What is the life expectancy rate for densely populated countries?

Team Members: Sanjida Nisha, Nour Elabbasy, and Cindy Weng Zhu

Motivation

While living in a developed country, we often don't think about life outside of our country. However, it is really important to understand how or what the lives of people are who don't have the luxury to live in a developed country where unemployment and poverty rate is low, where the living conditions are healthy and where there is an equal distribution of income. One way of understanding more about living conditions in developing countries is through life expectancy rates. Life expectancy refers to the number of years a person can expect to live. There are many factors that contribute to this number.

Let's take a look at the current pandemic situation around the world. With the recent deaths of COVID-19, it also gave rise to previous viruses and other health factors that have taken effect into human life expectancy. Certainly, health related and immunization factors have somewhat of a relation with life expectancy but socioeconomic factors of what each country possess can have some effects too.

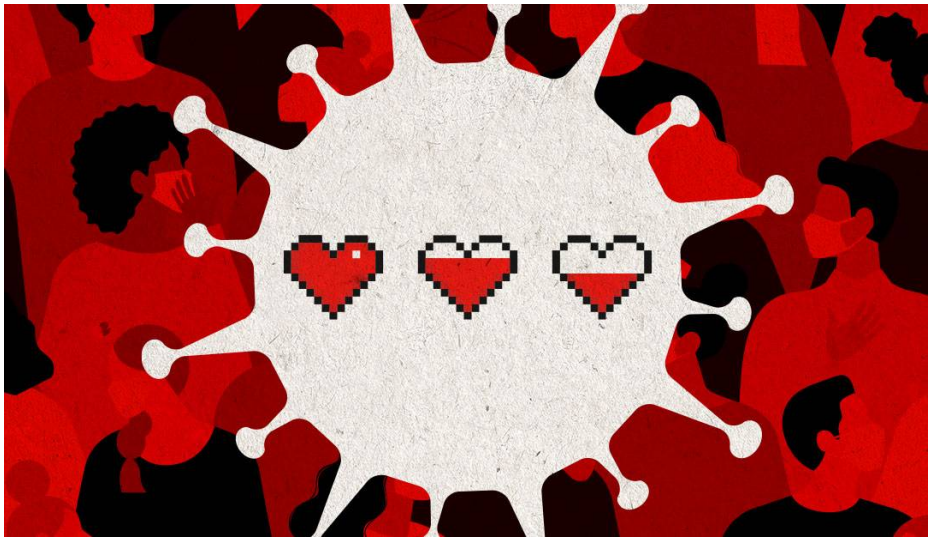

Figure 1. Decreasing Life Expectancy with COVID, cr: Princeton University.

A developing country can be defined as a poor agricultural country that is seeking to become more advanced economically and socially. They usually have a slow rate of industrialization and low per capita income. Unemployment and poverty is usually high in these countries. With this in mind, we will explore this dataset to help us learn about the life expectancy of people in densely populated countries. Therefore, through this blog post we will be exploring what the life expectancy rate is for more densely populated countries.

Data

We got our data set from Kaggle and it is called "Life Expectancy (WHO)" and it is about life expectancy and many health factors from 193 countries. The information was provided by the WHO data repository website and its corresponding economic data was collected from the United Nation website. The data covers information from 2000 - 2015 and it has 20 factors that are divided into the following general sections: Immunization related factors, Mortality factors, Economical factors, and Social factors.

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0

Figure 2. Raw Life Expectancy DataFrame

We first started cleaning the data by renaming some of the columns so they are uniform throughout the data. Then we decided to drop several columns because they would not help us in answering our question. After that we changed the column "Status" from a string value to integer, where developing countries are 0 and developed countries are 1. We encoded the countries using integers so each country got its own number and we assigned that value to a new column called "Country_int" and then dropped the country column.

	Status	Life Expectancy	Adult Mortality	GDP	Population	Schooling	Country_int
0	0	65.0	263.0	584.259210	34413603.0	10.1	0
1	0	59.9	271.0	612.696514	33370804.0	10.0	0
2	0	59.9	268.0	631.744976	32269592.0	9.9	0
3	0	59.5	272.0	669.959000	31161378.0	9.8	0
4	0	59.2	275.0	63.537231	30117411.0	9.5	0
...	...	...	...	...	...	...	...
2933	0	44.3	723.0	454.366654	12019911.0	9.2	192
2934	0	44.5	715.0	453.351155	11982219.0	9.5	192
2935	0	44.8	73.0	57.348340	11954293.0	10.0	192
2936	0	45.3	686.0	548.587312	11923906.0	9.8	192
2937	0	46.0	665.0	547.358879	11881482.0	9.8	192

2938 rows × 7 columns

Figure 3. Cleaned and Preprocessed Life Expectancy DataFrame

We chose this data set because we thought that it had enough accurate data that would help us answer our question. Additionally, since there were so many features we were able to focus on a couple that we thought would help us answer our questions and guide our project. However, one limitation is that the last time it was updated was 6 years ago which means that our conclusions are only appropriate for 2000-2015 not now. Also, we realized that there was an issue with how the population was represented so we decided to add another data set from the World Bank. We only took data from 2000-2015 and we renamed some countries.

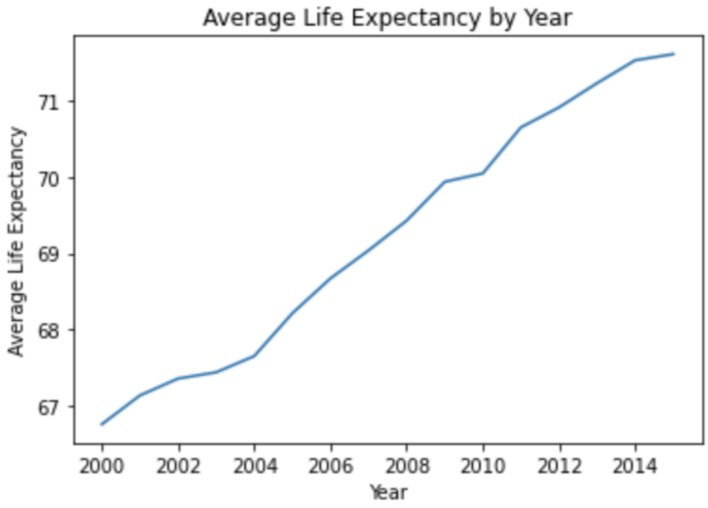

Figure 4. Average Life Expectancy Rate vs. Year

The line graph above shows that as the years passed on the average life expectancy increased linearly. We speculated that that would be the case because medicine has progressed a lot so diseases that were considered deadly are now curable.

Model

We built a Linear Regression model from scikit-learn's library because most of the values in our dataset are continuous values which are where linear regression models are a good start. In addition, we would like to predict the life expectancy rate for countries with more dense populations and linear regression models can help us answer this question.

Evaluation

After training our model, we use scikit-learn's library 'score' method for the linear regression models which calculates the coefficient of determination R^2 of the model where the higher the score, the better. Using the testing sections of the dataset, the model scored approximately a 0.7 R^2 score. This means that there is a decent variance between our features and the target and that most of our data points fall closely in the regression line.

We used a Predicted vs. Actual plot where we can explain this observation by using the Seaborn library and its `regplot` method. A predicted vs. actual plot lets us observe how closely our actual and predicted target values (in our case, life expectancy values) are from the dataset and from the model respectively.

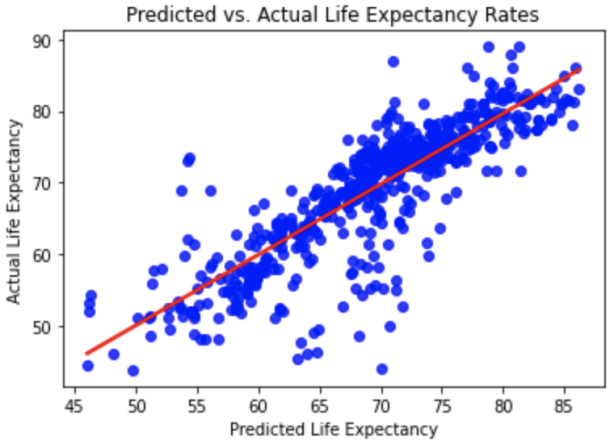

Figure 5. Predicted vs. Actual Life Expectancy Rates

As we can see in the graph, most of our data points are very close to the regression line which means that most of our predicted life expectancy rates correspond to the actual life expectancy rate.

Another scikit-learn method that we used to evaluate our model was the Mean Squared Error (MSE) which measures the average of a set of errors and shows us how fit our line is to our data where the lower the score, the better. The model that we built and trained scored approximately 25, our MSE score is low which shows that many of the data points can fit into our line of best fit but not all of them.

Last but not least, we used residual plots to evaluate our model. Seaborn's `residplot` method allows us to plot our residual values.

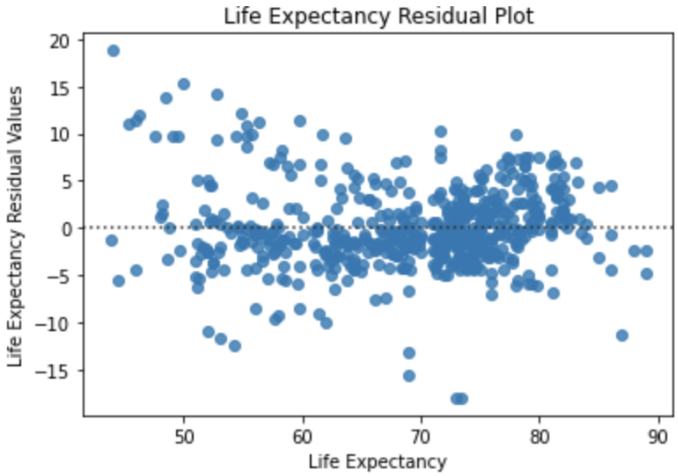

Figure 6. Life Expectancy Rate Residual Plot.

For a decent linear regression, the residual plot would show a random scatter of data points. For our model, we have somewhat of a random scatter but it shows more of a pattern than random. Since our model had an R^2 score of approximately .70, there is still room for improvement if we had more time for our project.

To answer our data science question, we can take a look at the dataset as it was given and also use our model. From the dataset provided, we can view two drastically different population densities and compare their life expectancies. In this case, we noticed that the most populated

country in our dataset was China with approximately 1.3B and the least populated country in our dataset was Narau with approximately 10K.

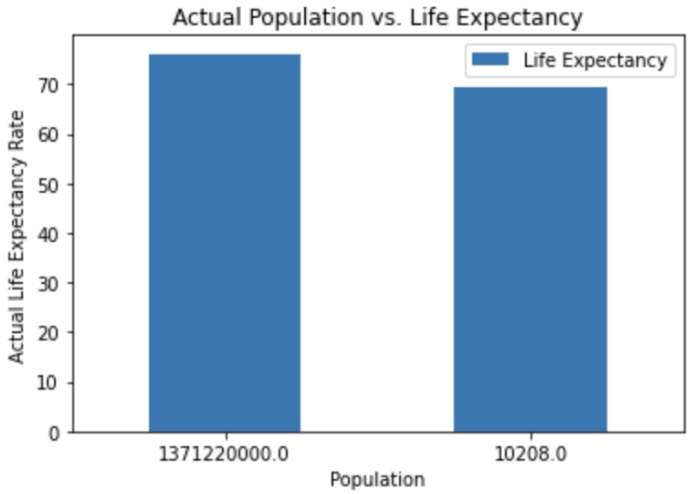

Figure 7. Population vs. Actual Life Expectancy Rates Bar Plot

From this graph, we can see that the country with the highest population density on the left has a higher life expectancy than the country with the lowest population density.

Furthermore, we can use our model to predict the life expectancy rate against the population. For this, we decided to choose a country (United States) to predict its life expectancy rate as we increase the population density.

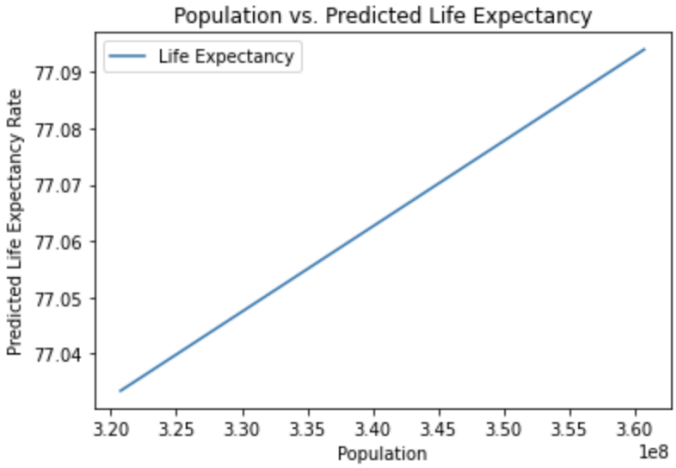

Figure 8. Population vs. Predicted Life Expectancy Rate Plot

As we can see in the DataFrame and the graph, when we increase the population (in other words, making the population denser), the rate of life expectancy increases (although by very little).

More densely populated countries have a higher life expectancy according to our model predictions and the data provided. Although other factors within our dataset affect the life expectancy rate more than population density, we can still observe how population density might affect it when keeping other variables constant. The reason why an increased population causes an increase in life expectancy rate might be because as the population increases for a country, there might be more resources and opportunities for its population such as health resources, economy increases, etc. So, population density affects other factors such as health factors and socio economic factors which directly and positively affect life expectancy rates in a country.

We do not feel confident about its performance, there is still a lot we can work on in our dataset to improve the coefficient of determination or a better scatterplot for the residual values.

Future Work

Throughout this project, we mainly focused on answering one question, which was how much the life expectancy rate is for more densely populated countries. However, just this one question doesn't justify the amount of information this dataset presents. There are many other answers we can get from this dataset and can also keep forming new questions. For future work, we will heavily focus on getting accurate data for every country. For the sake of our data model, we replaced some of the population Kaggle presented to us with populations from the World Bank dataset for every country. However, there were some countries in Kaggle that weren't presented in the World Bank dataset, so we left those populations as it is. Also, we would fact check the rest of the variables and make sure the correct number is presented in the dataset, as this will help us improve our model and form more accurate results.

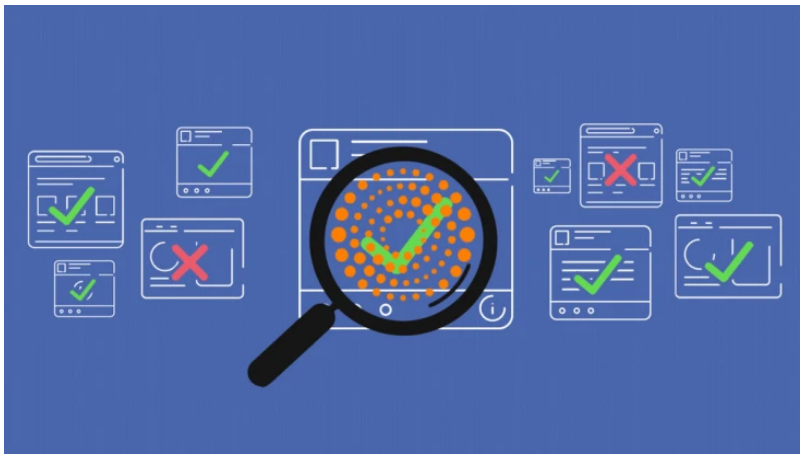

Figure 9. Fact Check, cr: TechCrunch

We would also like to get answers to our other questions, for example, which factors affect life expectancy the most and how this varies for each country. This is important to understand because it will help us understand the major issues that are causing death around the world and bring attention to it so the next step can be done to improve this issue. It will also help us understand whether or not these life expectancy rates are affected by immunization related factors, mortality factors, economical factors or social factors. Another future work we would like to do is predict results for upcoming years. For example, the life expectancy for upcoming years for every country for all of the different factors. We can also compare how each factor affected every country. For instance, what percentage of people around the world die because of alcohol, measles, hepatitis B etc, as this will help the the people of every country take precautions accordingly or what policies to make.

Figure 10. Life Expectancy Resources, cr: Gp Lah!

References

- *KumarRajarshi. "Life Expectancy (WHO)." Kaggle, 10 Feb. 2018, https://www.kaggle.com/kumarajarshi/life-expectancy-who/code.*
- *"Population, Total." The World Bank, Data, https://data.worldbank.org/indicator/SP.POP.TOTL.*