# Predicting Population Growth Rates from Socioeconomic and Demographic Indicators

DATA 110 Final Project Write-up

**Team Name: The Outliers**

Anderson Beck, Chase Brister, Jack Bingen, Raphael Koo

December 5, 2025

## 1 The Problem

Here's the big question this project addresses: What are the main socioeconomic and demographic factors that actually drive a country's population growth rate each year? This matters a lot because a country's annual population growth rate is the foundation for all of its long term planning, from deciding how many schools and hospitals to build, to figuring out retirement and social security needs. By looking at factors like a country's GDP per capita, its fertility rate, and life expectancy, we wanted to see how each correlates to population growth and see if any are true drivers of population change. Understanding these relationships helps explain the huge differences we see around the world such as why some countries are growing quickly and why others are dealing with a shrinking or aging population. This project uses real-world data from Our World in Data to offer policymakers and international organizations better insights into global challenges like the "fertility crisis" in developed nations and unsustainable growth elsewhere.

## 2 Data Collection

Starting the project wasn't the easiest because there wasn't a single, perfect dataset that had all the variables we needed standardized together. So, we had to come up with a way to create our own. We ended up combining several different datasets from Our World in Data into a single standardized dataset. We pulled records for countries' annual population growth rates, fertility rates, life expectancy from birth, and GDP per capita, over different years. Our main hypothesis in the beginning was that the trio of fertility rate, life expectancy, and GDP per capita would be the biggest forces influencing a country's overall population change.

# 3 Preparation

Once we had the raw data, the first job was cleaning and prepping it. This involved taking our four separate files and augmenting them together into one large dataset using a merge on the common fields: the country (Entity), the country Code, and the Year. We made sure to rename the long, technical column headers, for example, changing 'Period life expectancy at birth' to 'Life Expectancy' which was much easier to read. We also unified two different population growth columns into one clear 'Population Growth Rate' feature through using `.fillna()`. We had to drop any rows that had missing values, which standardized the dataset for analysis, bringing our rows from over 40,000 to only 11,545. Before feeding the data into our model, we removed the country name column ("Entity") since it was categorical and not necessarily useful for our numerical regression. While we could have applied one-hot encoding, doing so would have complicated our correlation analysis and added unnecessary dimensionality. By leaving out the country names, the model relies strictly on the numerical features of life expectancy, GDP per capita, and fertility rate, making it more effective in situations where the country may be unknown or when estimating values for a new region. We then normalized all numerical features to ensure that variables with large ranges (such as GDP per capita) didn't disproportionately influence the model.

# 4 Data Exploration

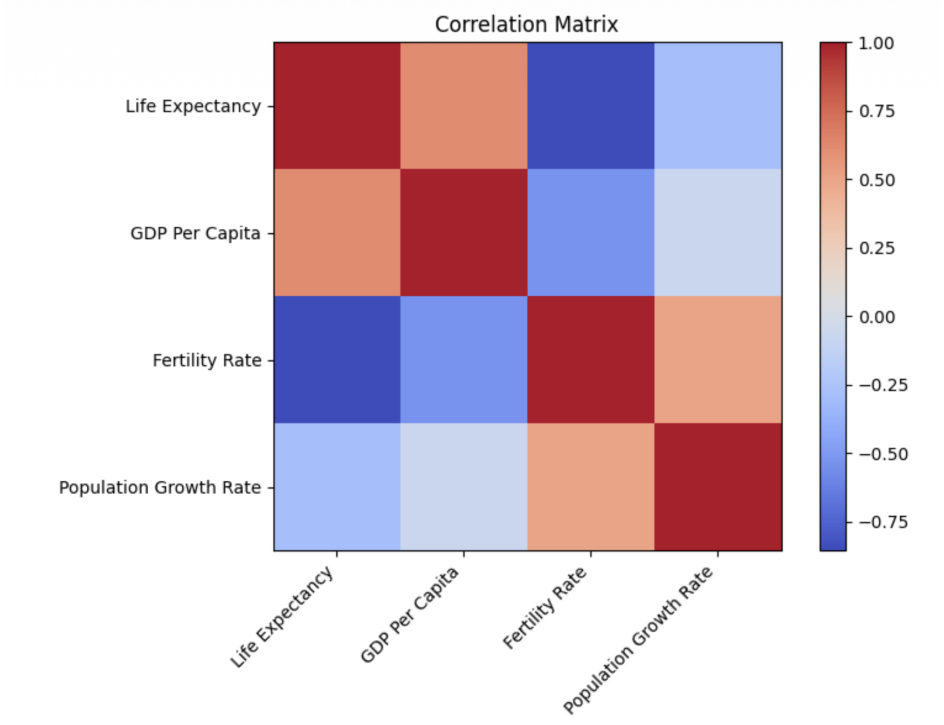


Figure 1: **Correlation Heatmap:** Showing relationships between variables.

Before building any models, we spent time exploring the data to see any patterns or correlations we could find. We used scatter plots to visualize relationships, which immediately showed clear patterns: countries with higher fertility rates also showed higher population growth, and places with a higher GDP per capita tended to have longer life expectancy. Our histograms revealed that the data for fertility rates and GDP per capita was skewed, while life expectancy had a much more typical, normal or gaussian distribution. The most important finding came from the correlation heat map, which clearly indicated that the fertility rate had the strongest, most linear connection to the population growth rate. This strong relationship made us confident that a Linear Regression model would be a great starting point. We also built time series plots that let us track these demographic and economic trends for any given country over many years.

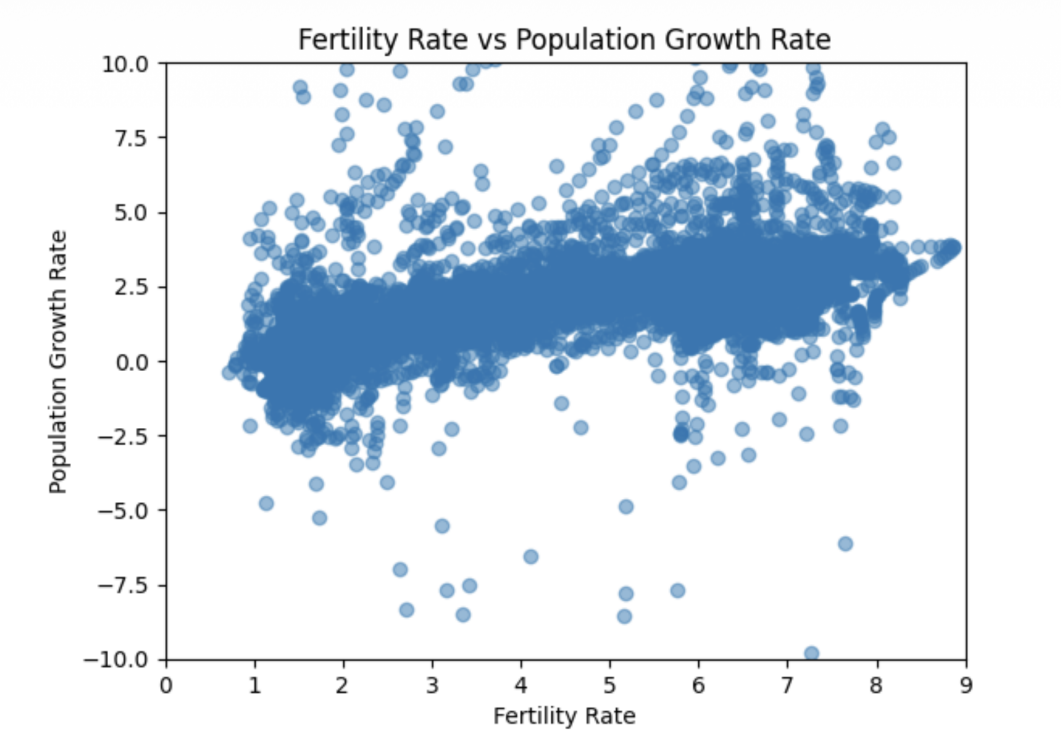


Figure 2: **Fertility Rate vs. Population Growth:** Visualizing the strong positive correlation.

# 5 Model Building

With our data prepared, we moved on to the modeling part of the project where we could use the relationships we had uncovered to predict population growth rates. We chose a Multiple Linear Regression model. Why? Because our exploratory analysis, particularly the correlation heat map, revealed a very strong, straightforward, and linear relationship between our most important feature (Fertility Rate) and our target (Population Growth Rate). Linear Regression is simple, highly interpretable, and perfect for testing our initial hypothesis. We used our four features of Fertility Rate, Life Expectancy, GDP Per Capita,

and the Year of the observation, to predict the population growth rate. We included Year not as a predictor of a trend, but as a way to account for large global changes over time that might affect all countries, like advances in medicine or shifts in global economic policy. This setup allowed us to see how much each of the three core demographic/socioeconomic factors truly influenced growth.

# 6 Model Evaluation

To see how well our model performed, we used the $R^2$ (R-squared) score, which tells us what percentage of the variability in population growth rate our features could explain. After training and testing, the results showed a Train $R^2$ of about 0.33 and a slightly better Test $R^2$ of about 0.41. This means our model, even a simple linear one, could successfully explain between 33% and 41% of the variation in population growth rates across all countries and years. While not perfect because population growth is super complex, this is a solid start for a model using only a few variables. The coefficients of the normalized features confirmed our original hypothesis that the Fertility Rate had by far the largest coefficient value, officially making it the single most influential driver in the model. We also visually compared the distribution of our model's predictions against the actual population growth rates, which showed they followed a similar shape, reinforcing that the model captured the general trend.

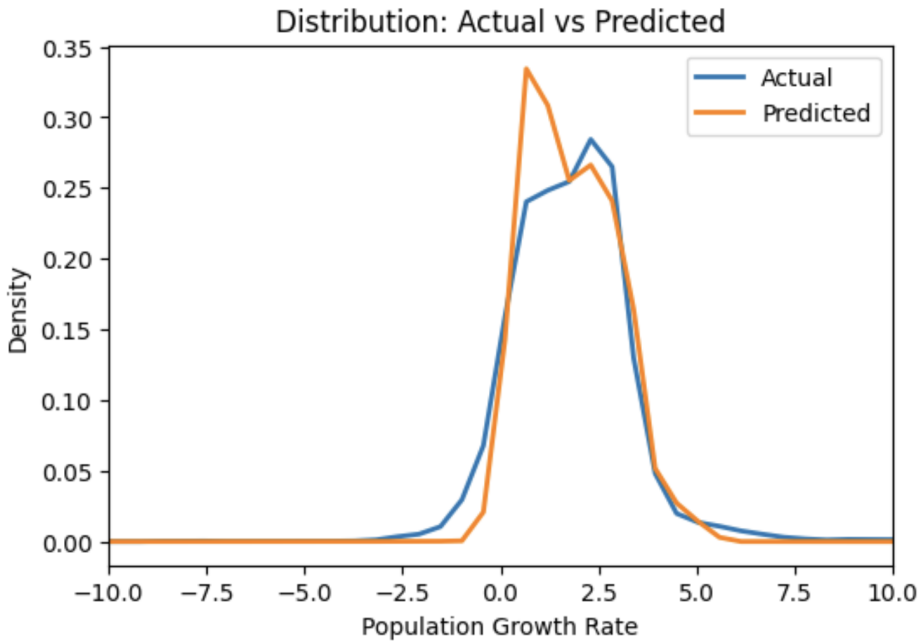


Figure 3: **Actual vs. Predicted:** Comparing model predictions against actual values.

# 7 Model Deployment

We imagine deploying this simple linear regression model as a quick and user friendly analytical tool for anyone involved in development or policy. It could be integrated into an interactive dashboard where policymakers could adjust future estimates for fertility rate, life expectancy, or GDP and instantly see the resulting estimated population growth. This would be fantastic for organizations trying to anticipate demographic shifts, plan where to allocate resources, or compare growth trajectories between different countries. However, we have to address the model's limitations. Since it's based on historical data, its predictions are only as good as the numbers we fed it, and any missing or inconsistent data could throw it off. More importantly, population growth is affected by deep, complex factors like wars, migration, political stability, and climate change that a simple linear model simply can't fully capture. So, while it's a great tool for initial exploration and education, it should be used with a healthy dose of caution for any major, real-world decision-making.

# 8 Project Meeting Log

| **Date** | **Location** | **Attendees** | **Focus** |
|---|---|---|---|
| Oct 15 | Fedex Center | All Members | Proposal Brainstorming |
| Nov 02 | Library | All Members | Data Cleaning |
| Nov 10 | Fedex Center | All Members | Model Selection |
| Nov 17 | Zoom | All Members | General Discussion |
| Dec 01 | Library | All Members | Final Report Compilation |
| Dec 02 | Library | All Members | Poster Creation |
| Dec 04 | Library | All Members | Final Presentation Practice |

Table 1: Log of group meetings and attendance.