Life Expectancy Around the World

Group 3

Abby Herrup, Mark Burton, Jim Worlein and Clarence Robinson

29 March 2021

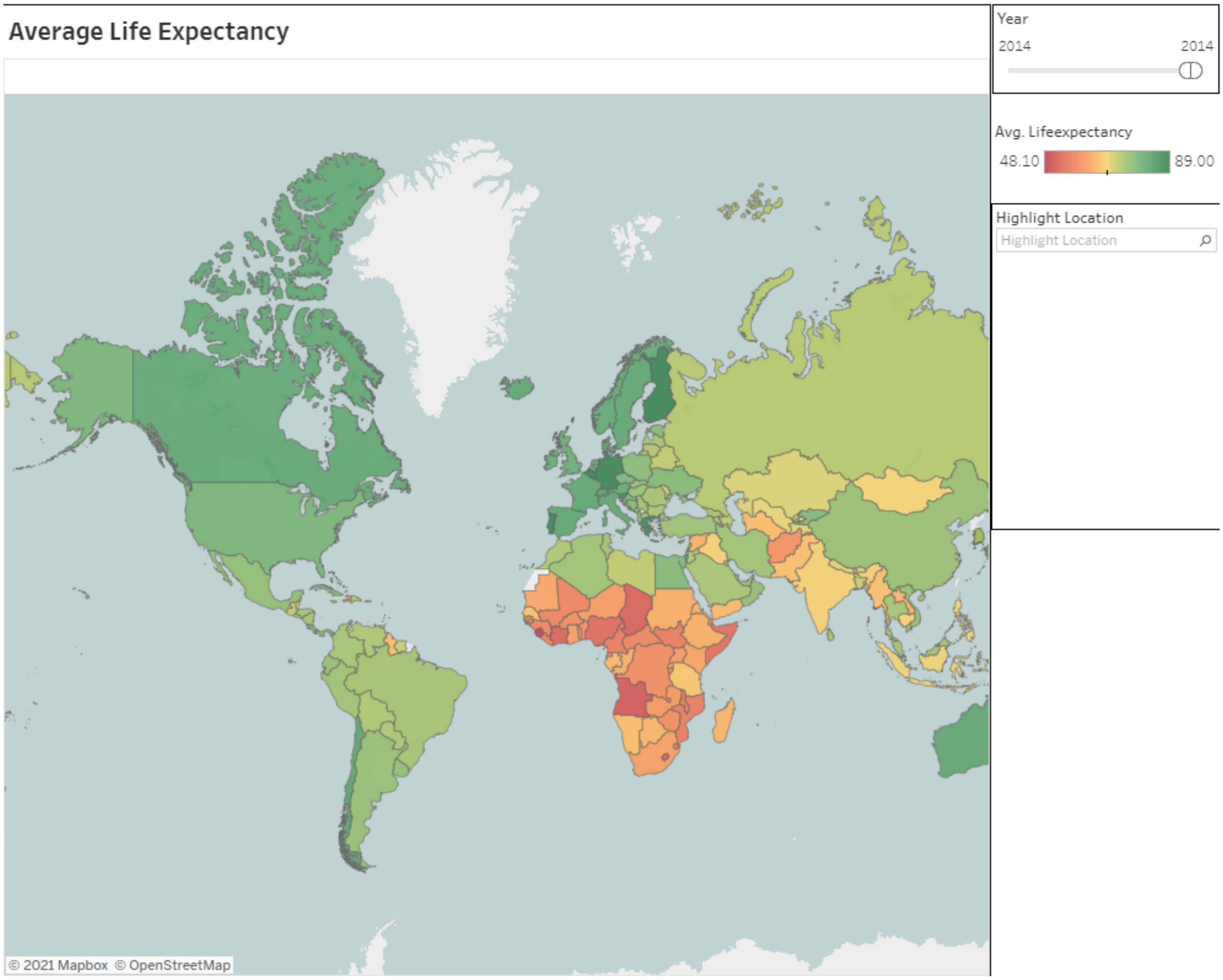
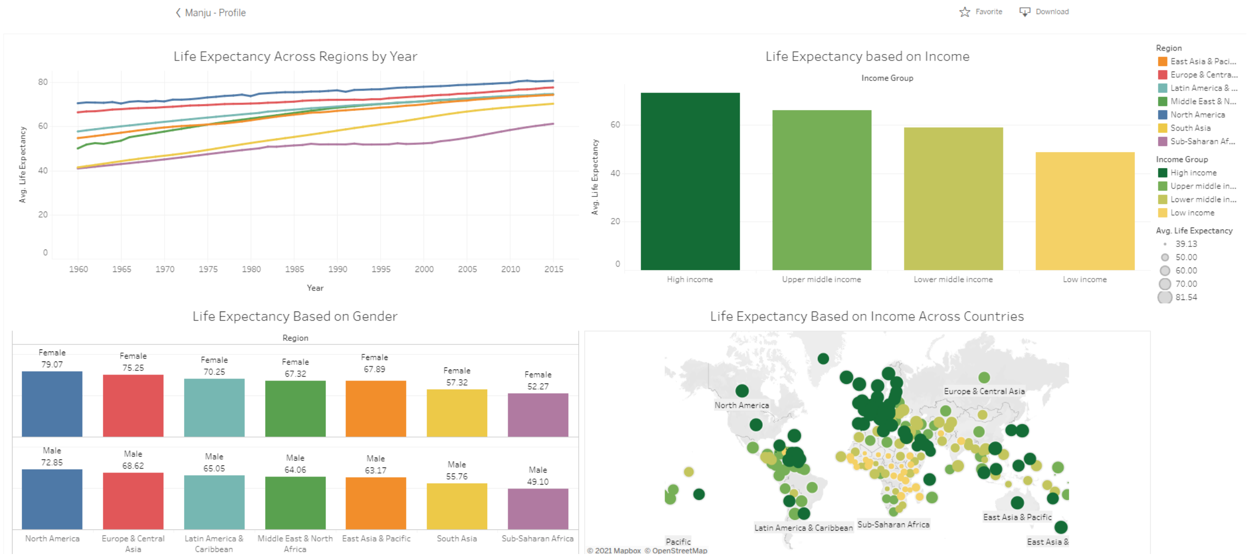


TABLE OF CONTENTS

Inspiration and Overview

For our capstone project, we decided we wanted to explore life expectancy around the world. We set out on a journey to unveil the data behind one of the most mysterious controversial questions we all ponder. The term "life expectancy" refers to the number of years a person can expect to live. Life expectancy is important for assessing population health. By understanding how the data is shaped by the demographics you can gain a broader more unbiased look at the world.When considering different topics, we were intrigued by the possibility of predicting how long the average person could live. We found this Tableau dashboard[1] by user Manju to be inspiring. The color scheme is quite interesting along with the combination of the map and line/bar charts.



Data & Modeling Approach

Our data set was posted on Kaggle and was created using health data from the World Health Organization and economic data from the United Nations. The CSV file covers 193 countries with 22 columns. The time period of the data goes from 2000 to 2015. The data set was compiled for users to explore how economic and health issues affect life expectancy throughout the world.

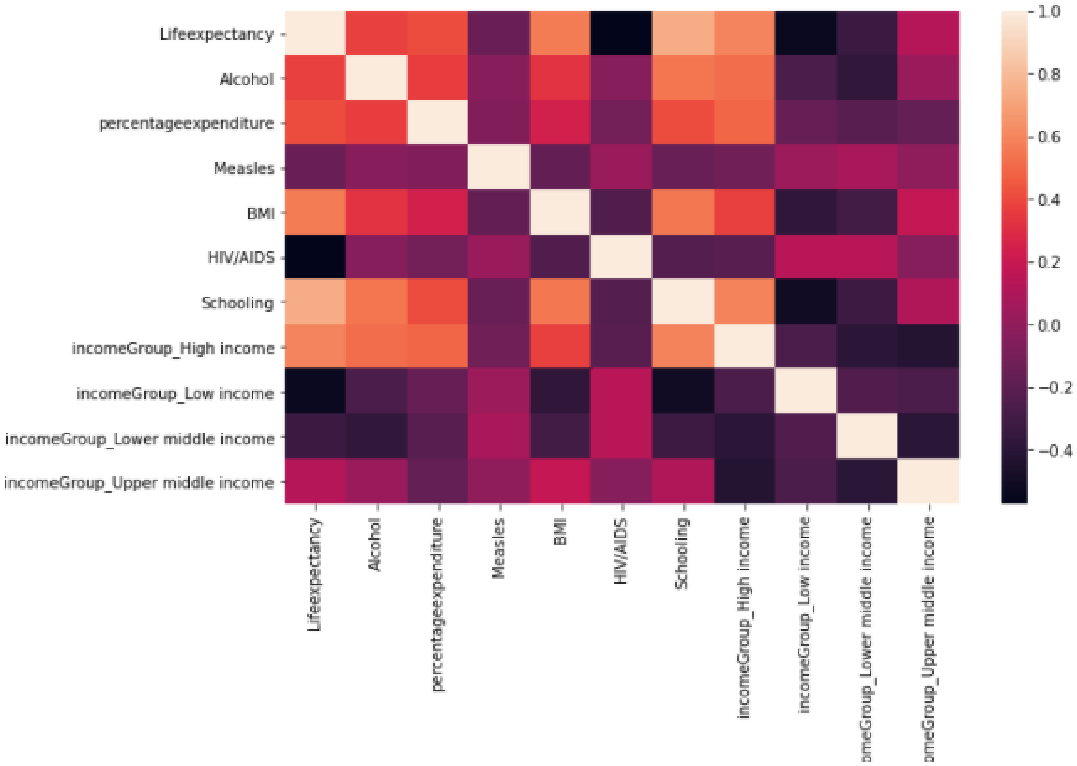
The data set column information can be categorized by economic, preventative health, and social/lifestyle factors. Economic factors would include: Gross Domestic Product, percentage of government expenditures on healthcare, educational opportunities, etc. Preventative health primarily covers vaccination coverage of such disease as polio, diphtheria, and hepatitis B. And lastly, social/lifestyle factors include body mass index and alcohol consumption.

[1] https://public.tableau.com/profile/manju8316#!/vizhome/Life-Expectancy_0/Dashboard1

We also found a CSV from the World Bank that had income groups for the countries over a similar timeline. We used Python to clean and combine the data into a single CSV. When cleaning the data, we combined the CSVs, dropped duplicate column names, and for the machine learning model, we also dropped all null values.

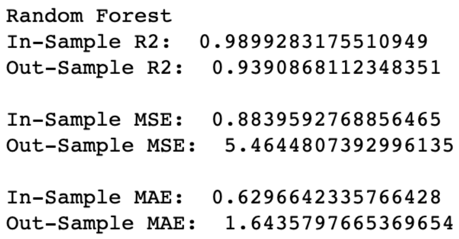
Numerous questions could be asked about this dataset including: (i) What health metrics impact Life expectancy? (ii) What other demographics impact life expectancy? (iii) What countries have the highest life expectancy and is it correlated to the human development index (HDI)? (iv) Do higher income regions have higher life expectancy? What is the relationship between life expectancy and average national income (per capita)?

Before starting the machine learning problem to answer our questions above, we wanted to take some time to explore the data. When exploring the data, we wanted to see how correlated each of our inputs are, to get a sense of what data may be most predictive for estimating life expectancy.



Results of Machine Learning Problem

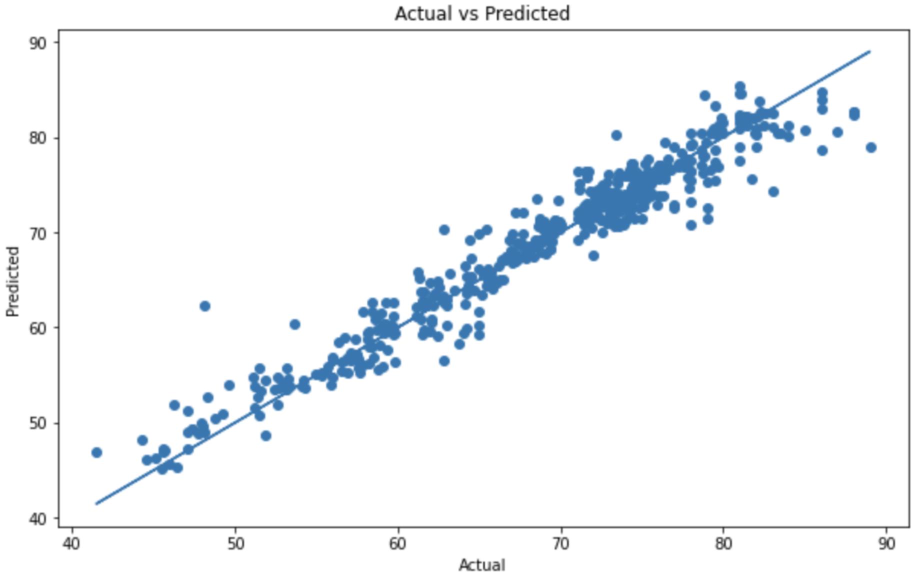
We looked at many different machine learning models, but in the end, we selected Random Forest Regressor. This model performed very well for our data with an Out Sample R-squared factor of 93.9%. This outperformed the Gradient Boosting Regressor and XGB Regressor models which had R-squared factors of 89.4% and 92.3% respectively.



```
Random Forest
In-Sample R2:  0.9899283175510949
Out-Sample R2:  0.9390868112348351

In-Sample MSE:  0.8839592768856465
Out-Sample MSE:  5.4644807392996135

In-Sample MAE:  0.6296642335766428
Out-Sample MAE:  1.6435797665369654
```

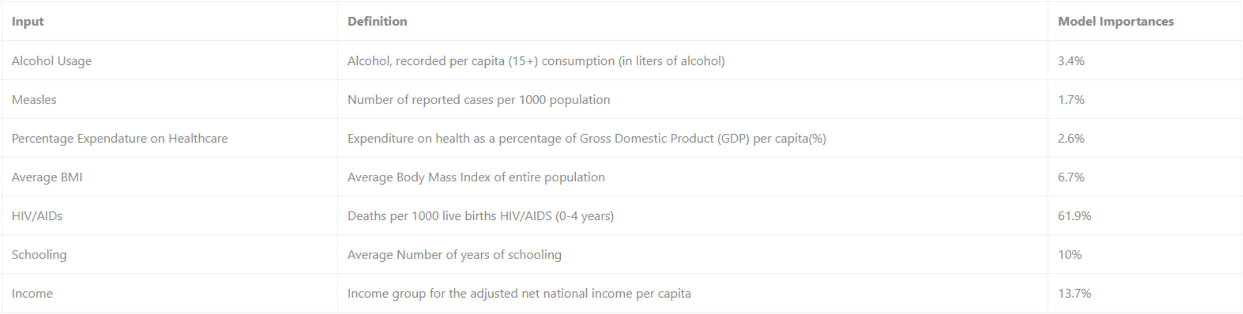


(i) What is the impact of health statistics on life expectancy?

Our model found that the rate of HIV/AIDS deaths is by far the most significant health indicator in predicting life expectancy of a country.

(ii) How does national income per capita affect the average life expectancy in a said country?

The income group of a country is also a significant indicator on the life expectancy of a country. On average, the countries grouped into the “High Income” tier of gross national income per capita had a higher life expectancy.

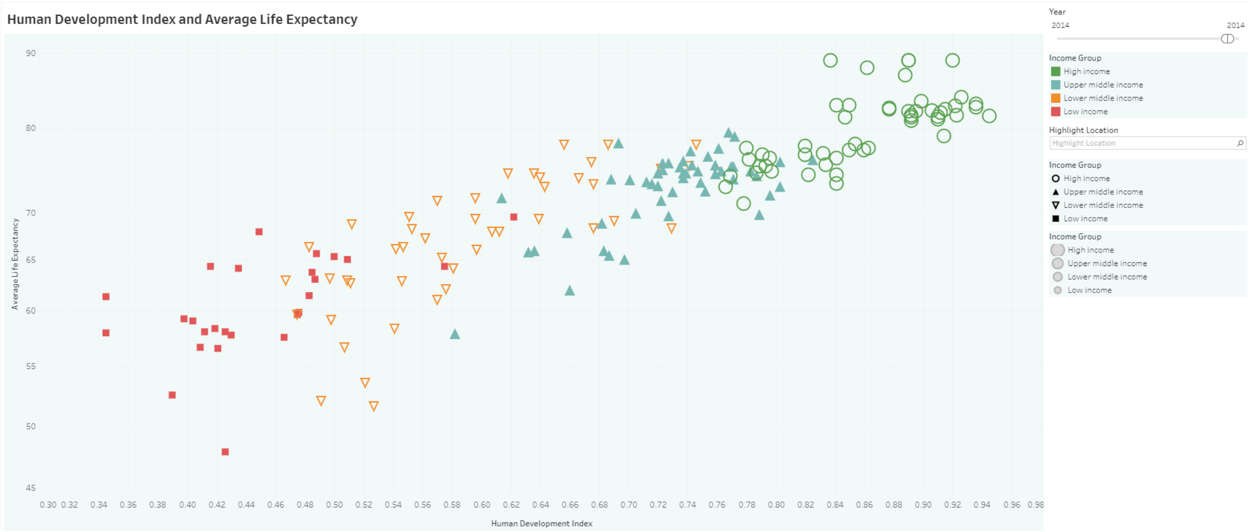


Input	Definition	Model Importances
Alcohol Usage	Alcohol, recorded per capita (15+) consumption (in liters of alcohol)	3.4%
Measles	Number of reported cases per 1000 population	1.7%
Percentage Expendature on Healthcare	Expenditure on health as a percentage of Gross Domestic Product (GDP) per capita(%)	2.6%
Average BMI	Average Body Mass Index of entire population	6.7%
HIV/AIDs	Deaths per 1000 live births HIV/AIDS (0-4 years)	61.9%
Schooling	Average Number of years of schooling	10%
Income	Income group for the adjusted net national income per capita	13.7%

Key for the Machine Learning Inputs (variable definitions provided by Kaggle)

Conclusions

There is a direct correlation between life expectancy and the Human Development Index as indicated by the chart below.



Countries with high HDI scores have longer life expectancies than poorer scoring ones. Access to healthcare, economic opportunities, and education have significant impacts on life expectancies throughout the world.

When analyzing our machine learning model- most of our variables we predicted would have an effect on life expectancy did- but there was one surprise. We hypothesized that expenditure on health care would significantly impact how long one would live, however, this variable had very low significance in the predictive power of our model.

Limitations/Bias

The first limitation we encountered is there are some null values in the data set. This was unsurprising as the size and developmental statuses of the 193 countries range from members of the G7 to tiny countries we were previously unaware existed. There were also inconsistencies across different columns, where some inputs did not have any values for random years.

Another limitation was the lack of data past 2015, it would be interesting to have more recent data, as socioeconomic changes happen often in developing countries, which would impact the different inputs we used.

Future Work Recommendations

There are many more areas one could explore in this topic, given more time/additional data. One area that could be further looked into is the relationship between healthcare and life expectancy. Going along with healthcare, one could analyze the access to vaccines and modern medicine remedies in the country and see how differences in that could possibly affect life expectancy. Future work could also be done to see how the pandemic has shifted life expectancy in various countries, especially for countries that were especially hard hit with Coronavirus cases, such as the United States.

Works Cited

Data:

1. Life Expectancy (WHO) Kaggle, 10 February 2018
2. World Bank staff estimates based on sources and methods in World Bank's "The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium" (2011) https://data.worldbank.org/indicator/NY.ADJ.NNTY.PC.CD