Fertility & Infant Mortality Rates

Leslie Cervantes Rivera & Kevin Le

Background

The Infant Mortality and Fertility Rates Dataset provides data on infant mortality rates (measured as the number of infant deaths per 1,000 live births) and fertility rates (measured as the number of births per 1,000 women per year) across U.S. states. The Infant Mortality Dataset covers the years 2003 to 2023, while the Fertility Rates Dataset spans 2016 to 2023.

Infant mortality refers to the death of a baby that occurs before their first birthday. It is a key indicator of a population's overall health, reflecting the social, economic, and healthcare conditions within a state. High infant mortality rates could indicate insufficient healthcare access or inequalities in medical services, particularly in low-income and rural areas. Over the years, improvements in screening and treatment for illnesses, better obstetric management, and neonatal care have contributed in the declining of infant mortality rates, but disparities still exist within state and demographic groups (Marino).

Fertility rate represents the total number of children a woman has during her reproductive years ("Fertility Rate."). It plays a crucial role in population growth and demographical planning, influencing services such as education, healthcare, and workforce development. From 2007 to 2022, the fertility rate has dropped by about 19%, influenced by the health of the economy, social, health, and historical events (Hickerson). These changes have long-term implications, as it leads to an aging population, smaller workforce, and economic strain on government budgets.

The Population Demographics dataset provides data on the years 2016 to 2023, providing information on total population, as well as the distribution of sex and age, marital status, educational attainment, SNAP participation, household median income, and health insurance coverage by ethnicity. These variables represent key social, economic, and healthcare conditions across U.S. states.

Ryabov (2015) found that the relationship between Total Fertility Rate (TFR) and various human development indicators "a negative association between selected human development indicators and TFR exists in suburban and rural counties, as well as in the United States as a whole." However, the analysis was inconclusive for urban counties; some indicators such as

median income were positively associated with TFR, whereas higher educational attainment negatively associated with TFR.

In 2019, Hamilton et al. reported a TFR of 1,705 expected births per 1,000 women for all women aged 15–49 in the U.S. The TFR decreased as the level of education increased, starting from women with a 12th-grade education or less through those with an associate's or bachelor's degree. However, the TFR began to rise again among women with a bachelor's degree and continued to increase for those holding a doctorate or professional degree.

Data Sources

Birth Rates and Infant Mortality Datasets

The datasets come from the Centers for Disease Control and Prevention (CDC), collected from anywhere a person receives healthcare. This may lead to limitations as it is up to the city, county, and state to decide what information is collected, and how and when it can be shared by the CDC ("Where Does Our Data Come From.").

Population Demographics

The dataset is sourced from the United States Census Bureau under the American Community Survey (ACS) category, specifically the *Selected Population Profile in the United States*. The Census Bureau collects data directly from respondents through censuses and surveys, with primary sources also including administrative data.

It should be noted that data for 2020 is not included due to disruptions caused by the COVID-19 pandemic. To protect the safety of their employees and respondents, the Census Bureau suspended ACS data collection operations, including mail surveys, in-person followups with non-respondents, and data collection from group quarters such as nursing homes, college dorms, and prisons.

Question of Interest

How do educational, and regional factors influence birth rates and infant mortality trends across the US from 2016 to 2023? Specifically do certain characteristics such as income, and public services like food stamps and health care have a negative, positive, or no effect?

As politicians continue to attack the Department of Education, propose cuts on SNAP benefits, and limit access to reproductive healthcare, we are wondering if this will affect future population growth.

A satisfactory answer is one in which we identify variables that explain or at least influence fertility and infant mortality data for state and regional levels.

Exploratory Data Analysis (EDA)

Fertility and Infant Mortality Rates

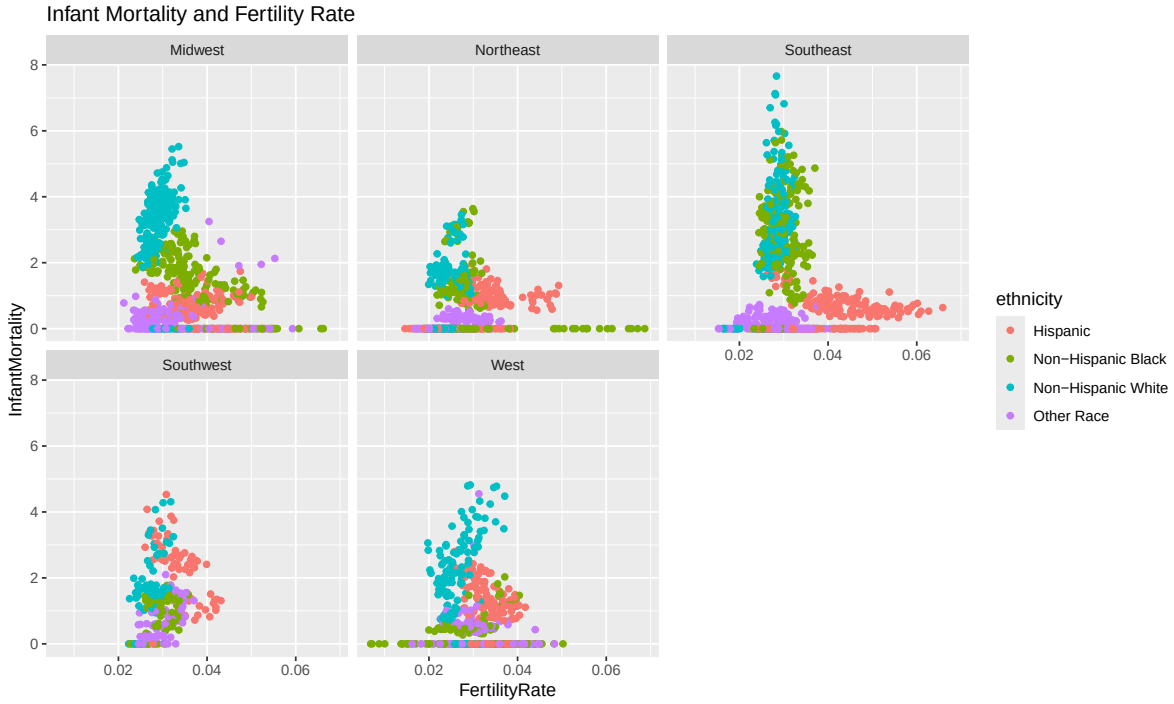
$$\text{fert\_age1524} = \frac{\text{births\_age1524}}{\text{pop\_age1524}} \times 1000$$
$$\text{fert\_age2534} = \frac{\text{births\_age2534}}{\text{pop\_age2534}} \times 1000$$
$$\text{fert\_age3544} = \frac{\text{births\_age3544}}{\text{pop\_age3544}} \times 1000$$
$$\text{total\_fer} = (\text{fert\_age1524} + \text{fert\_age2534} + \text{fert\_age3544}) \times 10$$

To calculate total fertility rate, we first calculated age-specific fertility rates by dividing the number of births in each age group by the corresponding population and multiplying by 1000, following CDC methodology. These rates were then summed across age groups and multiplied by 10, to estimate the average number of children a woman would have over her reproductive lifetime. Finally, we expressed the results as births per woman for easier interpretation.

$$\text{total\_mort} = \frac{\text{deaths\_total}}{\text{norm\_total\_births}}$$

To calculate total infant mortality rate, we divided total number of infant deaths by the total number of births. Since we combined the Fertility and Infant Mortality datasets, we had to normalize due to the datasets having a slight difference in ethnicity-based birth counts.

Initial Exploration of CDC Data



When analyzing birth rates, it's important to account for the relationship between fertility and infant mortality rates. Since higher fertility rates often correlate with higher raw infant mortality rates, we normalize the data by calculating standardized fertility and infant mortality rates to ensure accurate comparisons. To improve visibility, we grouped states by region, because faceting by state was unclear.

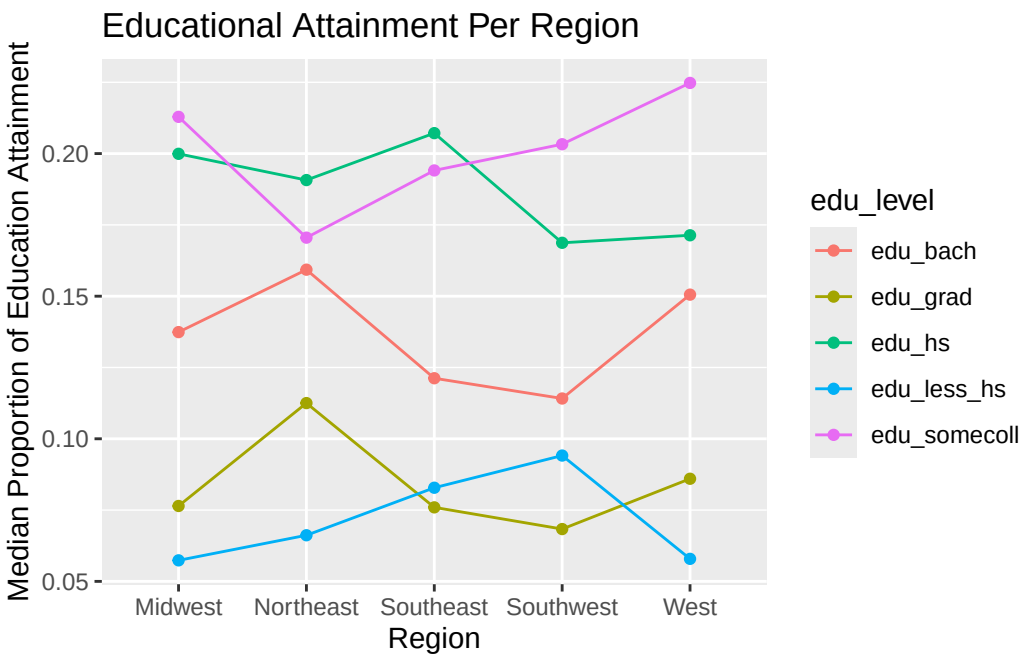
Across all five regions, the "Other race" category consistently shows both lower fertility and infant mortality rates. In contrast, fertility rates for Non-Hispanic White populations exhibit little variation, whereas their infant mortality rates show the highest variation among the four ethnic groups.

For Non-Hispanic Black populations, the relationship between fertility rates and infant mortality generally cluster towards lower fertility rates and lower infant mortality rates. However, there is a noticeable increase in the variation of infant mortality in the Southeast, similar to Non-Hispanic White populations.

Finally, among Hispanic populations, there is a trend of higher fertility rates, especially in the Midwest, Northeast, and Southeast regions. Interestingly, Hispanic groups also tend to have lower infant mortality rates compared to other ethnicities in all regions except for Southwest.

It should be noted that coloring and filtering by year did not yield a noticeable difference in the structure of each ethnicity.

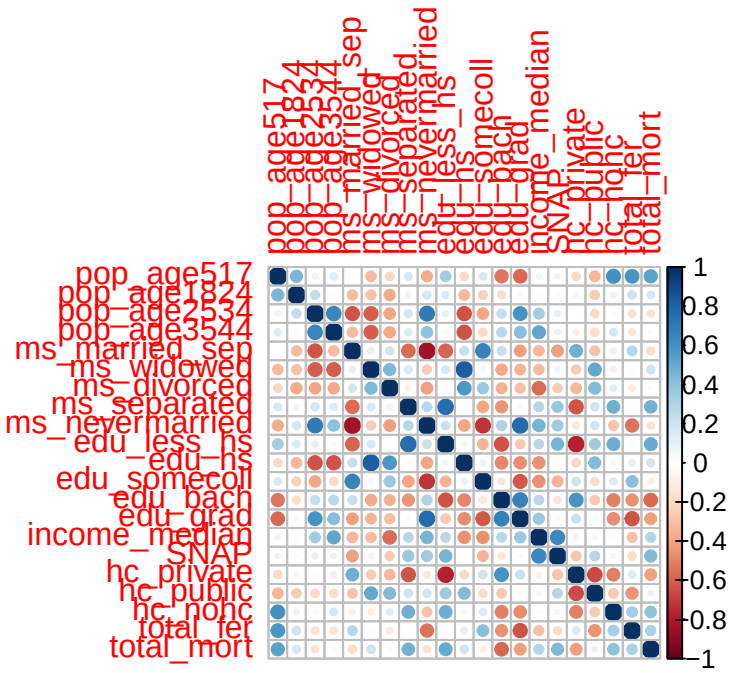
Educational Attainment Per Region



The graph shows that the Southwest region has the highest proportion of individuals with less than a high school education. The Southeast has the highest proportion of individuals whose highest attainment is a high school degree. The West region leads in having the highest proportion of individuals with some college education, while the Northeast has the highest proportion of individuals with a bachelor's and graduate degrees.

Correlation Matrix

The correlation matrix allow us to examine the relationships between `total_fer` and `total_mort` with various socioeconomic and demographic factors.



- `total_fer` and `edu_somecoll`: States with a higher proportion of individuals with some college education tend to have a slightly higher total fertility rate.
- `total_fer` and `ms_nevermarried`: States with a higher proportion of never-married individuals tend to have a moderately lower total fertility rate, suggesting marital status plays a role in fertility patterns.
- `total_fer` and `edu_grad`: States with a higher proportion of individuals with graduate degrees tend to have a lower fertility rate, possibly due to career and individual prioritization.
- `total_mort` and `SNAP`: States with a higher proportion of SNAP consumers tend to have a slightly higher total mortality rate, possibly reflecting low financial stability.
- `total_mort` and `no_hc`: States with a higher proportion of individuals with no health coverage tend to have a higher total mortality rate, suggesting lack of access to healthcare such as prenatal care.
- `total_mort` and `edu_bach`/`edu_grad`: There are moderate negative correlations between total infant mortality rates and higher education levels, suggesting states with a greater population of individuals with a graduate degree or higher tend to experience lower mortality rates, potentially due to better access to resources.
- `total_fer`/`total_mort` and `pop_age517`: In both total fertility and mortality rates, states with a higher proportion of individuals aged between 5 and 17 (particularly those aged 15 to 17) tend to have slightly higher rates.

Given the presence of correlated variables, we will apply Principal Component Analysis (PCA) to reduce dimensionality and identify the most important features that capture the most variance in the data.

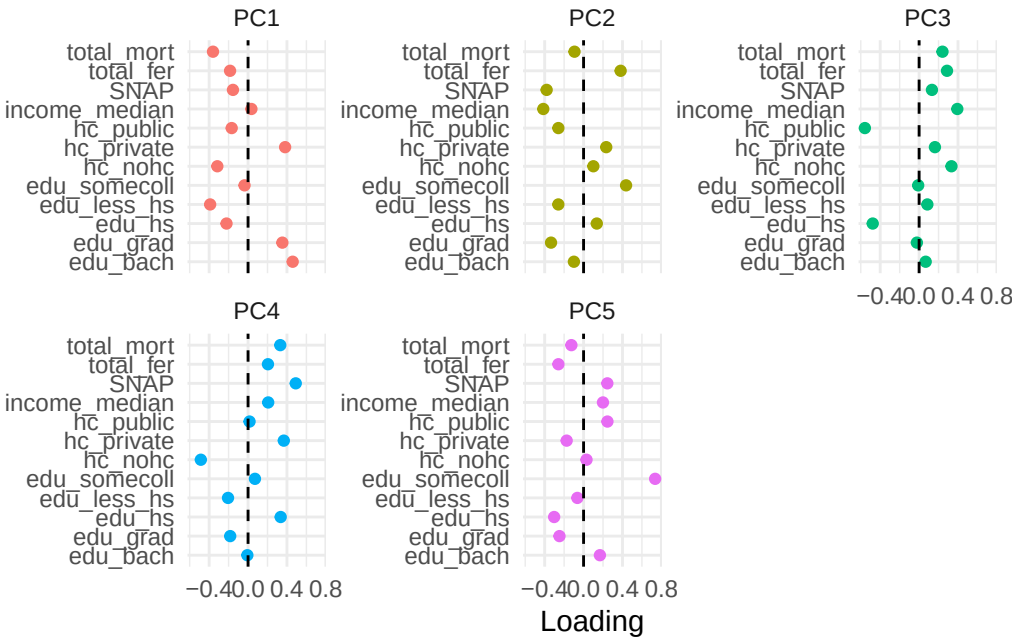
Principal Component Analysis (PCA)

We are going to identify the most significant patterns in the data while minimizing information loss.

To determine the optimal number of components, we plotted the proportion of variance explained and the cumulative variance. From the visuals, we identified that five components capture about 80% of the total variance.

Examining Loadings

Next, we will identify the variables that contribute the most to each principal component by examing their positive and negative weights. This will help us discover patterns.



Name for PC1: *"Education Attainment and Infant Mortality"*

The variables that are most influential are `edu_bach`, `edu_grad`, `ed_less_hs`, `total_mort`, and `hc_private`. States with high PC1 scores indicate a relatively higher than average proportion of residents with private insurance, bachelors and graduate degrees. While states with a lower

PC1 value indicates a relatively higher than average proportion of residents who have less than a high school education and higher infant mortality rate.

Name for PC2: *"Education and Income"*

The variables most influential for PC2 are `edu_somecoll`, `income_median`, `SNAP`, `total_fer`, and `edu_grad`. States with a high PC2 value tend to have a higher than average proportions of residents who have attained only some college education and have higher total fertility rates. However, states with low PC2 values have a higher than average proportion of individuals with graduate degrees, higher median income, and SNAP rate.

Name for PC3: *"Public/None Healthcare Coverage and High School Educated Population"*

The variables that are most influential for PC3 are `hc_public`, `edu_hs`, `income_median`, and `hc_nohc`. Lower PC3 scores are associated with states that have a higher than average proportion of individuals with public healthcare coverage and those with only a high school degree. Conversely, high PC3 scores are linked to states with higher than average median incomes and a proportion of individuals with no health insurance.

Name for PC4: *"Private/None Healthcare Coverage and SNAP"*

The variables that are most influential for PC4 are `hc_nohc`, `SNAP`, `hc_private`, `edu_hs`, and `total_mort`. State with high PC4 scores tend to have higher than average SNAP rates, a greater proportion of individuals with private healthcare coverage, a larger share of individuals with only a high school diploma, and higher infant mortality. Whereas, low PC4 scores indicate states with a relatively higher proportion of citizens without healthcare coverage.

Name for PC5: *"Education Attain and Fertility Rate"*

The variables that are most influential for PC5 are `edu_somecoll`, `edu_hs`, `total_fer`, and `edu_grad`. States with high PC5 scores indicates a higher proportion of residents who only attained some college education. In contrast, states with low PC5 scores have a higher than average proportion of individuals with only a high school degree or graduate degree and higher total fertility rate.

In the scope of our project, we are going to pick the variables that are the most correlated to total fertility and infant mortality rates. These variables are education attainment, economic indicators, and health coverage. These variables align with social, economic, and healthcare conditions.

Overall, it seems that educational attainment, followed by healthcare coverage, explained a lot of the variation in our data.

Total Fertility Rate Analysis (Estimates)

Year	Bachelor	Graduate	Public.Healthcare
2016	-4.093	-3.959	-3.869
2017	-4.566	-3.107	-3.913
2018	-3.589	-4.067	-3.154
2019	-4.203	-3.496	-2.858
2021	-4.034	-3.360	-2.441
2022	-3.595	-3.909	NA
2023	-3.887	-4.029	NA

Total Infant Mortality Rate Analysis (Estimates)

Year	Bachelor	Graduate	Median.Income
2016	-92.74	NA	NA
2017	-85.53	NA	1.879e-05
2018	-63.04	-44.20	2.392e-05
2019	-63.09	NA	NA
2021	-65.66	-40.97	1.416e-05
2022	-65.37	-56.14	1.238e-05
2023	-62.43	NA	1.319e-05

Linear Regression

In this part, we'll use linear regression to see which predictors are significant for each year. We are going to regress total fertility and total infant mortality rates on education attainment, economic indicators, and health coverage.

$$\text{total\_fer} = \beta_0 + \underbrace{\beta_1 \text{less hs}_i + \cdots + \beta_5 \text{grad}_i}_{\text{education attainment}} + \underbrace{\beta_6 \text{income\_median}_i + \beta_7 \text{SNAP}_i}_{\text{economic indicators}} + \underbrace{\beta_8 \text{private}_i + \cdots + \beta_{10} \text{no coverage}_i}_{\text{health coverage}} + \epsilon_i$$

$$\text{total\_mort} = \beta_0 + \underbrace{\beta_1 \text{less hs}_i + \cdots + \beta_5 \text{grad}_i}_{\text{education attainment}} + \underbrace{\beta_6 \text{income\_median}_i + \beta_7 \text{SNAP}_i}_{\text{economic indicators}} + \underbrace{\beta_8 \text{private}_i + \cdots + \beta_{10} \text{no coverage}_i}_{\text{health coverage}} + \epsilon_i$$

TFR Analysis (Estimates)

Highest Educational Attainment Gap

State	Region	TFR	TIMR	EDU Gap	Private H. Ins.	Public H. Ins.	No H. Ins.
WV	SE	1.745984	7.022246	-0.4303145	0.6117712	0.4708296	0.05933861
NV	W	1.855272	5.715390	-0.3767206	0.6240543	0.3341561	0.12823672
AR	SE	1.957765	8.081812	-0.3756445	0.6042494	0.4253394	0.08111635
MS	SE	1.856713	8.797333	-0.3747325	0.5710601	0.3848793	0.12820787
KY	SE	1.943188	6.679062	-0.3647445	0.6373239	0.4219584	0.05341831
LA	SE	1.953974	7.971727	-0.3562219	0.5981105	0.3828598	0.10491624

Note: NA indicates not significant

Running the linear regression for each year in 2016 to 2023 (not including 2020) for total fertility rate **Figure @ref(fertility_table)**, revealed that education attainment, particularly bachelor's and graduate degrees, was consistently significant and negatively associated with fertility rates. This indicates that a higher proportion of individuals with a bachelor's or graduate degree is associated with lower fertility rates. This may suggest adults are delaying traditional markers of adulthood such as securing a job, getting married, and having children. As a result, adults are having children later in life, which means having fewer due to reproductive health considerations. The magnitude of the negative effect for bachelor's seems to vary slightly each year, with the strongest negative effect in 2017 and a slightly weaker effect in 2021. In contrast, graduate degrees had the strongest negative effect in 2023 and a weaker effect in 2017.

Public health insurance also had a significant negative association with fertility rates from 2016 to 2021. States with higher proportions of individuals with public health insurance tend to have lower fertility rates. Public health's strongest negative effect occurred in 2016 and 2017, and dropped significantly in 2022, indicating it might be more muted over the years.

For total infant mortality rate **Table @ref(mortality_table)**, a similar pattern was observed, with bachelor's showing a consistently negative significance from 2016 to 2023. Graduate degree had a significant negative effect between 2018 to 2022. The strongest negative effect for bachelor's education was in 2016, and gradually decreases, reaching its lowest effect in 2023. However for graduate's education, the strongest negative effect occurred in 2022.

Income median was also statistically significant for total infant mortality, but is approximately 0, making no impact on infant mortality rates.

In conclusion educational attainment is significant to both total fertility and infant mortality rate, with health insurance also playing a role in fertility rates.

Lowest Educational Attainment Gap

State	Region	TFR	TIMR	EDU Gap	Private H. Ins.	Public H. Ins.	No H. Ins.
CT	NE	1.599438	4.551920	-0.1753801	0.6840320	0.3559672	0.06032768
VA	SE	1.801822	5.816454	-0.1703695	0.7385265	0.2654202	0.09254426
CO	W	1.723395	4.795118	-0.1693896	0.6623546	0.3323717	0.08537134
MD	NE	1.805444	6.575433	-0.1550811	0.7243955	0.3135246	0.06742054
MA	NE	1.530933	3.919695	-0.1123101	0.7150150	0.3721184	0.02597246
DC	SE	1.230983	0.000000	0.2140750	0.7148357	0.3497646	0.02692039

Education Gap

As observed in our initial exploration, states in the Southeast region **Table @ref(top_gaps)**, which have a lower proportion of highly educated residents, tend to exhibit higher than average total mortality rates per 1,000 live births. In contrast, states in the Northeast **Table @ref(bottom_gaps)**, which have a higher proportion of highly educated residents (despite the educational gap being negative), tend to show lower than average total mortality rates. While educational attainment was found to be a statistically significant predictor of total fertility rate, its effect remains inconclusive when analyzed at the region and state levels.

Conclusion

Fertility and infant mortality rates play a huge role in population growth and are key indicators of a country's overall health, reflecting the social, economic, and healthcare conditions. As we move toward the future, we question how major conflicts will impact our population. In this project, we explored fertility and infant mortality rates across U.S. states using various methods.

We identified trends across ethnic groups and found that Non-Hispanic White populations had the highest variation in infant mortality. The South region had a higher proportion of individuals with a high school degree or less, the West led in some college education, and the Northeast region had the highest proportion of bachelor's and graduate degree holders. In addition, we analyzed the relationships between total fertility (TFR) and infant mortality rates (IMR) with predictors such as age groups, marital status, economic indicators, and health coverage. PCA emphasized education attainment and healthcare coverage as the most significant predictors, which we tested using linear regression for both TFR and IMR. Using linear regression, it revealed the change of estimates of significant predictors over time. Given the importance of educational attainment, we examined the education gaps between higher education (graduate and above) and lower education (bachelors and below). We also used a table to examine trends in healthcare coverage categories.

Throughout our analysis, we encountered challenges with missing or omitted data, particularly for ethnic groups other than Non-Hispanic White. This resulted in occurrences where our techniques failed to identify statistical significance or produced missing values (NAs). Additionally, attempts to run state-level linear regression per year produced NA values due to having more predictors than observations.

For future research, we would like to obtain county-level data to better assess maternal and infant health disparities in the United States. It would also be very interesting to see if voting patterns are correlated with educational gaps for each state and county levels.

Sources

"Fertility Rate." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., www.britannica.com/topic/fertility-rate. (2025).

Hickerson, Ali. "The US Fertility Rate Is Decreasing: What It Means for the Nation's Future | News | Citynewsokc.Com." *United States Fertility Rate Is Dropping*, www.citynewsokc.com/news/the-us-fertility-rate-is-decreasing-what-it-means-for-the-nations-future/collection_81b09a28-e02f-59b9-b52a-08b15c647dcf.html. (2024).

Marino, Kate. "Infant Mortality Rates Declining, but Sudden Unexpected Infant Death Is on the Rise." *Infant Mortality Rates Declining, but Sudden Unexpected Infant Death Is on the Rise | VCU Health*, www.vcuhealth.org/news/infant-mortality-rates-declining-but-sudden-unexpected-infant-death-is-on-the-rise/#:~:text=Wolf%20attributes%20declining%20overall%20infant,media%2
(https://www.nichd.nih.gov/health/topics/infant-mortality/topicinf.) (2025).

"Where Does Our Data Come From." *Centers for Disease Control and Prevention*, archive.cdc.gov/www_cdc_gov/surveillance/data-modernization/basics/where_does_our_data_come_from.h
(2023).

Ryabov, Igor. "On the relationship between development and fertility: The case of the United States." Comparative Population Studies 40.4 (2015).

Hamilton, Brady E. "Total fertility rates, by maternal educational attainment and race and Hispanic origin: United States, 2019." (2021).

Variable Table

Fertility Dataset

Note: `fert_age1524` and below were created.

Variable Name	Meaning
births_age1524	Births to mothers aged 15-24.
births_age2534	Births to mothers aged 25-34.
births_age3544	Births to mothers aged 35-44.
pop_age1524	Population aged 15-24 (women age 15-54 population).
pop_age2534	Population aged 25-34 (women age 15-54 population).
pop_age3544	Population aged 35-44 (women age 15-54 population).
births_total	Total births across all categories.
fert_age1524	Fertility rate for females aged 15-24.
fert_age2534	Fertility rate for females aged 25-34.
fert_age3544	Fertility rate for females aged 35-44.
total_fer	Total Fertility Rate for women aged 15-44.

Infant Mortality Dataset

Variable Name	Meaning
deaths_total	Total births across all categories.
norm_total_births	Total normalized births.
total_mort	Total Infant Mortality Rate.

Census Dataset

Variable Name	Meaning (Proportional Total Population by State and Year)
pop_age517	Population aged 5-17 (women age 5-54 population).
pop_age2534	Population aged 25-34 (women age 5-54 population).
pop_age3544	Population aged 35-44 (women age 5-54 population).
pop_age2534	Population aged 15-24 (women age 5-54 population).
ms_married_sep	Marital status: individuals married but are separated.
ms_widowed	Marital status: widowed individuals.
ms_divorced	Marital status: divorced individuals.
ms_separated	Marital status: separated individuals.
edu_less_hs	Education: individuals with less than high school education.
edu_hs	Education: individuals with high school education.
edu_somecoll	Education: individuals with some college education.
edu_bach	Education: individuals with bachelors education.
edu_grad	Education: individuals with graduate education.
income_median	Household median income.
SNAP	Individuals with SNAP benefits.

Variable Name	Meaning (Proportional Total Population by State and Year)
`hc_private`	Health Coverage: individuals with private insurance.
`hc_public`	Health Coverage: individuals with public insurance.
`hc_nohc`	Health Coverage: individuals with no health insurance.

Number of Components