Data Management Systems

Course Number: 610

Course Term: Summer, 2020

Assignment number: 5

Project title: Team Analysis Assignment using Cognos Analytics

Contact Information:

Team 3

Lydia Waribo Email: [Lwaribo@gmail.com](mailto:Lwaribo@gmail.com) Phone: 202-997-8967

Tue Le Email: [tuele.1111@gmail.com](mailto:tuele.1111@gmail.com) Phone: 407-443-9934

Gerald Mcdonald Email: [mcdonaldg09@gmail.com](mailto:mcdonaldg09@gmail.com) Phone: 706-830-5255

Randall Brown Email: [Randall.Brown0001@gmail.com](mailto:Randall.Brown0001@gmail.com) Phone: 703-705-8458

Melissa Hunfalvay Email: [melissa@righteye.com](mailto:melissa@righteye.com) Phone: 240-357-0026

Instructors Names:

Dr. Moretto

Dr. Goldberg

Introduction

The data set chosen for this assignment is Life Expectancy data. This data set represents a multi-decade database of life expectancy from the year 2000 to 2015 across 193 countries throughout the world. The original data set, prior to any cleaning, contains 2,938 number of cases or rows of data. There are 22 columns (variables) of data (see Figure 1). Variables include strings, floats, Boolean and integers.

The dataset provides the following variables on life expectancy that may be used to develop predictive models; country (count; 193), year (the year in which the data is reported, ranging from 2000-2015), status (developing or developed country), country (name of country), life expectancy (averaged per year). Total counts for the following factors were also reported by year: adult mortality, infant deaths, Hepatitis B, Measles, BMI, deaths of children under the age of 5, Polio, Diphtheria, HIV/AIDS, thinness, (1-19 years and 5-9 years) and population. Other values were reported as percentages for example, percentage expenditure and income compositions. A maximum value for the schooling variable was reported.

After cleaning, refining, and modifying the data (outlined in the next section) there were 2,810 number of cases or rows of data. Twenty columns (variables) of unique values, with 193 countries life expectancy data remained. Therefore, this was a significant amount of unique data to make the dataset robust for exploration (see Figure 2).

This dataset is an important topic as it explores factors related to longevity of life. Insights from such data may help to inform global resources, like the International Finance Corporation (IFC; [https://www.ifc.org](https://www.ifc.org/wps/wcm/connect/corp_ext_content/ifc_external_corporate_site/home)) to undeveloped countries (Wallace, 2018). Furthermore, insights from such data may create awareness of important factors impacting health, wellness, and quality of life (Wallace, 2018).

Data Refinement

Given that predictive models are better developed with larger data sets that have many cases and possible inputs from which to select (Fortuny, Martens & Provost, 2014), only minimal and obvious data cleaning was conducted. The following data cleaning strategies were employed. Some factors were changed to whole numbers, for example, life expectancy. Several variables were filtered to include one decimal place, for example HIV/AIDS. Total expenditure and income composition factors were cleaned to show two decimal places for each value. A thousand separator was used on the population variable to make it easier to read.

The dataset had few null variables and therefore no variables were removed based on lack of data. Where null variables were found, such as in the total expenditure column, these rows of data were removed from the dataset.

After some initial sanity checks of the data it became obvious that there were some errors in the dataset. GDP and population variables were incorrect. For example, in 2015 the population of the United States as reported in the dataset was 54,587,661. However, according to the US Census Bureau on January 1st, 2015 the US population was 320,090,857 (<https://www.census.gov/newsroom/press-releases/...>) Similarly, the GDP was reported in the dataset as 5258.92275, however, according to the Worldometer the US GDP in 2015 was, when adjusted for inflation the GDP was $16,710,458,234,000 or nominal GDP as $18,219,297,000,000 (<https://www.worldometers.info/gdp/us-gdp/>). Therefore, due to errors in these variables the GDP and population variables were deleted from the dataset.

Additional errors in the dataset were found via outlier analysis. For instance, Hepatitis B showed four cells with negative variables and were replaced with a Null variable. The country of Lesotho (Figure 3), had the highest number of HIV/AIDS at 9.3 which affects its Life Expectancy average at 53.7. The country of Zimbabwe, with its HIV/AIDS and Life Expectancy at 6.2 and 67, respectively. These outliers were noted and observed but not removed as there was no indication that the data was anything but legitimate.

HIV/AIDS variable is a skewed feature when looking at the total number of cases by year from 2000 to 2015 (Figure 4), there is a small increase in 2001 and 2002, followed by a downtrend from 2003 to 2015.

Data Exploration

Within the dataset, there are two broad categories of data: geographic location (e.g. country, status) and health of the individual (e.g. BMI, HIV/AIDS). Health conditions and geographic location appear to influence one another. For example, Polio, a disabling and life-threatening disease caused by the poliovirus, has a vaccine (<https://www.cdc.gov/polio/what-is-polio/index.htm>). However, the vaccine is only available to those from certain (developed) countries. With this reasoning in mind, a set of initial questions were posed:

* 1. Does the geographic location predict life expectancy?
  2. Do health conditions predict life expectancy?
  3. Is there an interaction between the geographic location and health condition that predicts life expectancy?

Insights into these initial questions were analyzed via two calculations. First, Life Expectancy Condition and Total Deaths were added to the dataset. Total Deaths was calculated by adding together Adult Mortality, Under-Five and Infant to reflect the health condition of the country. The purpose of this calculation was to gain further insights into question 2 by exploring the idea that large total death counts relate to poor health conditions and in turn a lower life expectancy.

Second, Life Expectancy was then split into three different categories (50 and below as Low, from 51 to 74 as Mid and 75 and above as High). These values were used to separate and compare the data between those countries that are low to high life expectancy. These calculations allow us to further understand the relationship of geographic location and health condition (question 1 and 3). Total Deaths has a relationship to life expectancy, specifically, higher Total Deaths equal lower in Life Expectancy. For example, Norway has a high life expectancy and low HIV/AIDs (1.6) with low total deaths (1060). The calculations also reveal that countries with high life expectancy have lower total health conditions, that is health problems.

Filtering was used to separate the dataset into smaller, more refined groups of countries that have high Life Expectancy. This filtering was conducted to provide insights into question 3, the interaction between geographic location and health condition. A filter was added to Life Expectancy Condition whereby the value was high. Results revealed that countries that have high life expectancy are in North American, Europe and East Asia. These countries are also categorized as being developed countries which assume better health care.

Additional filters were added to the year variable whereby the years 2000 and 2015 were included and any other years or no data was excluded. This filter was added in order to compare the data over time, specifically the oldest data from the year 2000 to the most recent data, the year 2015. This filter was then used to further examine the relationship between Life Expectancy, HIV/AIDS over time during this period (2000, 2015).

Therefore, after further refinement of the dataset and additional insights into the value of such data, some new specific questions were formed to improve the relevancy of the analysis. These questions included:

* 1. Does HIV/AIDs affect Life Expectancy?
  2. Does the year in which the data is reported impact Total Deaths by Life Expectancy?
  3. Does geographic location predict Life Expectancy?
  4. Does developed versus developing countries have differing Life Expectancies?
  5. Does education level impact Life Expectancy?

Visualizations

To answer the questions posed from the data exploration process a series of visualizations were created. These are used to provide further insights into the dataset.

*Does HIV/AIDs affect Life Expectancy?* HIV/AIDS is a debilitating health condition. When compared with Life Expectancy (Figure 4) there is a downward trend of HIV/AIDS versus with a comparative upward trend of life expectancy over time. In 2005, there were 429.3 HIV/AIDS cases and the average of life expectancy in the year 2005 was 68.21. However, in 2015, there were only 120.9 HIV/AIDS and the average of life expectancy in this year was 71.62. Thus, there is an inverse correlational relationship between the health status of HIV/AIDs and life expectancy.

*Does the year in which the data is reported impact Total Deaths by Life Expectancy?* Results reveal that as life expectancy increases total deaths decrease, year after year (Figure 5). In the year of 2000, there were a total of 49,912 deaths worldwide and the average of life expectancy was at 67. However, in the most recent year 2015, there were a total of 36,959 deaths and the average of Life Expectancy was at 72. These results may indicate either a better allocation of resources to underdeveloped countries more recently, or a better understanding of how health plays a role in life expectancy.

*Does geographic location predict life expectancy?* Geographic location does predict life expectancy, the Figure 6 visualization shows countries with high life expectancy in dark blue color including North America, East Asia, East and North Europe and Australia. Japan has the maximum average value of life expectancy at 82.54. On the other hand, countries with lighter color including South America, Africa and Iran have low life expectancy. One example is Sierra Leone, a country in Africa that has a minimum average value of life expectancy at 46.11.

*Does developed versus developing countries have differing Life Expectancies?* Developed countries have a higher Gross Domestic Product (GDP) per capita and tend to have higher life expectancy (see Figure 7). In the Developed vs Developing visualization, the average life expectancy of developed countries is 79.2 whereas the average life expectancy of developing countries is 67.8.

*Does education level impact Life Expectancy?* Education vs Life Expectancy (see Figure 8) shows the correlational relationship between the two variables. Countries that have a better education system also have higher life expectancy. In this visualization, Schooling numbers were used as education and life expectancy numbers were broken down into three different groups (Low, Mid and High). Low life expectancy countries have a maximum schooling score of 13, Mid-life expectancy countries have a maximum schooling score of 17, and High life expectancy countries have a maximum schooling score up to 21. This insight shows that mid-tier countries have more schooling than lower tiered countries, and high-tier countries have more schooling than mid-tier countries. This could reveal that through education people are more informed on ways to care for themselves and others, thereby potentially mitigating health problems.

Decision Tree Models Development and Discussion

Decision Tree 1 (Figure 9) used Total Deaths to determine what effect other variables have on death. In other words, “What variables have the highest values in the Total Deaths calculation? Variables including percentage expenditure, total expenditure, life expectancy, adult mortality, infant deaths, under-five deaths, and life expectancy condition were filtered out from being drivers during this decision tree. Total deaths are a product of three of those variables (adult mortality, infant deaths, under-five deaths). The largest predictor of total deaths with a predictive strength of 53.1% is Measles, alcohol, and BMI. The most common predictor of death was a BMI rating of greater than 29.2%. The most effective rule for this tree is BMI less than 29.2%, measles greater than or equal to 741, diphtheria less than 88 and Hepatitis B less than 64. This insight reveals that BMI is a large predictor of Life Expectancy. Results also help explain which variable have larger impacts on Life Expectancy and can therefore help guide health policies and educational strategies.

Decision Tree 2 (Figure 10) targeted Life Expectancy and drivers that help predict longevity of life. For this decision tree, all available the variables were used. Income composition of resources, Adult mortality, HIV/AIDs, under-five deaths predicted life expectancy with a predictive strength of 83.1%. The best predictor was the income composition of resources being greater than or equal to in value .796 with an adult mortality rate less than 122 and a thinness 5-9 years less than 2.5. This produced 16% of all records. The next largest rule predictive value was 76.03% was income composition of resources greater than or equal to .796 and an adult mortality rate of greater than 122 but less than 258. The largest predictor in life expectancy in almost every rule was income composition of resources. In summary, from decision tree 2 it becomes clear that the composition of resources within a geographic location is a strong predictor of Life Expectancy.

Dashboard Assembly

Two dashboards were created to further understand and communicate the result of this dataset (Figure 11, 12). The dashboards illustrate how several variables (health, nutrition, education, and lifestyle) contribute to Life Expectancy. The composition of resources, known as Gross National Income (GNI) is the total value of a country’s income in a year, divided by its population and reflects the average before tax. The GNI, is “a measure of all money, goods and services that come into or stay in a country” (Master Class, 2019). Often expressed in terms of Purchasing Power Parity (PPP), it is an important determinant of health status because the more income available to the citizens of a country, the easier it is for the citizens to improve on housing, nutrition, education and other living conditions (Global Finance, 2020). A country with higher income will have better access to housing, education, and health services. This leads to improved health, lower rates of mortality and a higher life expectancy. The opposite is true for a country with lower income.

Figure 11 dashboard displays the correlation between Expenditure and Life Expectancy. The higher the expenditure, the higher the standard of living. This, in turn, suggests an opportunity for a higher life expectancy. Figure 11 dashboard also shows the key drivers of life expectancy which includes thinness and refers to the nutritional level of the country. Malnutrition has a direct correlation to the physical and cognitive development of the population. Poor nutrition has a negative impact on life expectancy. The dashboard provides visual depiction comparing countries, when the variables are adjusted, using the interactive map.

On the other side of a weight scale is high BMI, which assesses the extent to which fatty tissue can cause serious diseases and health conditions (CDC, 2020). BMI is calculated by dividing a person’s weight in kilograms by the square of their height in meters (CDC, 2020). The average BMI is at 38 in the overweight range and therefore not considered as a healthy BMI, study shows that adults with extreme obesity have increased risks of dying young from cancer, heart disease, stroke, diabetes. (NIH, 2014).

Figure 11 dashboard displays the cost-effect relationship between education and life expectancy. The higher the level of education attained, the higher the average life expectancy. In addition, Figure 11 dashboard shows that average HIV/AIDS is at 2%. This insight is somewhat true as the survival rate has dramatically improved over the last two decades (Scaccia & Maddel, 2020). However, early testing and timely treatment are important in managing the virus which can extend life expectancy and reduce the risk of further transmission (HIV.GOV, 2017).

The dashboard in Figure 11, also indicates that developed countries have a higher life expectancy compared with developing countries. In developing countries, low standard of living, access to healthcare, malnutrition, infant deaths and adult mortality, low level of education among other factors are responsible for low life expectancy. While in developed countries, these factors are high and thereby contribute to high life expectancy.

Another interesting insight, seen in Figure 11 dashboard, is that the average life expectancy between 2009 and 2015, was between 70 and 72. This indicates that overall, the average life expectancy has been fairly consistent, and possibly driven by developed countries,“due to rapid declines in mortality, particularly infant and maternal mortality, caused by infectious diseases in childhood and early adulthood” (WHO, n.d)

In the Figure 12 dashboard, we can see that the combined effect of GNI, Adult Mortality, HIV/AIDs, under 5 deaths and thinness contribute 83% to life expectancy. Indicating a need to focus on these factors to extend life expectancy. On the other hand, the combined effect of several factors (health, social, lifestyle, education) contribute to total deaths at 63%. With this insight, public education on the importance of maintaining a healthy body and lifestyle is needed.

A resource whose mission it is to increase Life Expectancy, is the International Finance Corporation (IFC). This organization works in developing and emerging economies, to make financial investments, with the goal of lifting people out of poverty. The second dashboard is therefore a very useful tool for this organization as it would help track the developmental impact of the various investments the IFC makes.

Figure 12 dashboard also provides a quick snapshot to decision makers and helps them prioritize and focus their interventions. For instance, since there is a correlation between Life Expectancy and factors such as health, nutrition, education, economic status, the organization may be able to use the dashboard to identify where to deploy scarce resources among competing needs. Without a dashboard such as this, it would be difficult to allocate the appropriate resources to address issues of life expectancy in developing countries. In summary, the organization may not know when it has achieved its objectives, unless such an interactive and up-to-date dashboard can display the outcomes of its investment in developing countries.

Storybook Presentation

In summary, the Life Expectancy dataset reveals that geographic location, health condition and the combination of both affect Life Expectancy. This is a compelling story and can be seen in the storyboard presentation (<https://youtu.be/AVUVQsdyxlc>). Such results help to inform global resources and provide informative insights to help people live longer and healthier lives.

References

Centers for Disease Control (2020). The Health Effects of Overweight and Obesity. Retrieved from <https://www.cdc.gov/healthyweight/effects/index.html>

Fortuny, E. J., Martens, D., & Provost, F. (2014). Predictive modeling with big data: Is bigger really better? Big Data, 1(4). [http://doi.org/10.1089/big.2013.0037](https://doi.org/10.1089/big.2013.0037)

Global Immunization (2019). *What is Polio?* Retrieved from: [https://www.cdc.gov](https://www.cdc.gov/polio/what-is-polio/index.htm)

Global Finance Magazine (n.d) GDP-GNI-Definitions (n.d) Retrieved from: <https://www.gfmag.com/global-data/glossary/gdp-gni-definitions>

Hall M. (2020) In Investopedia. Retrieved from: <https://www.investopedia.com/updates/purchasing-power-parity-ppp/>HIV.gov (2017).

HIV.gov (2017). HIV Care and Treatment Activities. Retrieved from [https://www.hiv.gov](https://www.hiv.gov/federal-response/federal-activities-agencies/hiv-care-and-treatment-activities)

International Finance Corporation (IFC). (2020). Retrieved from: https://www.ifc.org

Master Class (2019). What is Gross Income? Definition and Formula for GNI-2020. Retrieved from: https://www.masterclass.com

Regent, D.(2020). *Worldometer.* Retrieved from <https://www.worldometers.info/gdp/us-gdp/>

Scaccia, A., & Madell, R. (n.d.). Facts About HIV: Life Expectancy and Long-Term Outlook. . Retrieved from: https://www.nih.gov

United States Census Bureau (2014). *Census Bureau Projects U.S. and World Populations on New Year’s Day.* Retrieved from: [https://www.census.gov](https://www.census.gov/newsroom/press-releases/2014/cb14-tps90.html#:~:text=Census%20Bureau%20Projects%20U.S.%20and%20World%20Populations%20on%20New%20Year's%20Day,-December%2029%2C%202014&text=As%20our%20nation%20prepares%20to,1%2C%202015.)

Wallace, E. (2018). Knowing life expectancy to improve public health. Policy Map. Retrieved from: [https://www.policymap.com](https://www.policymap.com/2018/10/knowing-life-expectancy-to-improve-public-health/)

Appendix

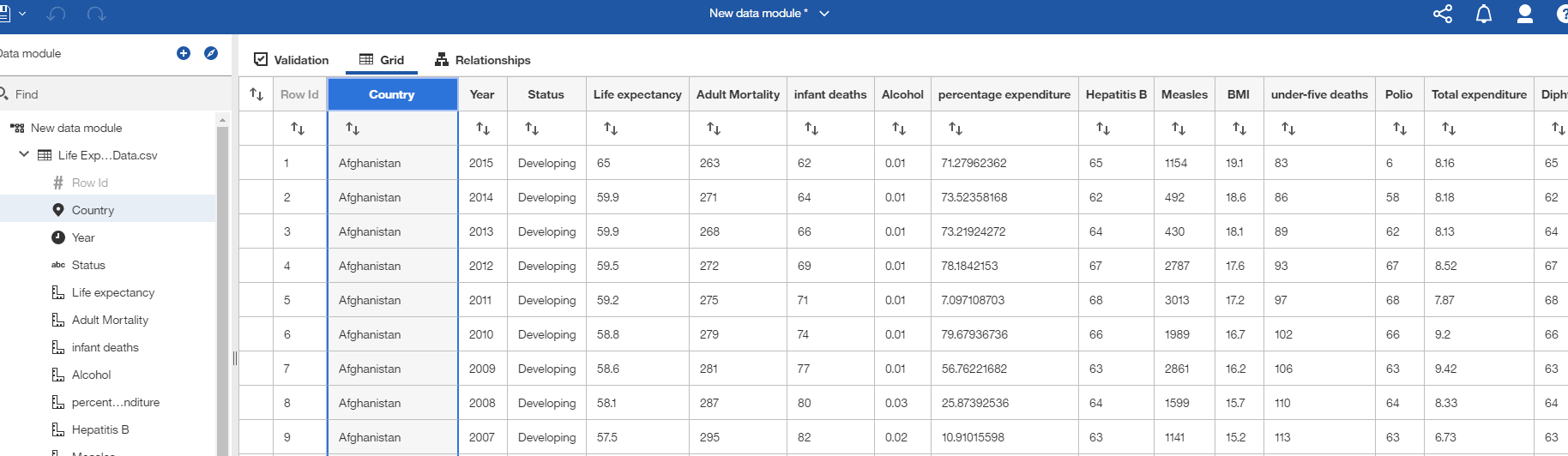


Figure 1: Life Expectancy Dataset

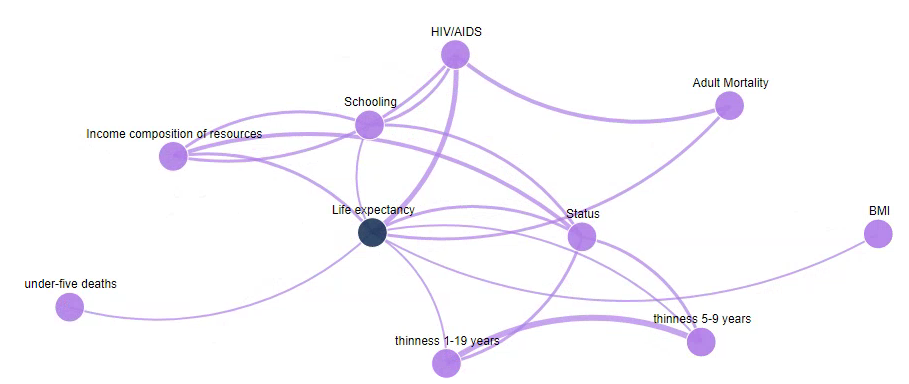


Figure 2: Relationship map after data cleaning

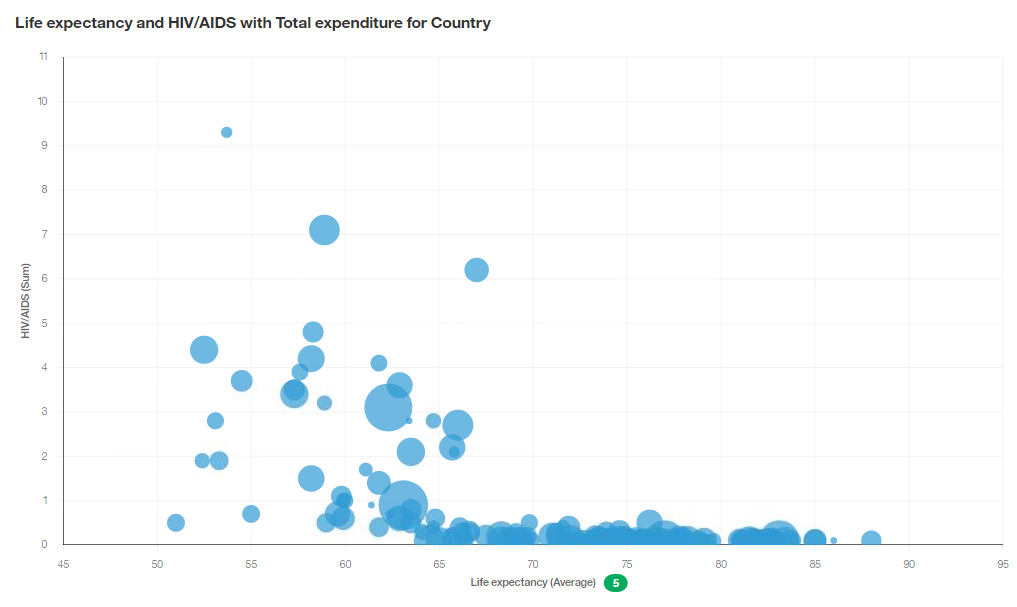


Figure 3: Outliers HIV/AIDS vs Life Expectancy

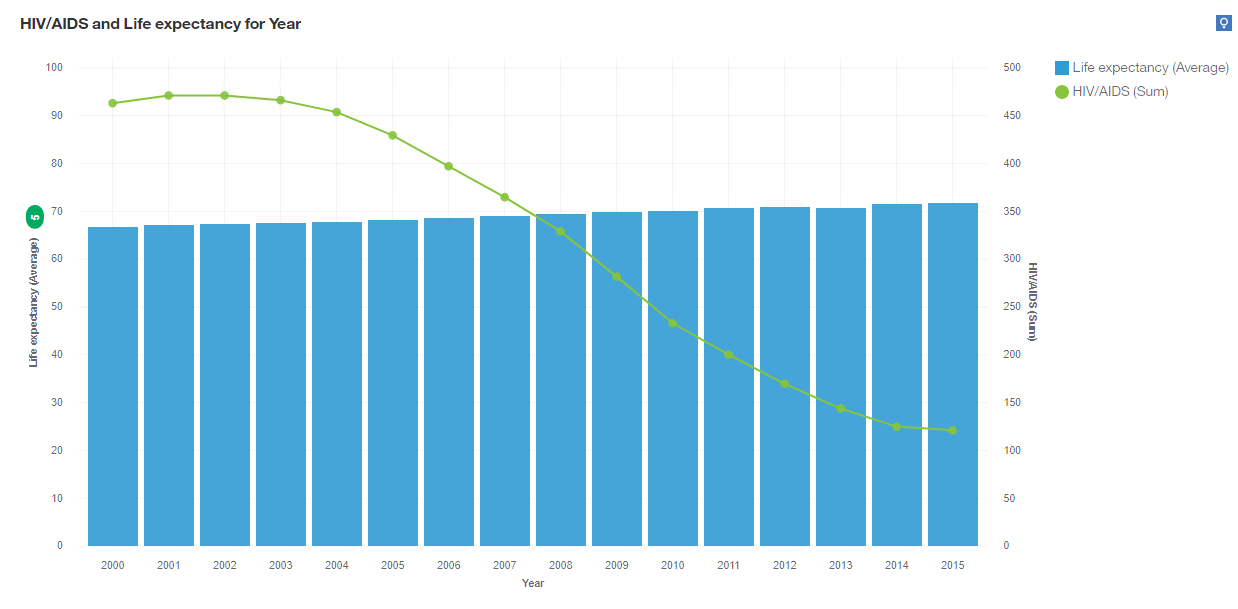


Figure 4: HIV/AIDS vs Life Expectancy

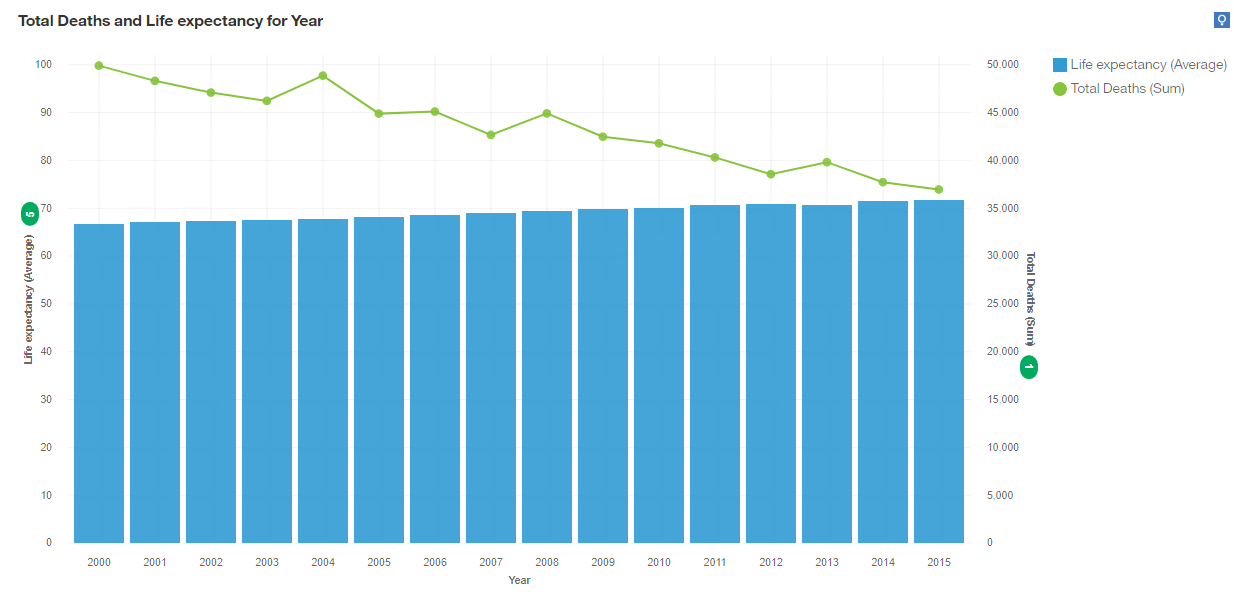


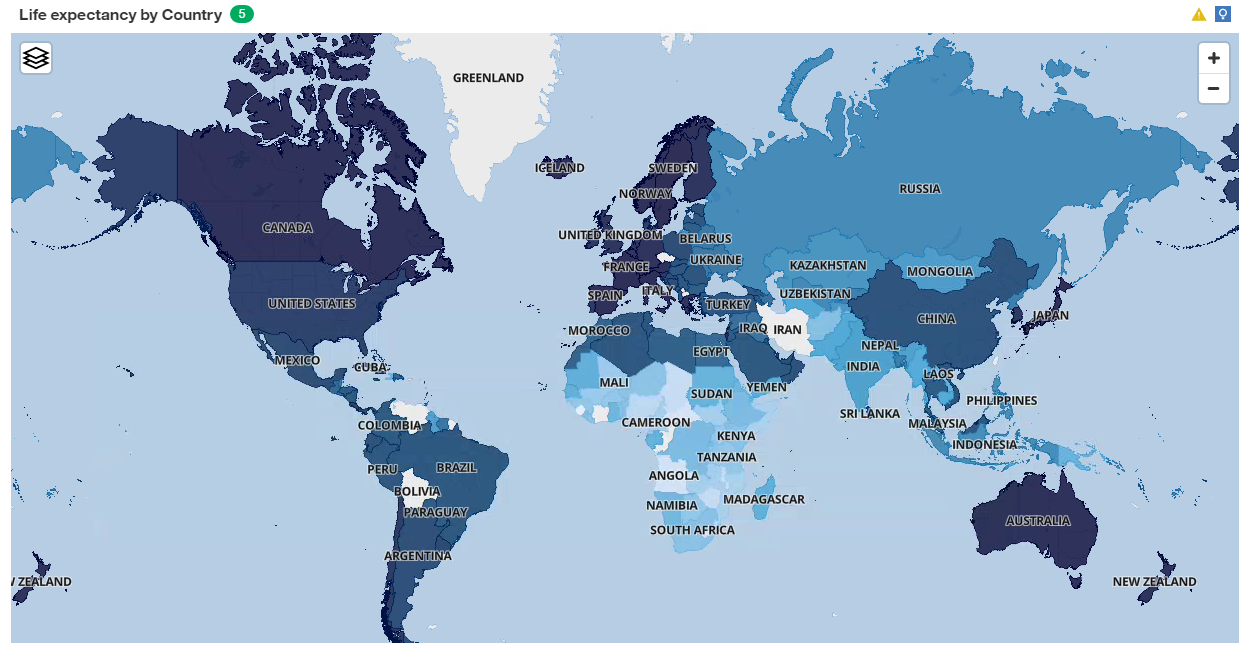
Figure 5: Total Deaths and Life Expectancy by year

Figure 6: Life Expectancy by Geographic Location

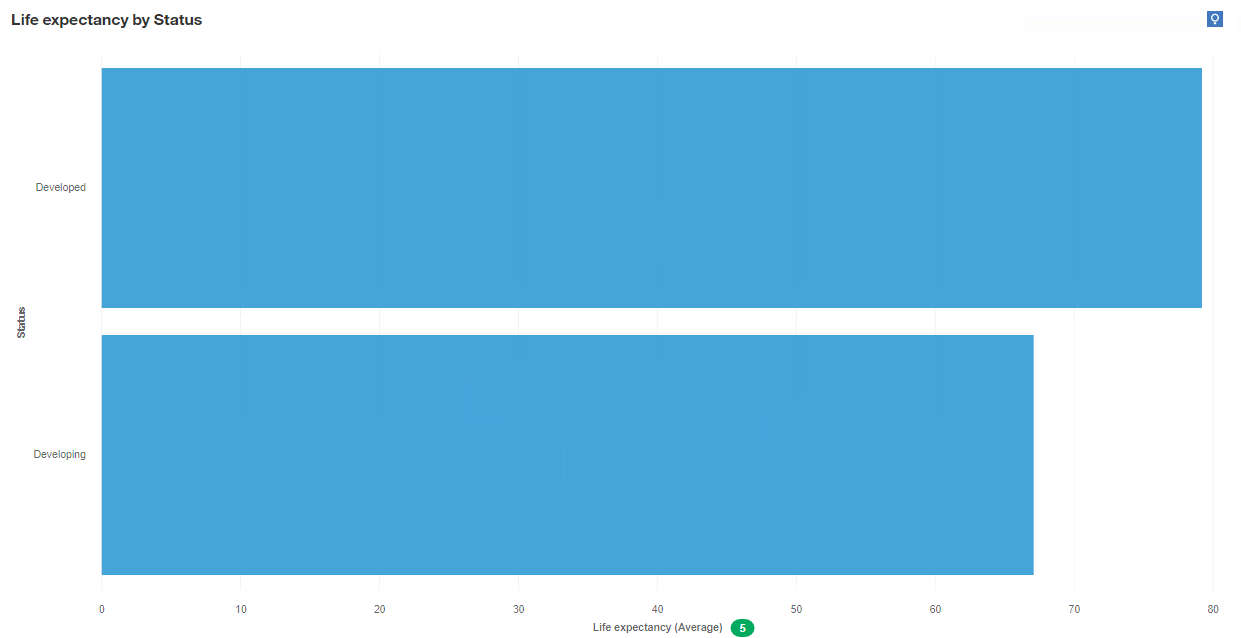


Figure 7: Developed vs Developing

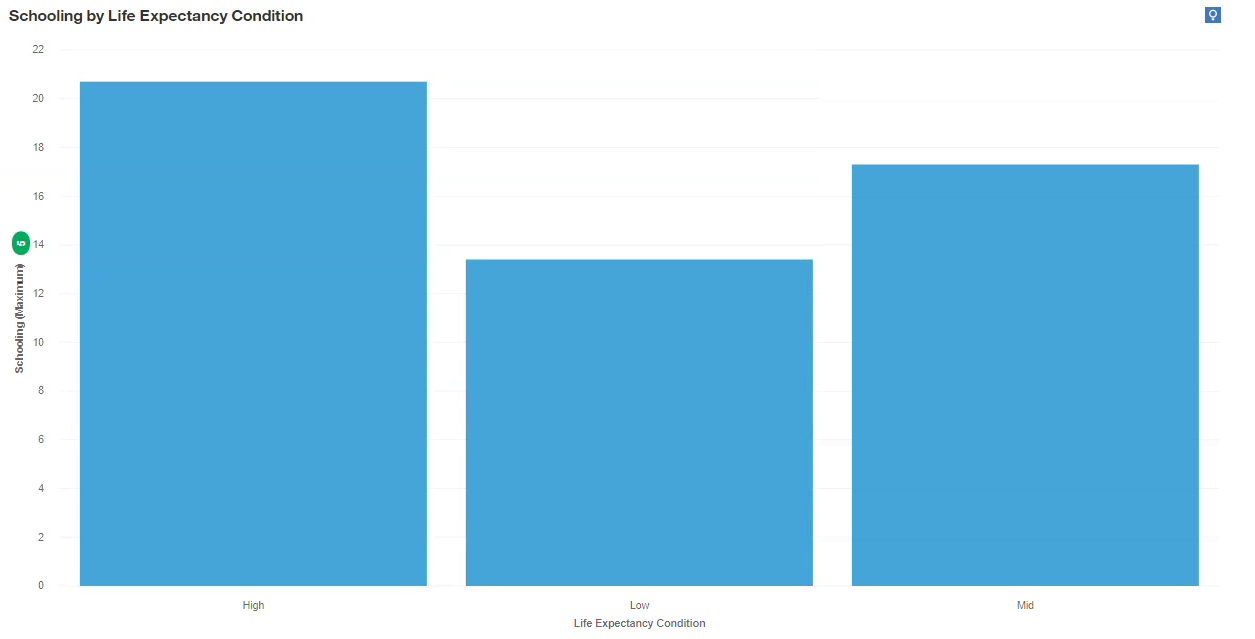
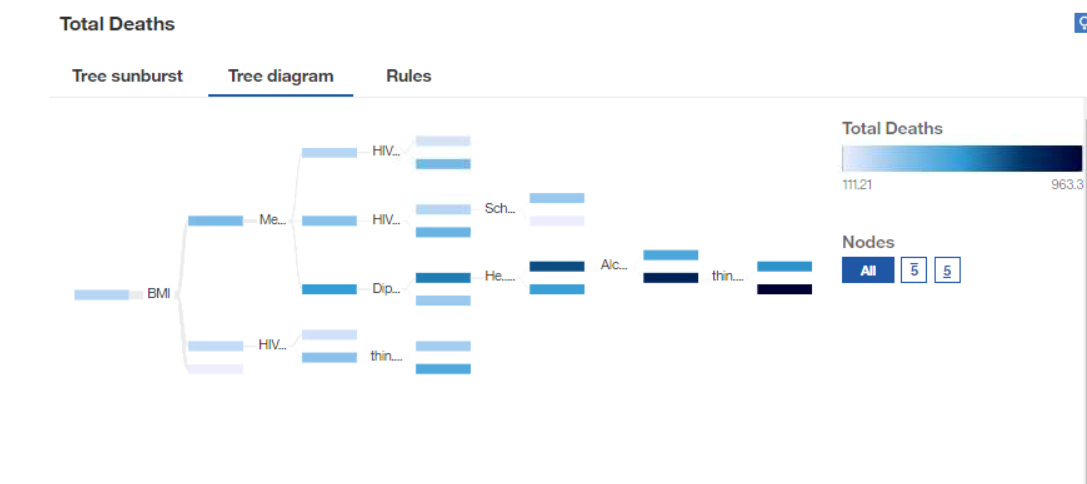


Figure 8: Education vs Life Expectancy

 Figure 9: Decision Tree 1

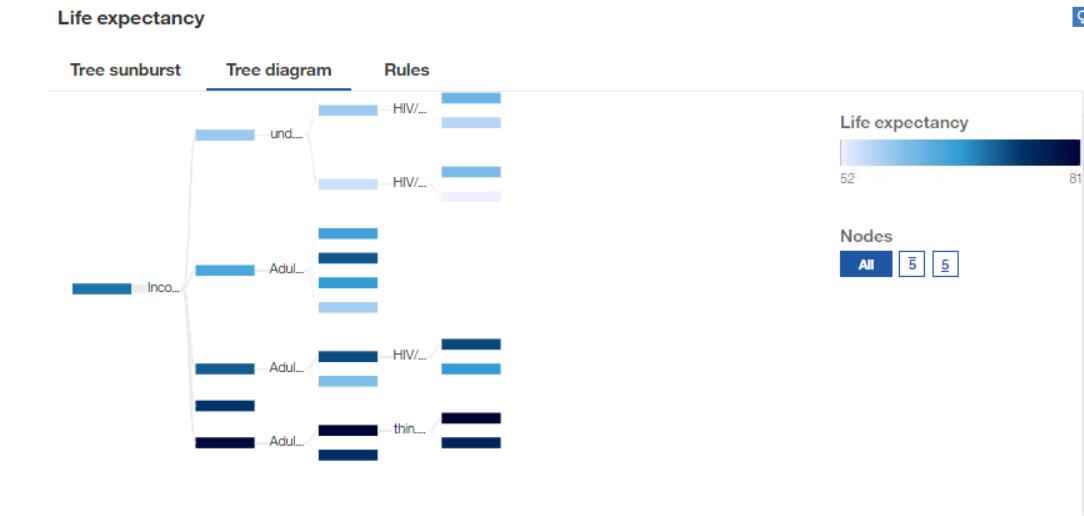


Figure 10: Decision Tree 2

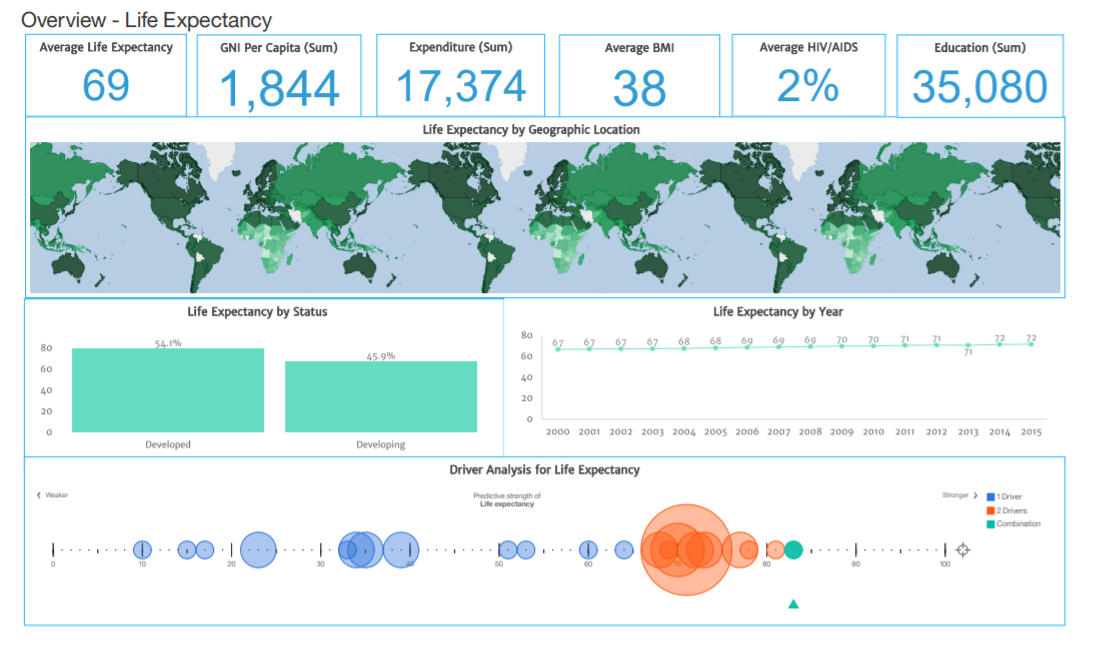


Figure 11: Overview Life Expectancy Dashboard

Figure 12: Target Dashboard