Predicting Populations: Modeling Demographic Predictions for Nations Around the World Using Population Pyramids and Demographic Transition Models

Michael Zheng

Abstract -

Examining and analyzing demographic statistics are a key foundation for analyzing the makeup of a population for a given nation. Some of the most important population demographics include total population, crude birth rate(CBR - births/1000 people), crude death rate(CDR - deaths/1000 people), and populations of different age groups and genders. These help provide a basis into understanding the overall population growth and stage of development of a particular country, which can ultimately be used to shape economic, political, and social policies for different nations around the world. In this research project, I used multiple machine learning models(neural network and linear regression) in order to predict key demographic statistics over the next 5 years for each nation. These models yielded accurate results, as most sets of predictions had accuracy scores above 0.85. Additionally, using advanced visualization techniques, I modeled these predictions using popular human geography models such as population pyramids and demographic transition models. I found that for more developed nations such as countries in Western Europe and East Asia, the population of older individuals is going to rise significantly over the coming years. For less developed nations such as those in Africa, the overall population is going to skyrocket, and the proportion of young adults who will enter the labor force will significantly surge. In terms of real-world implications, these results could lead to more developed nations starting to spend more on social security and less developed nations starting to implement policies that limit population growth.

Introduction -

My research question that I looked to answer through this project is: How will population demographics change for nations around the world over the next 5 years? Finding answers for this is incredibly useful because it provides valuable insight into each nation's demographics for the future and the trends that will occur, which allows nations to plan ahead for these changes and implement proactive policies in order to effectively deal with the ever-changing population. One example of this in action is the one-child policy in China, in which China restricted couples to giving birth to only one child in order to combat overpopulation, before it was ultimately

repealed due to concerns about a future shortage in working-class individuals. In my project, since I am going to be predicting statistics, my project uses regression and works with numerical figures. I use supervised learning, in which I train my neural network and linear regression models to ultimately make accurate predictions for various demographic statistics over the next 5 years.

Background -

Overall, much of the previous research on using machine learning for demographic predictions has involved a variety of models to make mostly small-scope predictions using minimal data. For example, one such research project involved making population forecasts for major cities in Taiwan, in which the authors used LSTM and XGBoost models to make predictions, and using mean absolute percentage error(MAPE) they found that XGBoost models were more effective(Wang and Lee et al, 2021). However, one shortcoming was how minimal their data was, as they only used demographic data from the last five years. Thus, the model forecasted some demographic statistics such as immigration very poorly.

The vast majority of research papers on this topic make predictions only for the total population of a specific region or nation. For example, one research paper specifically modeled total population projections in Azerbaijan, using a variety of models(KNN, Decision Tree, Random Forest) and metrics(MAPE, R^2)(Hajirahimova and Aliyeva et al, 2023).

Overall, much of the previous research has been exceptional in terms of model variety and comparing results of predictions from different models. However, one overarching shortcoming is that the minimal data that is used in these predictions leads to errant predictions, as the models aren't trained with enough data. Additionally, these previous models make predictions for very specific regions but aren't generalized to other nations around the world. Thus, through my approach, I looked to build upon the positive aspects of previous research by using multiple models and validation techniques(neural network, linear regression, r2_score, accuracy_score), while also counteracting the shortcomings of previous research by using extensive data from the past 60 years to train my models and by creating expansive models that can be generalized to nations all around the world.

Dataset -

I extracted demographic data from the World Bank into a CSV file, which I used as the initial dataset for my project. The dataset includes all of the statistics that I want to predict, such as total population, total male population, total female population, crude birth rate, crude death rate, and dependency ratio(% of working age population). It also included statistics of the percentage of each age group's population for a given gender. Each age group is in intervals of 5 years and is split by gender. For example, ("male ages 0-4, % of total male population", "female ages 75-79, % of total male population", etc.) are each a statistic for each country in this dataset. I used 198 countries in total, as that was the number of countries in which population data was available over the last 62 years. My initial dataset is mostly numerical and is 10494 rows by 64 columns. The rows represent each statistic for each country. I used 53 statistics for each country from 198 countries, which makes 198 times 53 = 10494 rows. Of the 64 columns, 2 represent categorical variables, as they denote the country name and variable. The next 62 columns represent the data from each year from 1961-2022(62 years).

Here is what the first five rows of the initial dataset look like:

	Country Name	Series Name	1961 [YR1961]	1962 [YR1962]	1963 [YR1963]	1964 [YR1964]	1965 [YR1965]	1966 [YR1966]	1967 [YR1967]	1968 [YR1968]	...
0	Afghanistan	Birth rate, crude (per 1,000 people)	5.044300e+01	5.057000e+01	5.070300e+01	5.083100e+01	5.087200e+01	5.098600e+01	5.108100e+01	5.114800e+01	...
1	Afghanistan	Death rate, crude (per 1,000 people)	3.134900e+01	3.084500e+01	3.035900e+01	2.986700e+01	2.938900e+01	2.887200e+01	2.839600e+01	2.789500e+01	...
2	Afghanistan	Age dependency ratio (% of working-age populat...	8.022234e+01	8.040194e+01	8.071532e+01	8.121956e+01	8.199482e+01	8.295472e+01	8.390807e+01	8.484415e+01	...
3	Afghanistan	Age dependency ratio, old (% of working-age po...	5.078088e+00	5.049558e+00	5.022020e+00	4.999719e+00	4.985287e+00	4.979162e+00	4.976931e+00	4.976000e+00	...
4	Afghanistan	Age dependency ratio, young (% of working-age ...	7.514425e+01	7.535238e+01	7.569330e+01	7.621984e+01	7.700953e+01	7.797556e+01	7.893114e+01	7.986815e+01	...
...	...	...	...	...	...	...	...	...	...	...	...

Then, I used data preprocessing to craft a multitude of datasets. First, I created three datasets for predicting solely the age group population percentages("male ages 0-4, % of total male population", "female ages 75-79, % of total male population", etc.). I created an input

dataset(x variable) that would comprise age group statistics from 10-year intervals, and I created an output dataset(y variable) that would comprise age group statistics from the ensuing 5-year interval. This is because essentially, **I am training my model to make predictions for the next 5 years based on the previous 10 years**, and I use these 10-year and 5-year intervals over the last 62 years to repeatedly train my model. Both of these two aforementioned datasets are strictly for model training only. Then, I create a third dataset that contains data from the previous 10 years(2013-2022), which I use to ultimately predict the next 5 years(2023-2027). The reason why I chose 5 years is because the model starts to yield relatively inaccurate predictions past that point.

Here is a look at the input dataset(x variable) for model training:

*10 year intervals for each nation(Afghanistan_1963-1972, Zimbabwe_1988-1997, etc.)

	Population ages 00–04, female (% of female population)_year1	Population ages 00–04, male (% of male population)_year1	Population ages 05–09, female (% of female population)_year1	Population ages 05–09, male (% of male population)_year1	Population ages 10–14, female (% of female population)_year1	Population ages 10–14, male (% of male population)_year1	Population ages 15–19, female (% of female population)_year1
Afghanistan_1963-1972	17.758778	16.938428	13.543510	12.926272	11.672353	11.002399	10.746217
Afghanistan_1968-1977	18.236691	17.671171	13.971734	13.551710	11.684286	11.337764	10.032217
Afghanistan_1973-1982	18.613482	18.258943	14.361766	14.098073	11.946299	11.735827	9.956299
Afghanistan_1978-1987	18.900686	18.725819	14.690599	14.557110	12.189574	12.081642	10.107864
Afghanistan_1983-1992	18.842613	19.003940	14.956103	15.119666	12.488520	12.598265	10.356533
...	...	...	...	...	...	...	...
Zimbabwe_1988-1997	17.500841	18.362543	16.458917	16.878289	14.066749	14.307306	11.547052
Zimbabwe_1993-2002	14.800768	16.021688	16.000256	16.366869	15.199321	15.216557	12.632678
Zimbabwe_1998-2007	13.977304	15.570476	13.211204	14.122752	14.597698	14.834634	13.754373
Zimbabwe_2003-2012	15.141712	16.882249	12.694865	13.912200	12.211846	12.835939	13.306484
Zimbabwe_2008-2017	14.819415	16.615082	13.944932	15.252407	11.956620	13.009578	10.989064

Output dataset(y variable) of all the age group percentages:

*Ensuing 5 year intervals for each nation(Afghanistan_1973-1977, Zimbabwe_1998-2002, etc.)

	Population ages 00–04, female (% of female population)_year1	Population ages 00–04, male (% of male population)_year1	Population ages 05–09, female (% of female population)_year1	Population ages 05–09, male (% of male population)_year1	Population ages 10–14, female (% of female population)_year1	Population ages 10–14, male (% of male population)_year1	Population ages 15–19, female (% of female population)_year1
Afghanistan_1973-1977	18.613482	18.258943	14.361766	14.098073	11.946299	11.735827	9.956299
Afghanistan_1978-1982	18.900686	18.725819	14.690599	14.557110	12.189574	12.081642	10.107864
Afghanistan_1983-1987	18.842613	19.003940	14.956103	15.119666	12.488520	12.598265	10.356533
Afghanistan_1988-1992	18.961469	19.614596	14.777227	15.328524	12.651028	13.108685	10.604869
Afghanistan_1993-1997	19.870548	20.687133	14.930632	15.514009	12.255435	12.744828	10.483943
...	...	...	...	...	...	...	...
Zimbabwe_1998-2002	13.977304	15.570476	13.211204	14.122752	14.597698	14.834634	13.754373
Zimbabwe_2003-2007	15.141712	16.882249	12.694865	13.912200	12.211846	12.835939	13.306484
Zimbabwe_2008-2012	14.819415	16.615082	13.944932	15.252407	11.956620	13.009578	10.989064
Zimbabwe_2013-2017	15.352441	17.394693	13.203936	14.512547	12.726118	13.823210	10.524313
Zimbabwe_2018-2022	14.271254	16.182629	13.595811	15.142543	11.830499	12.835855	11.213290

For predicting other statistics such as total population, CBR, CDR, etc., I didn't create any new datasets and instead used the aforementioned initial dataset described above.

Methodology/Models -

For predicting the age group population percentages("male ages 0-4, % of total male population", "female ages 75-79, % of total male population", etc.) for nations over the next 5 years(2023-2027), I used a neural network model, specifically MLPRegressor. Remember, for predicting age group percentages, I created three datasets: an input dataset(x variable), a corresponding output dataset(y variable), and a dataset with input data from the last 10 years(2013-2022) for making future predictions. All of the datasets are 2-D Pandas DataFrames. First, I split the first two datasets into x_train, x_test, y_train, and y_test. My test ratio was 0.2, so essentially 20% of the input and output data was dedicated to testing or validation, and 80% of the input and output data was dedicated to training the neural network.

The way the neural network is trained is that it takes in each row of the corresponding input and output training data one at a time. Each row of the input dataset is all of the age group percentages for a given country during a given 10-year interval(e.g.: Afghanistan_1963-1972, Zimbabwe_1988-1997, etc.). Each corresponding row of the output dataset is all of the age group percentages for a given country during the 5-year interval afterwards(e.g.: Afghanistan_1973-1977, Zimbabwe_1998-2002, etc.). There are over 1,000 of these rows, as we took multiple corresponding 10-year and 5-year intervals from each nation.

For the neural network configuration, I did not implement any hidden layers because the additional hidden layers leave the model prone to overfitting. Because the input and output datasets aren't particularly convoluted, not using hidden layers simplified the model and made it more effective. For maximum iterations, I set it at 7,000 because the neural network needs ample repetitions in order for it to be fully trained properly. Then, after training the model, I used the third dataset(data from the past 10 years) to make predictions over each of the next 5 years for each nation's age group population percentages. These predictions were ultimately stored in a comprehensive 198 by 170 DataFrame, as I made predictions for 198 countries, and for each country, I predicted 34 age group percentages(17 for each gender) for each of the next 5 years, thus totaling 34 times 5 = 170 columns.

Here is the DataFrame storing the predicted values:

	Population ages 00–04, female (% of female population)_2023	Population ages 00–04, male (% of male population)_2023	Population ages 05–09, female (% of female population)_2023	Population ages 05–09, male (% of male population)_2023	Population ages 10–14, female (% of female population)_2023	Population ages 10–14, male (% of male population)_2023
Afghanistan	15.806836	16.258599	13.993252	14.386427	12.625034	12.745938
Albania	4.685114	5.041833	5.232487	5.779973	5.603071	5.794536
Algeria	10.383177	10.483466	10.370067	10.649694	9.142702	9.573218
Angola	16.800507	17.300638	14.535380	15.070055	12.807623	13.085785
Antigua and Barbuda	5.316073	5.599748	5.389643	6.159802	6.387972	7.129607
...	...	...	...	...	...	...
Vietnam	6.855214	7.669619	7.079308	8.054385	7.049818	7.749833
Virgin Islands (U.S.)	5.614879	6.220776	5.713649	6.962688	6.422270	7.016528
Yemen, Rep.	13.735584	14.052515	12.865522	13.261426	11.986691	12.202559
Zambia	15.265719	15.580645	13.754924	14.230703	12.902498	13.041527
Zimbabwe	13.187766	14.771760	12.712394	14.161001	12.241327	13.170573

198 rows × 170 columns

Then, I used linear regression to predict statistics such as CBR, CDR, total population, total female population, and total male population. For each variable, I modeled the line of best-fit using data from the past 10 years(2013-2022). This is because 10 years is the perfect balance between a good enough sample size and also relevance of data towards predicting the future. For each statistic, I extracted that corresponding row from the initial dataset for each nation and used the last 10 data points(last 10 years) of that row to train a linear regression model. My x variable was the year, and my y variable was the statistic that I wanted to predict. I used a new linear regression model for each given nation and each given statistic.

I used linear regression because its simplicity matched the type of data that was being used to make predictions, as linear regression is not prone to overfitting unlike other methods and produces consistently reliable predictions. The way linear regression works is that it takes in x and y data, and it calculates a line that minimizes the squares of the residuals of the plot. This line is used to interpolate and extrapolate predictions. While it isn't good at handling complex and fluctuating plots, it consistently produces reasonably valid predictions and does not overfit the data. Initially, I intended to do linear regression on the multipliers of these statistics for each of the last 10 years. A multiplier is how many times greater a quantity was from one year to the next. Trying to account for the curved nature of population graphs, I initially thought that predicting the future multipliers and using those multipliers to compute the numerical values for each statistic was a good idea. However, it turned out that trying to account for curved growth led to wildly erratic predictions and incredibly low accuracy. Thus, I stuck with linear regression

on the numerical statistics rather than the multipliers, as that was incredibly more reliable and consistently produced relatively high accuracy.

Then, using the line of best-fit calculated for each statistic and each nation over the last 10 years, I used that to predict the corresponding data for each of the next 5 years. For my line of best-fit equation, the x variable represented the year, and the y variable represented the predicted value of that statistic. For example, if I wanted to predict a particular value in a country for the year 2027, then my x value would be 2027, and my y value would be the corresponding predicted value for that statistic in that given year. I stored all of these predictions in separate DataFrames for each statistic, so I have five DataFrames for these five statistics(CBR, CDR, total population, total male population, and total female population), and each DataFrame has predictions of that statistic for each nation over the next five years.

Results and Discussion -

In terms of accuracy, both my neural network and linear regression models produced mostly accurate results and were effective predictors. For my neural network for predicting each of the age group percentages, I used the r2_score metric from scikit-learn as my validation technique for modeling the accuracy of the predictions. I compared my predictions and the actual y_test in order to produce an accuracy score, which ended up being around 0.98, a very strong correlation.

For the linear regression models I used on each of the five statistics(CBR, CDR, total pop., total female pop., and total male pop.), the accuracy of the models was high for each statistic except for CDR. As stated earlier, for each statistic, I reused and refitted a linear regression model for each country. Thus, the way that I calculated accuracy for each statistic was by averaging the coefficient of determination(R-squared) scores that the linear regression model produced for all nations. For example, for calculating the accuracy of using linear regression to predict CBR, I averaged all of the coefficients of determination scores for each country's CBR plot. Ultimately, for total pop., total female pop., and total male pop., the average coefficient of determination scores were each between 0.904 and 0.910. For CBR, the average coefficient of determination score was around 0.842. However, for CDR, the average coefficient of determination score was drastically lower, at around 0.618.

Here is a table of all of the accuracy results for each statistic:

Statistic:	Accuracy(nearest ten-thousandth): calculated using R^2 score(coefficient of determination)
Age group percentages(e.g.:“male ages 0-4, % of total male population”)	0.9792
CBR	0.8425
CDR	0.6185
Total population	0.9093
Total female population	0.9048
Total male population	0.9048

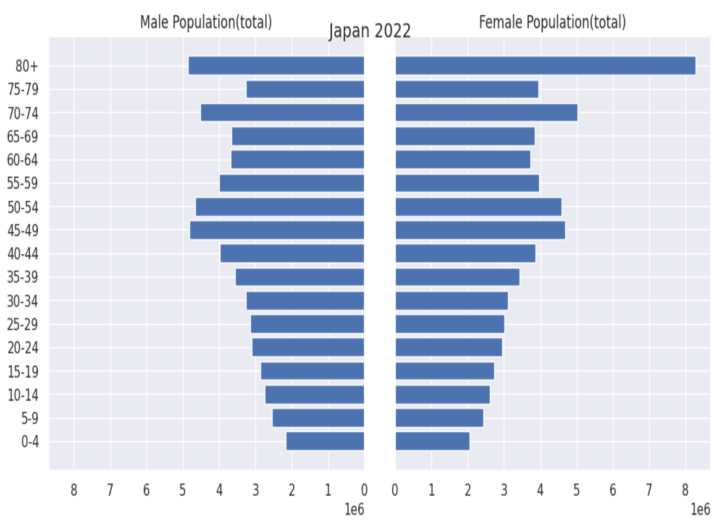
One potential reason for the CDR linear model having much lower accuracy than the other metrics is due to the rampant increases in deaths during the pandemic, which distorted the shape of the plot. One way my model can better account for this is to gather data of countries that had the most deaths attributed to COVID-19 and subtract those deaths from the total deaths. This would make the data and predictions less skewed by the uncharacteristically high death rates caused by the virus. Additionally, CBR and CDR are expected to have lower accuracy in general because these statistics fluctuate more compared to total population, total female population, and total male population.

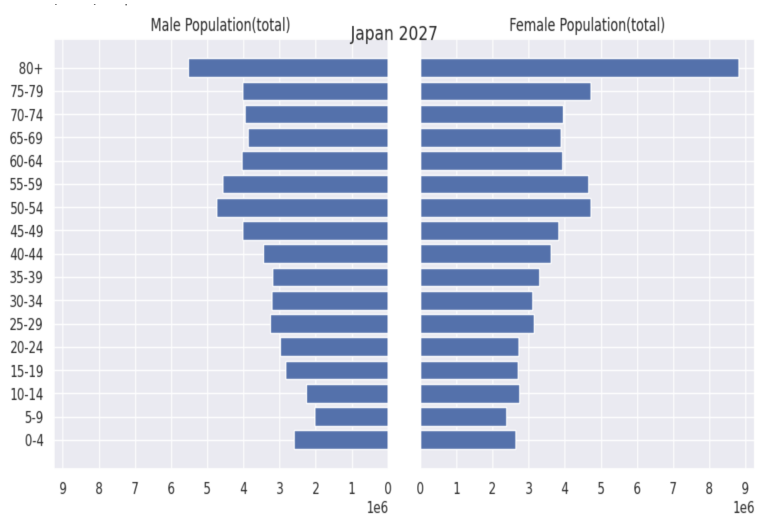
For visualizing predictions, I created two types of plots: population pyramids and demographic transition models. Population pyramids display the total population of each age group for each gender in a back-to-back horizontal bar graph. For calculating the total population of each age group for each gender, I multiplied the predicted age group percentages by the predicted total female and male populations. Demographic transition models display the total population, CBR, and CDR on a line plot, with the year as the x variable.

In terms of big-picture trends or takeaways from the predictions, the results largely depended on a nation’s level of development. For more developed nations such as those in Western Europe or East Asia, the proportion of older people(65+ years) is going to surge over the next 5 years. In these nations, the population will stagnate, and the CBR and CDR are going to

be roughly around the same level. For particularly developed nations with significantly old populations such as Japan, the death rate may even rise.

Here is the most recent population pyramid for Japan vs that projected for 2027:

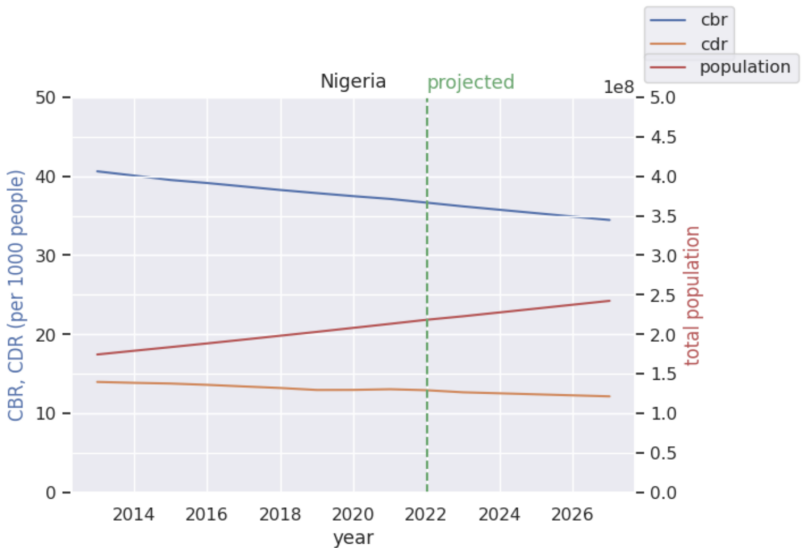




Key trends for Japan over the next 5 years: The 0-4 age group population will increase significantly(more births), the 80+ population will increase, and the 5-9 and 10-14 age groups will decrease(less future workers).

For less developed nations such as those in Africa or Latin America, the proportion of young adults will tremendously increase, leading to a growing workforce. In these nations, the CBR is going to decrease as the nations start to become more developed. However, the population will still continue to grow rapidly, as CBR is much greater than CDR in these countries.

Here is the demographic transition model for Nigeria(2013-2027):



Key takeaways: the CBR and CDR will continue to decrease, with CBR decreasing at a faster rate, and the total population will continue to increase over the next 5 years.

In terms of actionable policy insights that can be derived from these results, one such example would be increasing expenditures on social security in Japan. As seen from the Japan population pyramid, the elderly population will significantly increase over the next 5 years, especially those over 80+ years of age. Thus, this would require more pensions and more government spending to take care of these elderly citizens.

Conclusion -

Ultimately, using neural network and linear regression models to predict various major demographic statistics for each nation over the next 5 years, my predictions were largely valid and accurate, which can be applicably used to extract insights about the demographics of various nations and to inform economic, social, and political policy decisions for national governments. Using a neural network to predict age group percentages was highly successful due to the complexity of the data, while using linear regression to predict the other statistics(CBR, CDR, total pop., total female pop., total male pop.) was also relatively successful due to the simplicity and linear nature of these plots. One way these predictions can be improved upon is that I could have taken into account other variables when making these predictions, such as environmental or economic factors. This would have added more nuance and could have potentially led to even more accurate predictions. Overall, this work can be built upon by making predictions for other related demographic statistics that could take into account other factors besides age such as race or socioeconomic status.

Acknowledgments -

Special thanks to my mentor Kasra Koushan for giving me valuable advice and tips on this project, as well as giving me insightful information on publishing an effective research paper.

References -

"Data Bank World Development Indicators." *The World Bank*,

databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=.

Accessed 16 Nov. 2023.

Hajirahimova, Makrufa, and Aybeinz Aliyeva. "Development of a Prediction Model on Demographic Indicators based on Machine Learning Methods: Azerbaijan Example." *MECS Press*, 8 Apr. 2023, www.mecs-press.org/ijeme/ijeme-v13-n2/v13n2-1.html. Accessed 16 Nov. 2023.

Wang, Chian-Yue, and Shin-Jye Lee. "Regional Population Forecast and Analysis Based on Machine Learning Strategy." *National Library of Medicine*, 24 May 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8225119/. Accessed 16 Nov. 2023.