Global Population Projections

HarvardX PH125.9x Data Science Capstone Project

Gerry Henstra

December 2019

INTRODUCTION

Objective

Global population impacts every resource on our planet. It impacts agricultural yields, global ecomomics, climate, infrastructure, standard of living, and more. The past century has seen the global population expand beyond any previous period in the history of our planet, and it appears it will continue into the future.

The goal of this project is to use past population levels and variants to predict what the global population will be in 2050. Understanding this can help determine the global resources needed to sustain ourselves in the future.

Approach

A look at the trends in population by country since 1950 and the analysis of the various impacts or variants will be used to predict the population to 2050.

A look at the variants since 1950 will identify trends required to extrapolate and predict the population for each year between 2020 and 2050.

A custom predictive algorithm will be developed versus using linear regression, RSME or other methods learned in this course. The variants require some special adjustments to be more accurate, so by building custom logic, assumptions and adjustments can be catered to the variants needed to predict the future population. Plus I also wanted to use the skills developed in this course to their fullest. From the organization of the R scripts, to the development of more complicated R functions, I am employing as many skills learned a possible from the course.

Just three of the many variants provided by the UN will be used. Birth rate, death rate and life expectancy will be the main variants used as it is felt they provide the most impact.

The UN provides it's own prediction of population beyond 2019, and this will be used as the validation set. The results of the custom predictive algorithm will be compared to that of the UN predictions and see how close the predicted population comes to that of the UN.

Data

The data used comes from the United Nations World Population Prospects data files.

These can be found at https://population.un.org/wpp/Download/Standard/Population/.

The files used in this report are:

1. WPP2019_PopulationBySingleAgeSex_1950-2019.csv : actual population by age, country and year from 1950 to 2019

2. WPP2019_Period_Indicators_Medium.csv : variants in 5 year intervals by country including the birth rate, death rate and life expectancy

3. WPP2019_PopulationBySingleAgeSex_2020-2100.csv : UN predicted population by age, country and year from 2020 to 2100 (validation set)

4. WPP2019_F01_LOCATIONS.XLSX : hierarchy of the countries used in the above sets

Country specific population will be used in the analysis as well as the global population, however the prediction will be at the global level only.

Project Structure

The project has been organized into six (6) R script files.

1. Main
2. Packages
3. Global
4. Data
5. Plots
6. Prediction

Each contains methods related to the main components of this project.

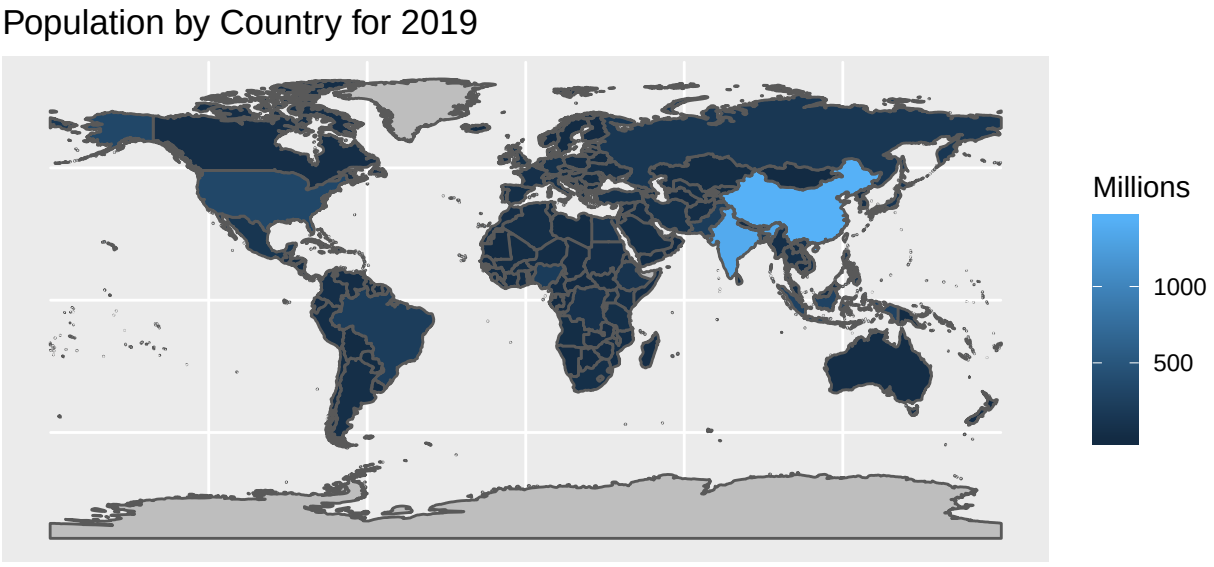
population_main : imports each of the script files and executes the data import
population_packages : imports each of the required libraries
population_global : contains common routines and global variables used throughout
population_data : contains all data sets and routines related to data processing, including downloading, importing and wrangling
population_plots : contains all reporting routines including plots
population_prediction : contains all routines used to build and test our prediction algorithm

When a function is called, the name of the file is used to denote the script where that function can be found. For example in the main function the import data routine is called like this: population_data.import_data(). This allows you to know exactly where to find the routine being executed when looking at the scripts. Only functions called from outside the script use this convention. If a script is used only as a sub-routine, it will have a simple name.

DATA EXPLORATION

The current world population as of `2019` is approximately `7.7` billion people, compared to a population of just `2.5` billion in `1950`. This is a `204`% increase over the past `69` years.

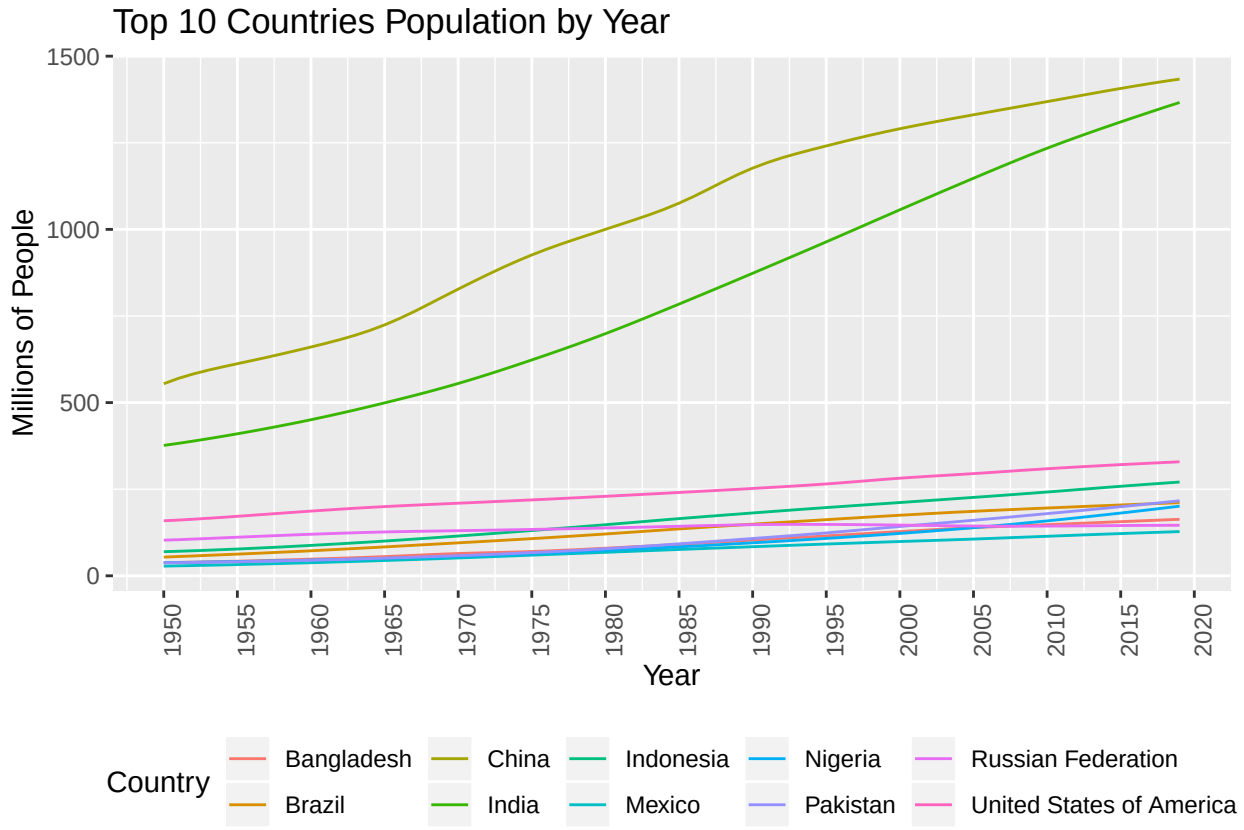
The following map shows the distribution of population by country for 2019. You can see there are just a few countries with much larger populations than the rest of the world.



Code	Country	Population
156	China	1,433,783,692
356	India	1,366,417,756
840	United States of America	329,064,917
360	Indonesia	270,625,567
586	Pakistan	216,565,317
76	Brazil	211,049,519
566	Nigeria	200,963,603
50	Bangladesh	163,046,173
643	Russian Federation	145,872,260
484	Mexico	127,575,529

The top 10 most populated countries in `2019` make up `58`% of the global population with a total of `4.5` billion people.

Since 1950, China and India alone, have contributed an additional `1.9` billion people to the worlds population.

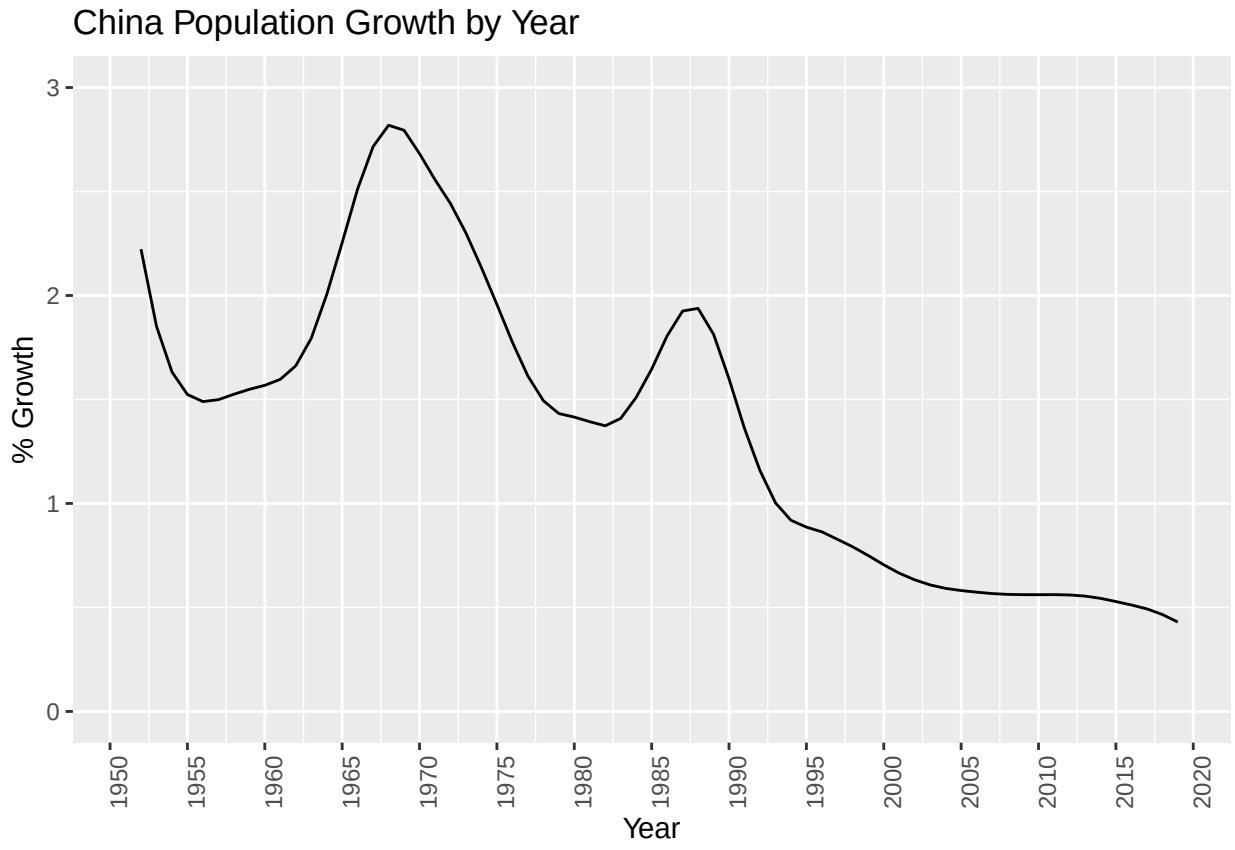


The chart above shows how China and India have a steep increase in population, while the remainer of the the top 10 most populated countries have a gradual, more modest increase.

China's One-Child Policy

In 1980, China instituted a one-child policy, which limited the number of children a family could have to one. Previously families were limited to two children, however they recognized that they needed to control the size of their population and limiting the number of children was a means to do this. Between 1950 and the 1976, China's population almost doubled to just under 1 billion people. They announced the one-child policy in 1979 and instituded in September of 1980.

By 2011, China's fertility rate (births per woman) fell to 1.4 and in 2015 they increased the single child policy to two. Althought the population continued to grow, the growth per year declined.

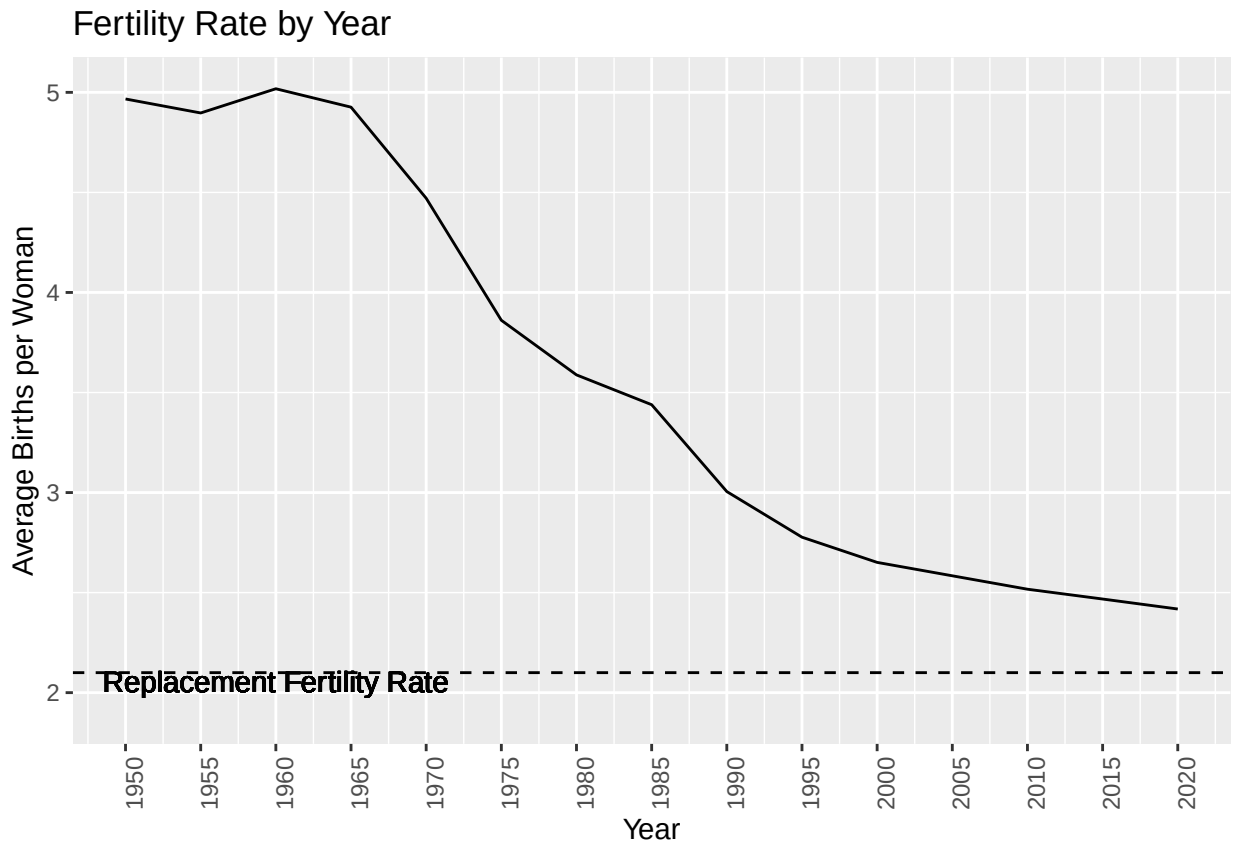


With China as a major contributor to the worlds population, impacts to their population impacts the world as a whole significantly, so China, and also India, require close attention to their future population trends.

Replacement Level Fertility

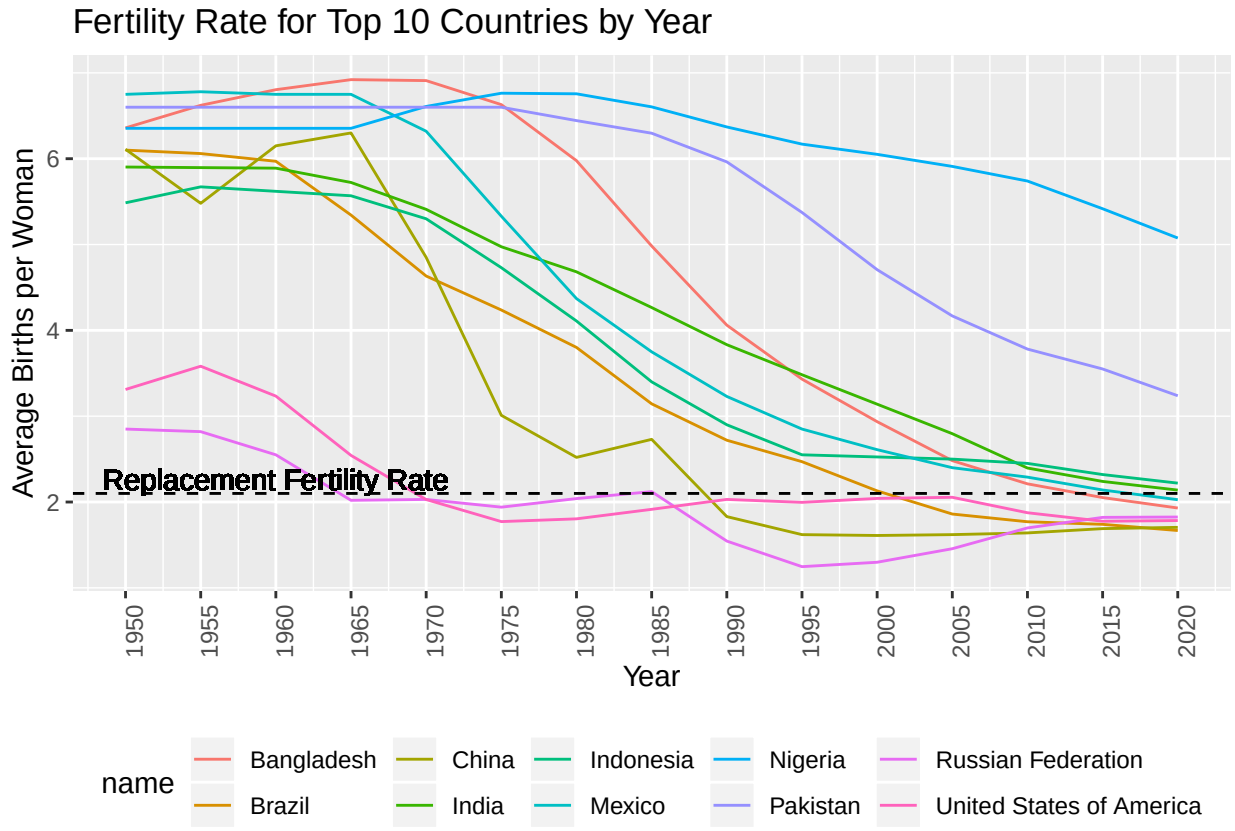
It is estimated that a country needs to have a replacement level fertility rate of 2.1. This represents the average number of children born per woman, at which the population replaces itself from one generation to the next, without migration.

The following chart shows that the global average fertility rate is steadily decreasing, nearing the replacement fertility rate.

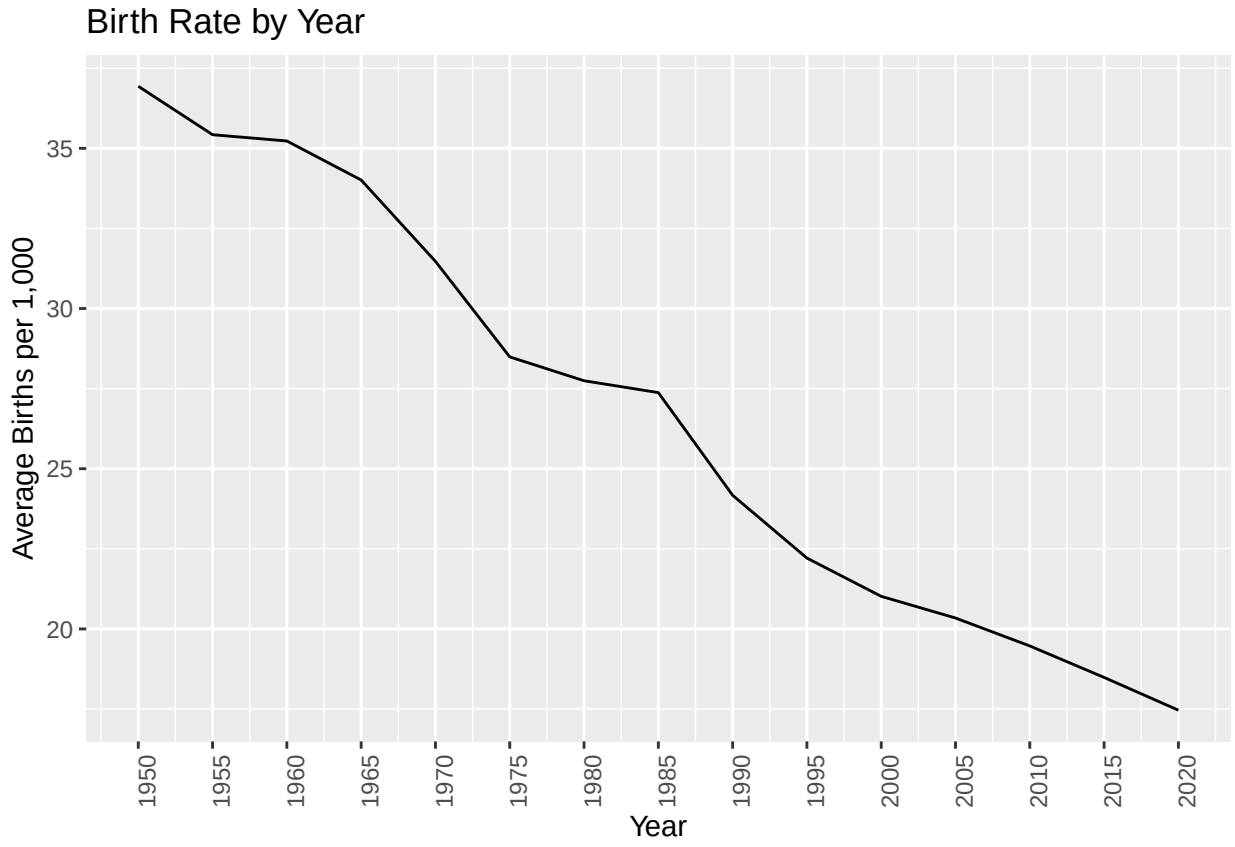


The global fertility rate dropped from `4.897` in the 50's to `2.418` currently in 2019.

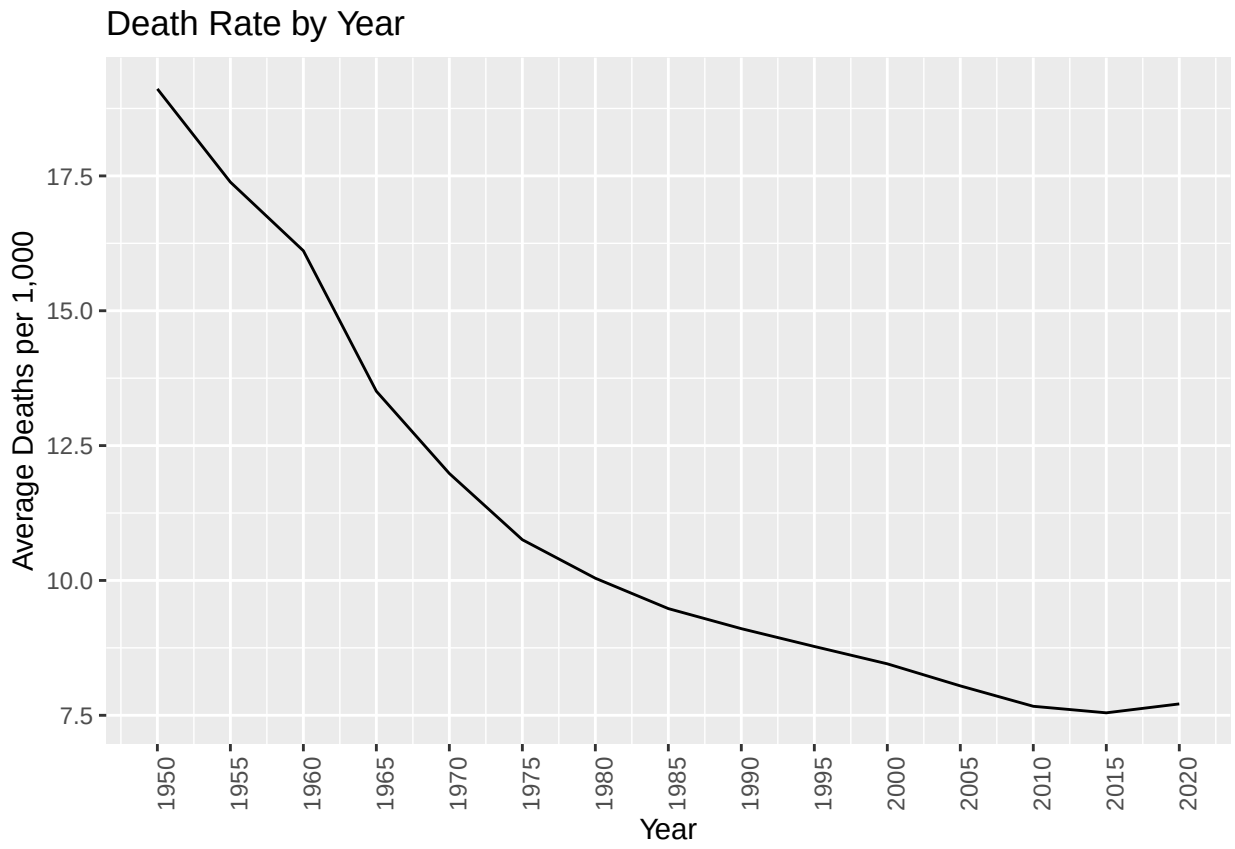
When looking at the fertility rates of the top 10 most populated countries, a signifant decrease for all countries is apparent, with six of these countries below the 2.1 level required to sustain the current population (excluding migration).



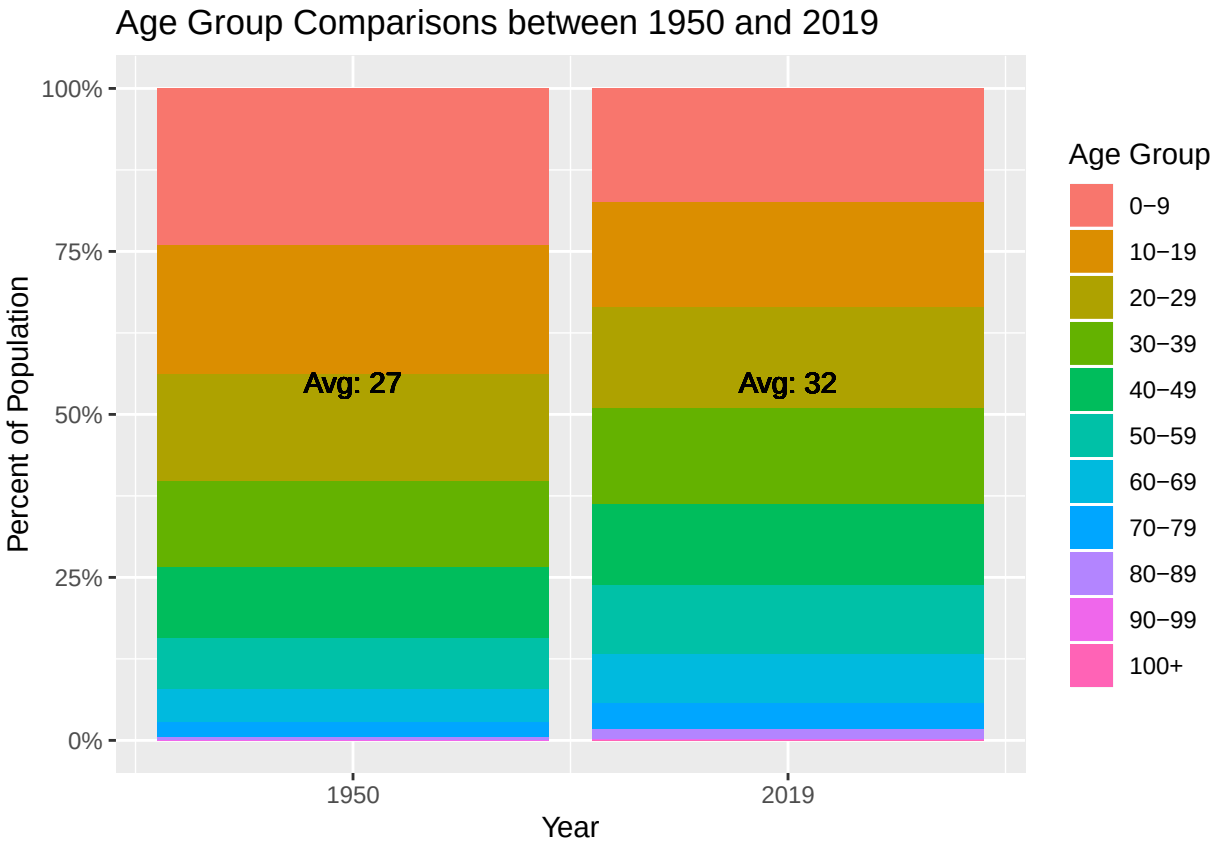
As a result of the dropping fertility rates the births per 1,000 women are steadily decreasing as shown here.



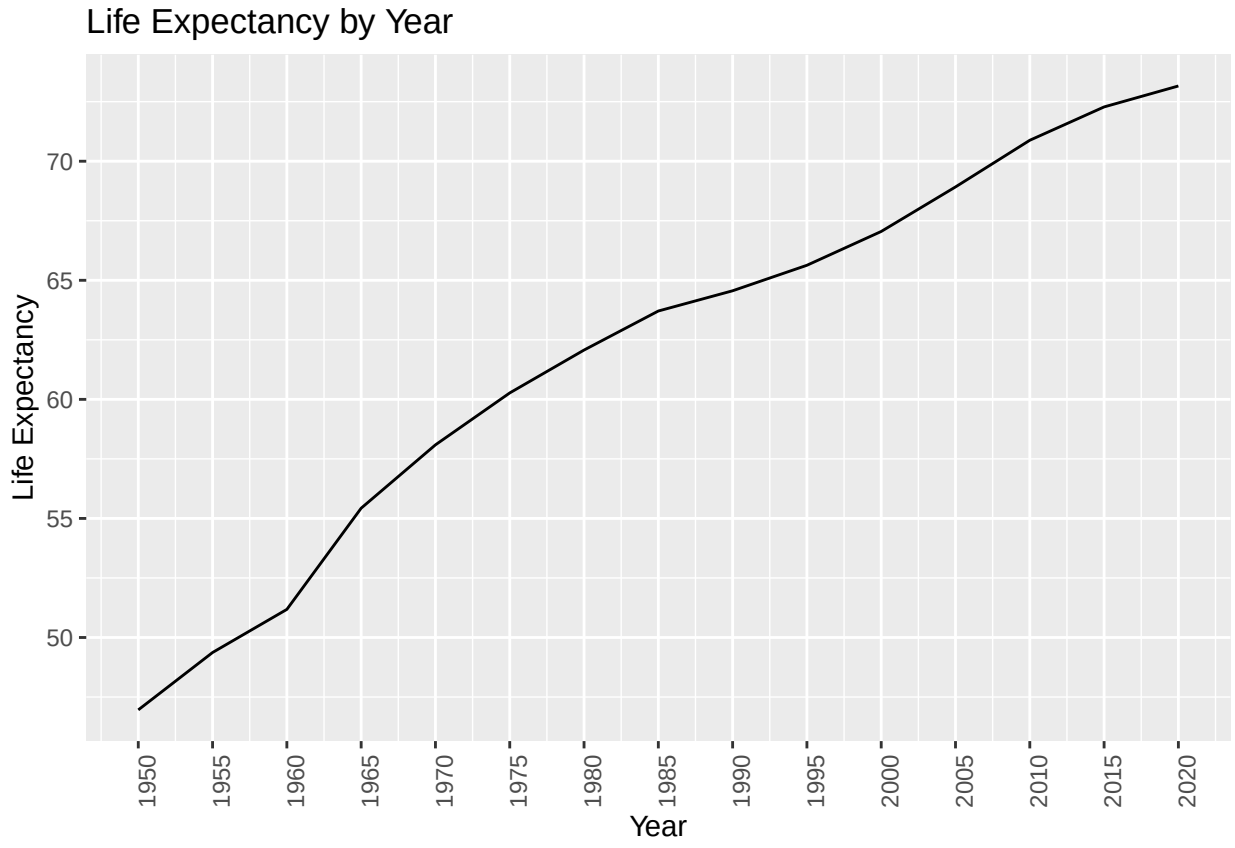
The global death rate has also decreased significanlty since the 50's. This can be contributed to improved medicine and a decrease in infant mortality rates around the world. There is better health monitoring and surveillance, improved nutrition, better access to health care, improved standards of living and increases in education. All of these are factors that contribute to a healthier society and one where people live longer.



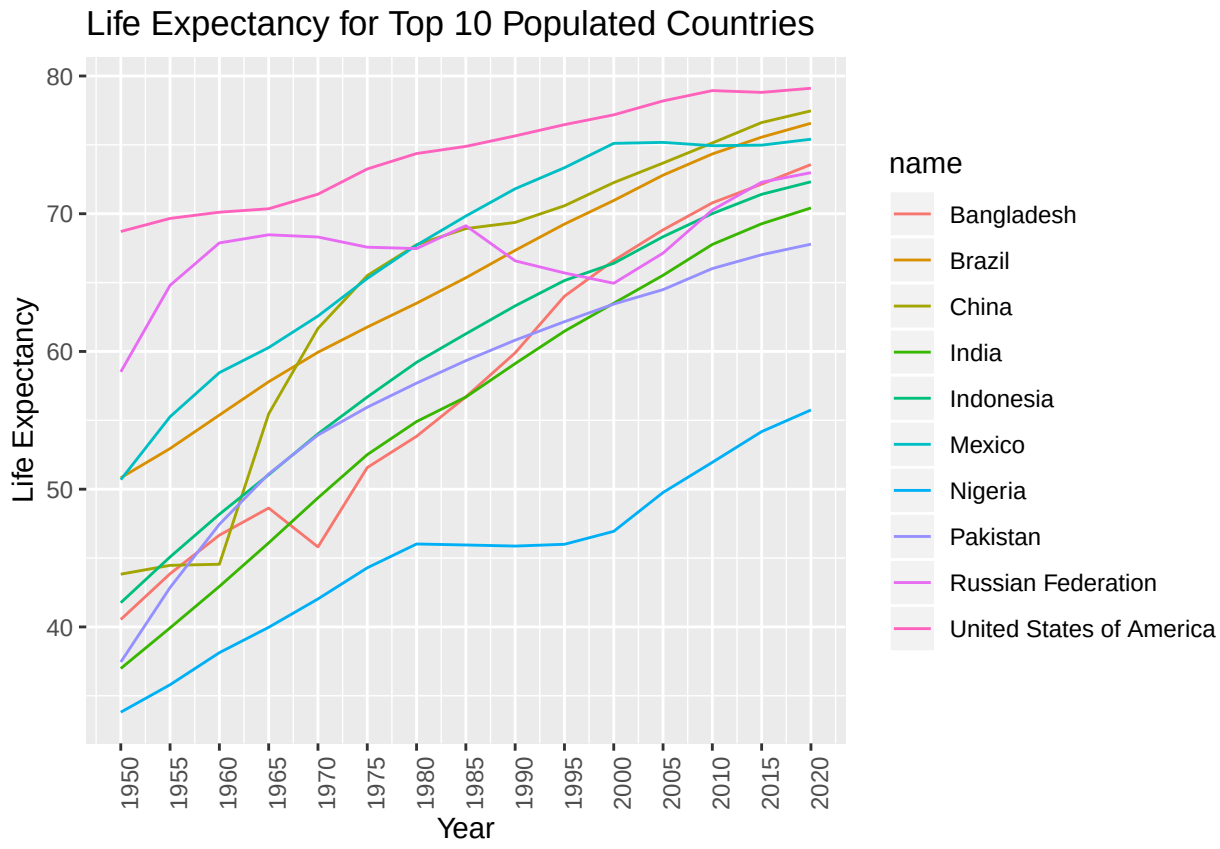
The impact of decreasing birth and death rates is causing an oveall increase in population. The average age of the population has increased from 27 to 32 years of age. Fewer births and fewer deaths is resulting in an aging population. This is contributing to the ongoing population growth.



The results of better overall health can be seen clearly by the increase in the average life expectancy around the world.



The countries with the highest population all have an increasing life expectancy.



OBSERVATIONS

The overall birth rate is decreasing globally with some large populated countries having a fertility rate less than the replacement fertility rate. This means that, ignoring migration, those countries will see a decrease in their younger population.

The global death rate is decreasing, resulting in an aging population as seen with the average global age increasing from 27 to 32 years of age and the average life expectancy also increasing significantly over the past decades.

Even though there are fewer births per 1,000 women, the fact that people are living longer has a greater impact on the overall population resulting in a consistant increase in overall population globally.

As the population ages there will be a point in the future where the death rate will increase as a result of the older population reaching their end of life. At that time, if the birth rate is still low, the global population will begin to level off and even decline.

China and India have a disproportionate percentage of the population, so what happens in these two countries can have a serious effect on the overall global population. China attempted to curb their population with the One-Child policy which had an effect, but not without it's consequences, one of which is the disproportionate number of males to females. If either of these countries have a significant shift in their population, either in birth rates or death rates, it can imact the overall global population.

The impact of global population has serious impacts on many other global elements, including economy, argriculture and climate. Understanding the impact of the population on these are crucial if life on this planet is to continue with some level of sustainability. The percentage growth over the past decades has been exponential and this level of increase cannot be sustained.

MODEL CREATION

The population was analyzed from 1950 to 2019. The data used was from the United Nations World Population data sets.

The population data set which is by age, year and country from 1950 to 2019 was the main file used for the analysis. The variant data set that contains impacts such as birth rates, death rates and life expectancy was key in understanding what was causing the trends in global population. These will be combined in order to predict the population from 2020 until 2050.

The UN also produces a data set that predicts the population from 2020 to 2100. This will be used as the validation set. It's not actual, but it is based on expert statisticians that spend serious resources on predicting population to assist other agencies and organizations to help plan for the upcoming population.

A custom predictive algorithm will be developed, that incorporates what was learned in our analysis to produce a model that tries to predict the population as closely to the UN prediction as possible. The goal is to try to be within 2%.

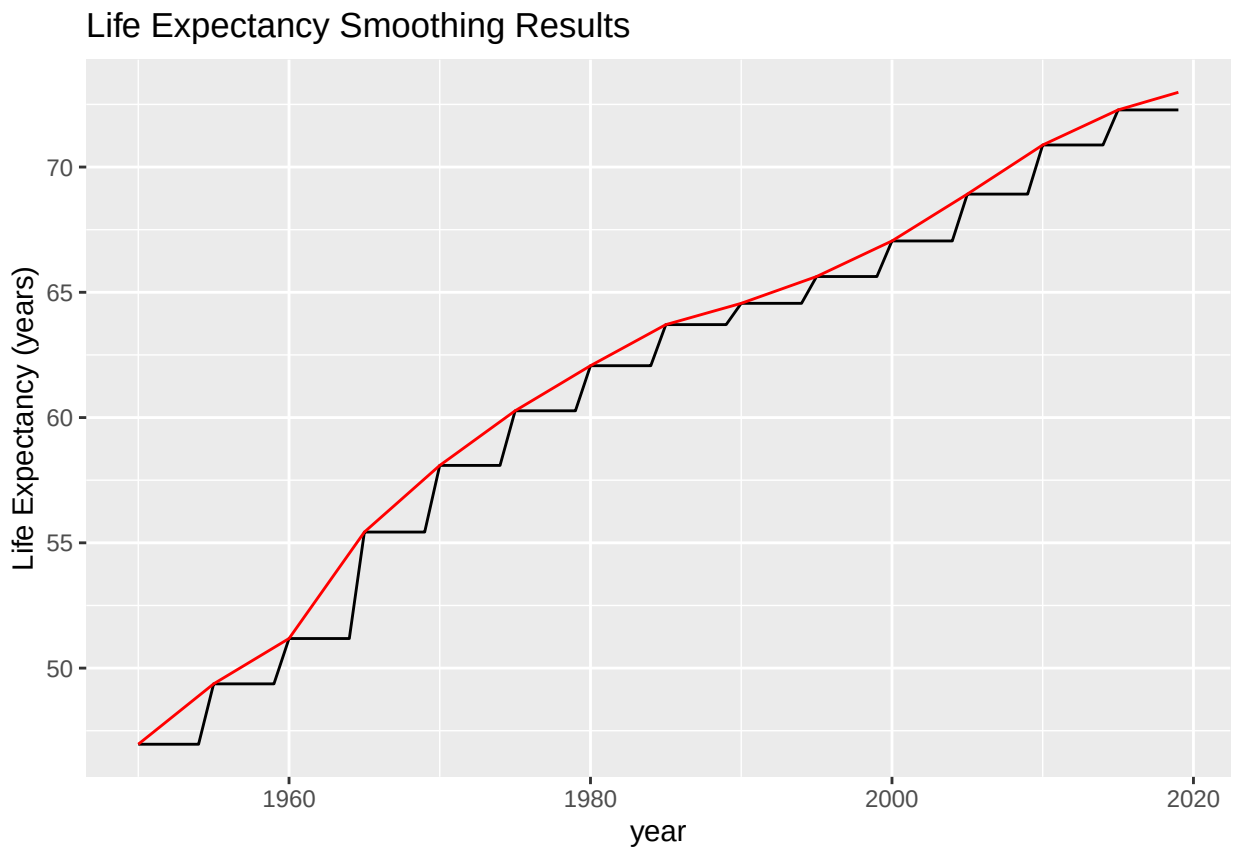
There are many effects on population, but focus will be on only three: birth rates, death rates and life expectancy.

The model will be custom. Existing algorithims like RMSE will not be used. Some of the variants and impacts to population may not cleanly fit into existing models that have been used in the course. This will be explained later as we describe the approach.

Step 1 : Convert Variant data to yearly

One of the current issues with the data is that the population data sets are yearly and the variant data sets are for five year intervals.

The five year intervals need to be evenly spread into single year totals. The life expectancy variant results are shown below as an example of the results of smoothing them out.



This will provide a clean join to the population data, which is provided by year, when building the algorithm.

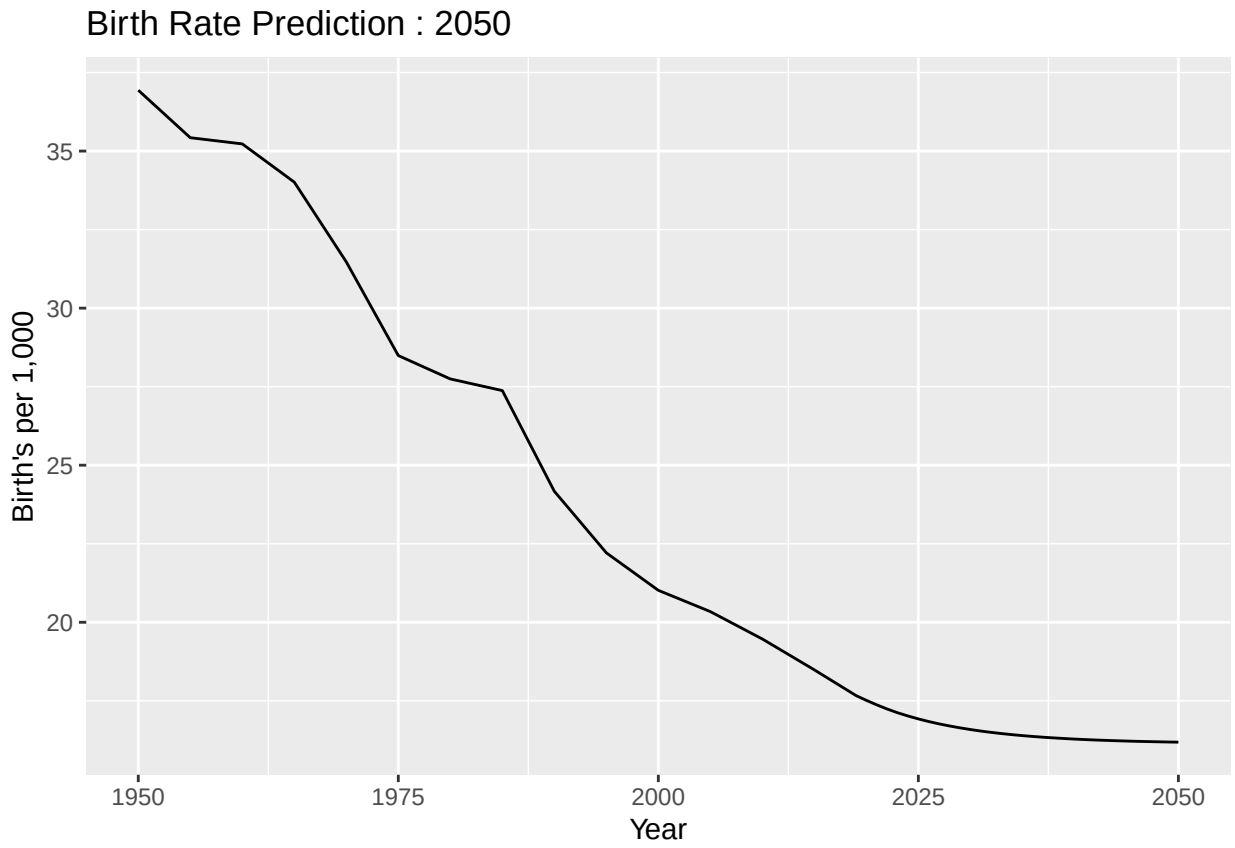
Step 2 : Predict Future Variants

The next step is to use the variants from 1950 to 2019 and predict their values from 2020 to 2050. The approach is to take the previous five year values, calculate the difference between each and take the mean value, representing the average yearly rate of change as a predictor. Five years intervals help spread out any single year anomolies and yet isn't impacted by years too far back.

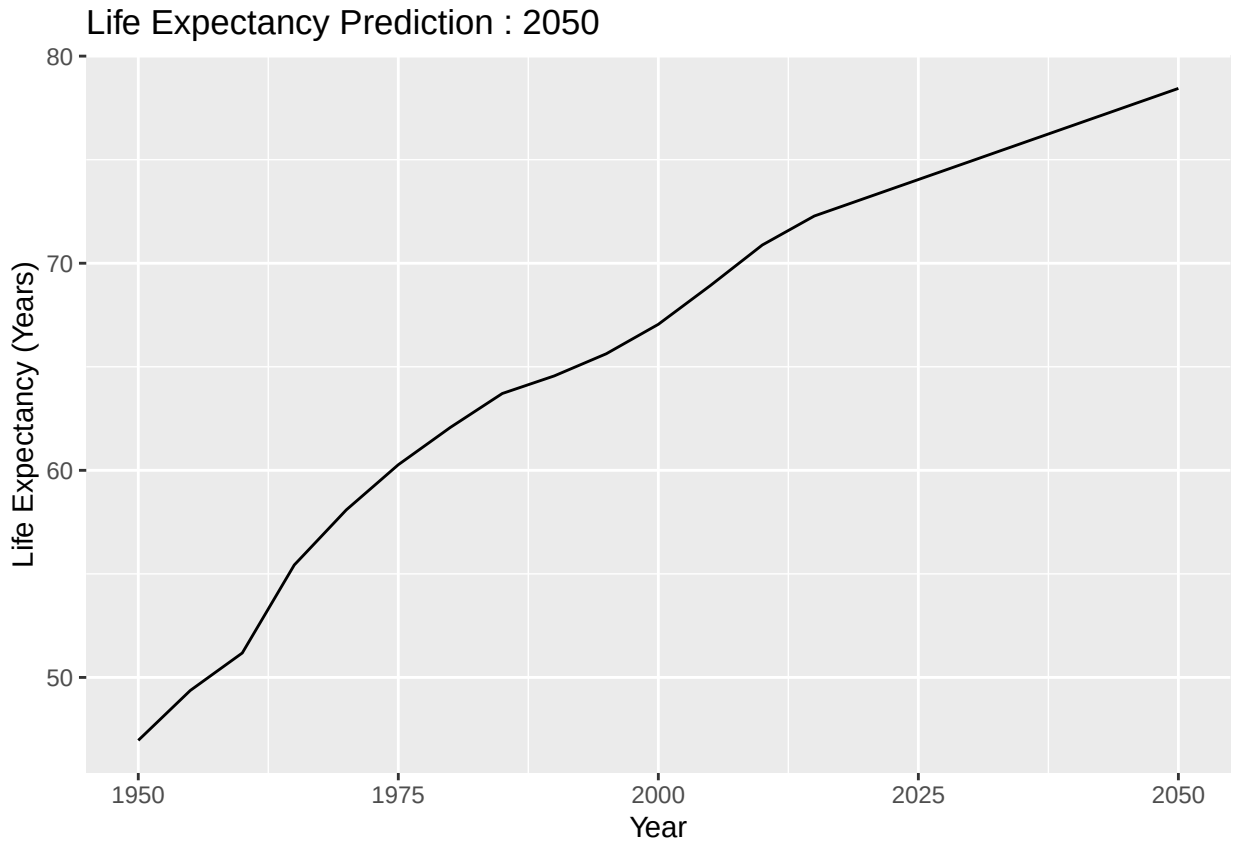
A smoothing factor is needed to this predictor because the trends are not going to be continuous. For example, if the current trend is extrapolated all the way out, there would evenutally be a birth rate of less than zero and that's just not feasible. When looking at a few factors, .25 seemed to have a gradual change when applied to each years difference. This resulted in a slowly reducing trend for both birth and death rates. The same rate was used for both, even though it is likely that they will change at different rates. Because this rate is applied to the existing rate of change, the individual trends are maintained proportionally.

The results of the three variant predictions shows us how the variants will impact the population prediction.

First, the algorithm has smoothed the birth rate to a point where it begins to level off.

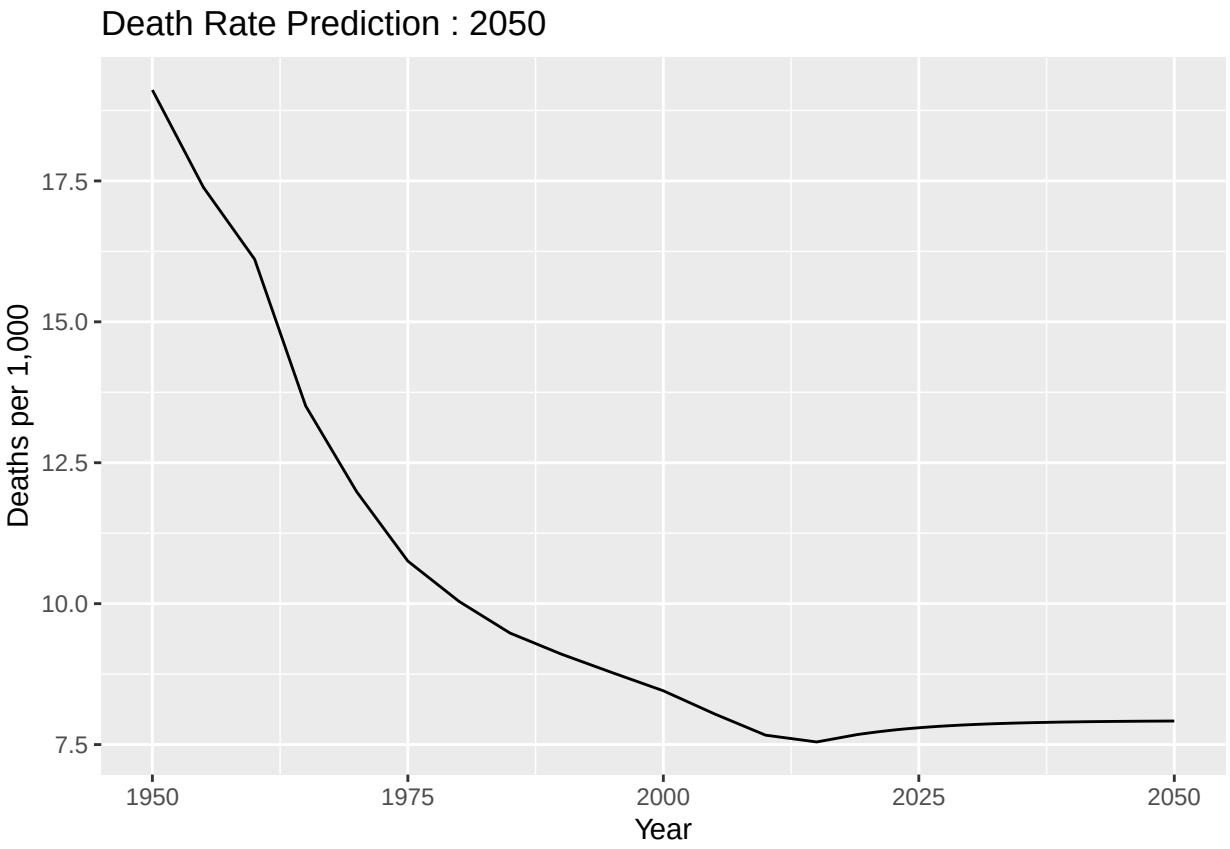


When the smoothing rate of .25 was applied to the life expectancy variant, it leveled out almost immediately. Based on research, sources say that the average life expectancy is expected to continue to increase, so the smoothing factor was removed and only the average rate of change was applied to each year.



By applying the algorithm to the death rate, the decreasing trend reversed and began to increase. An increase in the population is expected as it ages, so the death rate algorithm will remain as is. In addition to the increasing of the death rate, logic will be added to take the life expectancy into consideration. This will more accurately predict how the aging population impacts the death rate. This will be explained further in the next step.

Below is the result of the death rate prediction.



Step 3 : Predict the Population

The final step is to take the newly predicted variants of birth rate, death rate and life expectancy and apply it to our population prediction algorithm.

First, a training set of all population to 2019 is created. The set contains just the global population numbers by year and age.

Next, a validation set is created, which includes all global population predictions by age from 2020 to 2050, provided by the UN.

The final step calculates the next year in the training set and adds it back to it.

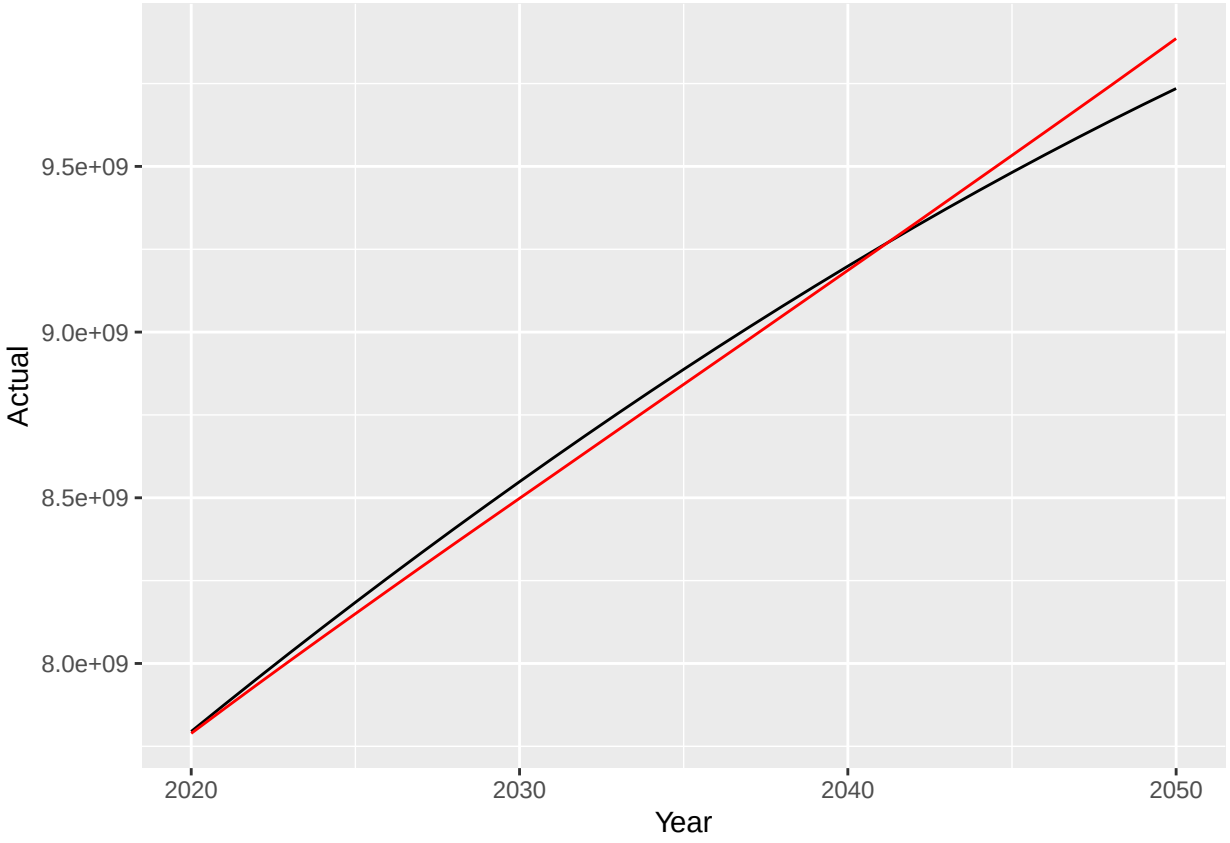
This final step executes the following adjustments and calculations:

1. The most recent year in the training set is used as a base for the calculation.
2. One year is added to the age, representing the aging of the population.
3. The total population from the most recent year is used with the predicted birth rate to calculate the new births.
4. The population over the life expectancy is extracted and the percent that it represents is calculated.
5. The rate of change in life expectancy between the recent year and the year before is used to determine the rate of the aging population.
6. The rate of change is applied to the current percent of the population over the life expectancy.
7. The death rate and life expectancy adjustment is appied to the population totals.
8. The newly predicted year is added to the training set so that this can be used in calcualting the next year

These steps are then executed for each year until it reaches 2050.

The result is a training set that contains the predictions that can now be compared to the validation set. In review, the validation set represents the predicted values from the UN.

The result of the custom predictive algorithm is a slightly lower population estimates until about 2042, when it starts overstating, however, during the 30 year prediction period, the variance is quite low overall.



Year	Prediction	UN Prediction	% Variance
2020	7,789,034,643	7,794,798,729	-0.07
2021	7,862,758,343	7,874,965,732	-0.16
2022	7,935,721,988	7,953,952,577	-0.23
2023	8,007,883,448	8,031,800,338	-0.30
2024	8,079,294,397	8,108,605,255	-0.36
2025	8,150,038,121	8,184,437,453	-0.42
2026	8,220,653,885	8,259,276,651	-0.47
2027	8,290,711,362	8,333,078,318	-0.51
2028	8,360,244,492	8,405,863,301	-0.54
2029	8,429,380,423	8,477,660,723	-0.57
2030	8,498,244,681	8,548,487,371	-0.59
2031	8,566,930,894	8,618,349,454	-0.60
2032	8,635,889,555	8,687,227,873	-0.59
2033	8,704,747,663	8,755,083,512	-0.57
2034	8,773,520,097	8,821,862,705	-0.55
2035	8,842,237,431	8,887,524,229	-0.51
2036	8,910,856,915	8,952,048,885	-0.46
2037	8,979,366,059	9,015,437,616	-0.40
2038	9,048,412,087	9,077,693,645	-0.32
2039	9,117,394,803	9,138,828,562	-0.23
2040	9,186,315,853	9,198,847,382	-0.14
2041	9,255,247,994	9,257,745,483	-0.03

Year	Prediction	UN Prediction	% Variance
2042	9,324,209,441	9,315,508,153	0.09
2043	9,393,848,938	9,372,118,247	0.23
2044	9,463,543,229	9,427,555,382	0.38
2045	9,533,278,191	9,481,803,272	0.54
2046	9,603,144,591	9,534,854,673	0.72
2047	9,673,182,102	9,586,707,749	0.90
2048	9,743,452,743	9,637,357,320	1.10
2049	9,814,505,133	9,686,800,146	1.32
2050	9,885,839,284	9,735,033,900	1.55

Overall, our average variance to UN's predictions is just `0.5`% or `99.5`% accuracy.

We can see that our algorithm starts to quickly deviate from the UN predictions after 2042. It is likely due to other variants or assumptions to the variants that we are not making. We are still reasonably accurate when we hit 2050, however, it is apparent that our algorithm needs adjustments if we were to predict beyond 2050.

SUMMARY

Our objective was to predict the global population to the year 2050. We used data sets published by the United Nations Department of Economic and Social Affairs World Population Prospects.

There are two main data sets available, a population data set and a variants dataset. We analyzed the population to 2019 to understand the trends of global and country specific population trends.

Our first major observation is that the global population has increased 204% in just over 69 years from 2.5 billion in 1950 to 7.7 billion in 2019. In 2019 the top 10 most populated countries represented 58% of the global population. China and India have a disproportionate percentage of the population of 1.8 billion combined. The Unites States is the third most populated country with 329 million, over 1 billion less than India or China.

The overal birth rate has been decreasing since the mid 1960's and in 6 of the top 10 most populated countries, the fertility rate is below the replacement fertility rate. The replacement fertility rate is the rate needed to sustain the current population of a country, ignoring migration.

Death rates are decreasing as the population gerneally live in better conditions and medicine is able to prolong life by curing or minimizing the impact of various diseases. As a result people are living longer and therefore the life expectancy has been increasing.

Even though the birth rates are decreasing, the population has been increasing because the number of deaths is significanly less. This is resulting in population growth, but also an aging population. Life expectancy has risen from an average of below 50 to over 70 and it is anticipated to increase to close to 80 globally by 2050.

Our model used three varaints out of many available, to try to predict the population from 2020 to 2050. We decided to develop our own algorithm and not use existing methods as we learned in the course. We did this for two reasons. First to learn more by building it ourselves and secondly the variants are more complicated than just applying them. It was felt that we needed to take special approaches to predicting such as applying smoothing factors, but more importantly, by aging the population each year in our algorithm.

This resultsed in a reasonably accurate estimate, however it became apparent that the algoithm would not sustain it's accuracy much beyond 2050. In order for that we would have to make some manual adjustments and we do not have enough knowledge to be able to so at this stage.

In the end we were able to predict the yearly global population from 2020 to 2050 with an average `99.5`% accuracy and a final 2050 prediction of 9.89 billion compared to the UN's 9.74 billion. A variance of 1.55% and an accuracy of 98.45%.