpose of country selection.



Map serves as the primary entry point for interaction enabled across the dashboard. The countries with the red outlines are the selected countries, the data points associated with them are used in our visualizations.

2. **Parallel Coordinates**

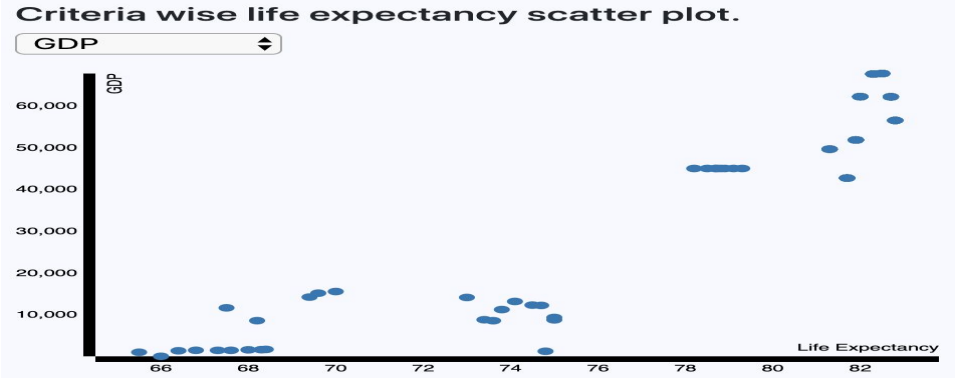
We have chosen Adult mortality, Polio, BMI, Schooling, Income as our dimensions in our parallel coordinates plot. These attributes encompass diverse aspects affecting human health such as literacy rate, healthcare system, government expenditure on health care. The order of Adult mortality, Body Mass Index values, Schooling trends, Income composition of resources, Diphtheria immunization was chosen by trying different combinations. The chosen combination showed the correlation trends in the best possible way.



A bowtie(X) shape can be observed between life expectancy and adult mortality suggesting a clear negative correlation. Using the brushing technique, upon selecting the countries with higher life expectancy tend to have lower adult mortality rates, higher values in polio and Diphtheria immunization rates, BMI, Schooling and Income.

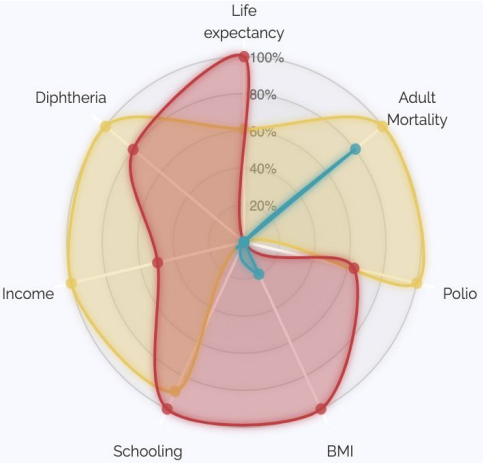
3. ScatterPlot

Using a scatter plot, we show correlation trends between various factors and life expectancy. The points in the scatter plot represent the country's(from Map) values in the selected year. We can observe a clear positive correlation between GDP and Life expectancy from the above scatterplot.



4. **Radarchart**

   We have chosen a radar chart to compare Life expectancy and other factors from 2008 to 2015. The data was first scaled to make it suitable for comparison before visualization.



   For 2015, we can see a comparative plot between Russia, USA and India. These three countries were specifically chosen to represent a high, medium and low life expectancy data point.

Conclusions:

We have summarized the major findings as mentioned below.

1. We found a strong positive correlation between literacy rate and life expectancy, implying that a country with high education tended to have a higher average life expectancy, this supports the intuition that better education leads to better life choices and hence healthy lifestyle.
2. A country with higher per-capita expenditure on healthcare tends to have a higher life expectancy.
3. Adult mortality rate represents the probability that a 15 year old person will die before reaching his/her 60th birthday, a country with a high adult mortality rate results in a lower life expectancy for the country.

4. Alcohol consumption did not seem to have a clear negative correlation as alcohol consumption was higher in developed countries in which factors like schooling, income and gdp tended to be high.
5. Higher population density was found to be inversely correlated with life expectancy as it was mostly observed in under developed countries lacking adequate resources.