PSTAT 100 Final Project - Fertility and Infant Mortality 2003 - Present

Arthur Kim and Wesley Kim

March 8th, 2025

Abstract

This project examines fertility and infant mortality trends in the United States using two datasets spanning from 2003 to 2023. The analysis focuses on understanding how socioeconomic and healthcare-related factors - specifically the mother's education, Medicaid coverage, and maternal age - are associated with infant mortality rates over time. Key variables explored include the proportion of births to mothers with college degrees, proportion of births covered by Medicaid, and the age distribution of mothers at childbirth. We analyzed how these demographic factors correlate with overall infant mortality rates (IMR), revealing potential patterns and disparities in maternal and infant health outcomes. This report aims to offer insights that can help inform health policymakers and promote equity in maternal and child health in the US.

Question of Interest

Throughout this report, we will be diving into the question:

What socioeconomic factors are most predictive of high infant mortality rates within the States?

Sub-Questions

- Sub-question 1: What role does education of the mother play in predicting infant mortality rates?
- Sub-question 2: How does medicaid coverage influence overall infant mortality rates across states?
- Sub-question 3: What significance does age of the mother play when compared to infant mortality rates?

Variable Descriptions

- state: Which U.S. state infant statistic was taken
- year: Calender year of the data
- bmcode: Bimonthly period code (1 = Jan-Feb, 2 = Mar-Apr,...)
- bacode: Biannual code (1 = First half of the year, 2 = Second half of the year)
- time: Time period indicator
- start_date: Starting date of the observation period, typically the same as the time variable
- end_date: Ending date of the observation period

Population Statistics

- pop_total: Total population in the area
- pop_nhblack: population of Non-Hispanic Black women age 15-44
- pop_nhwhite: population Non-Hispanic White women age 15-44
- pop_hisp: population of Hispanic women age 15-54 population
- pop_otherraceeth: Other races/ethnicities women age 15-54 population
- births_nhblack: Births to non-Hispanic Black mothers
- births_nhwhite: Births to non-Hispanic White mothers
- births_hisp: Births to Hispanic mothers
- births_otherraceeth: Births to mothers of other races/ethnicities
- births_total: Total births across all categories

By Insurance:

- pop_medicaid: Population covered by Medicaid (women age 15-54 population)
- pop_nonmedicaid: Population not covered by Medicaid (women age 15-54 population)
- births_medicaid: Births covered by Medicaid
- births_nonmedicaid: Births not covered by Medicaid

By Education:

- pop_nohs: Population without high school education (women age 15-54 population)
- pop_hs: Population with high school education (women age 15-54 population)
- pop_somecoll: Population with some college education (women age 15-54 population)
- pop_coll: Population with college degree (women age 15-54 population)
- births_nohs: Births to mothers without high school education
- births_hs: Births to mothers with high school education
- births_somecoll: Births to mothers with some college education
- births_coll: Births to mothers with college degree

By Marital Status:

- pop_married: Married population (women age 15-54 population)
- pop_unmarried: Unmarried population (women age 15-54 population)
- births_married: Births to married mothers
- births_unmarried: Births to unmarried mothers

By Age:

- births_age1524: Births to mother ages 15-24
- births_age2534: Births to mother ages 25-34
- births_age3544: Births to mother ages 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)

Mortality

- deaths_nhblack: Number of deaths among non-Hispanic Black population

- deaths_nhwhite: Number of deaths among non-Hispanic White population
- deaths_hisp: Number of deaths among Hispanic population of any race
- deaths_otherraceeth: Number of deaths among other racial/ethnic groups
- deaths_con: Number of congenital-related deaths (present at birth)
- deaths_noncon: Number of non-congenital deaths
- deaths_neo: Number of neonatal deaths (within first 28 days of life)
- deaths_total: Total number of deaths across all categories

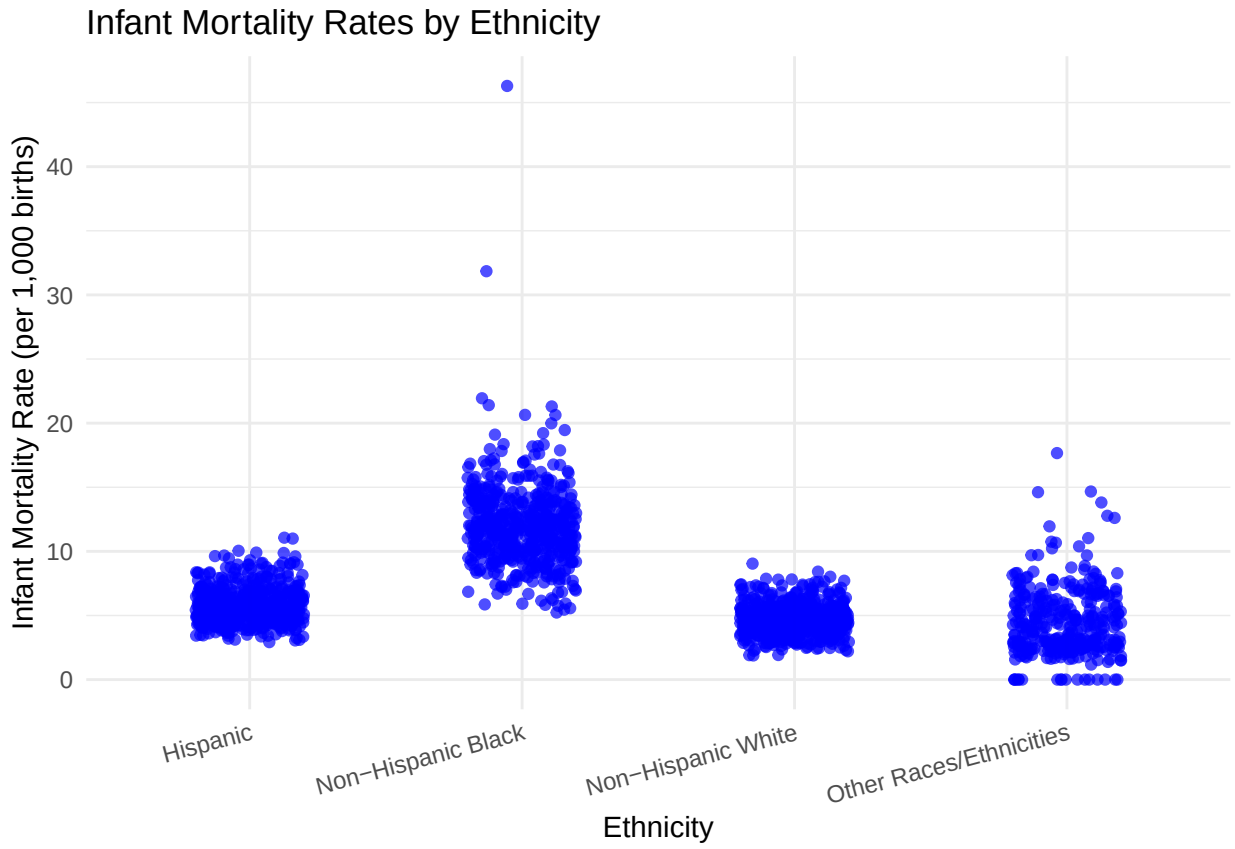
Data Cleaning & Preparation

We began by cleaning and aggregating the data to ensure consistency before merging datasets. We wanted to merge the two datasets, however they were formatted slightly differently. The mortality dataset was labeled by bacode (biannual code), while the fertility dataset used bmcode (bimonthly code). To align these, we recoded bmcode into bacode by mapping bmcode values 1, 2, and 3 to bacode 1 and bmcode values 4, 5, and 6 to bacode 2. We then aggregated the fertility data by summing and averaging relevant statistics to match the biannual structure of the mortality dataset. Finally, we merged the two datasets on state, year, and bacode, handling duplicate columns appropriately to create a clean and structured combined dataset for analysis.

Next, we used mutate to create new columns, including total infant mortality rate (IMR) and the proportion of births from mothers covered by Medicaid. These additions allowed us to quantify key relationships between socioeconomic factors and IMR. By calculating these metrics, we were able to more effectively analyze how maternal healthcare access, education, and other demographic variables influence infant mortality outcomes.

Next, we pivoted the dataset into a long format to facilitate visualization and analysis of infant mortality rates (IMR) across different ethnic groups. The original dataset stored IMR values in separate columns for each racial/ethnic group (e.g., IMR_nhblack, IMR_nhwhite, IMR_hisp). To make this data easier to work with, we used pivot_longer() to convert these multiple IMR columns into two columns: one for Ethnicity and another for IMR values. This restructuring allowed us to analyze IMR trends more efficiently and create clearer visualizations. To improve readability, we also renamed the ethnicity categories from their original variable names to more intuitive labels (e.g., “IMR_nhblack” was changed to “Non-Hispanic Black”). This transformation made it easier to compare IMR across racial and ethnic groups while preparing the dataset for further statistical analysis and visualization. With this long-format dataset, we were able to create our first plot, illustrating the distribution of IMR across different demographic groups.

Infant Mortality Rates by Ethnicity




```
pairwise_results <- pairwise.t.test(IMR_long$IMR, IMR_long$Ethnicity, p.adjust.method = "bonferroni")

pairwise_results
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  IMR_long$IMR and IMR_long$Ethnicity
##
##                          Hispanic Non-Hispanic Black Non-Hispanic White
## Non-Hispanic Black       < 2e-16  -                  -
## Non-Hispanic White       4.2e-16  < 2e-16            -
## Other Races/Ethnicities < 2e-16  < 2e-16            0.59
##
## P value adjustment method: bonferroni
```

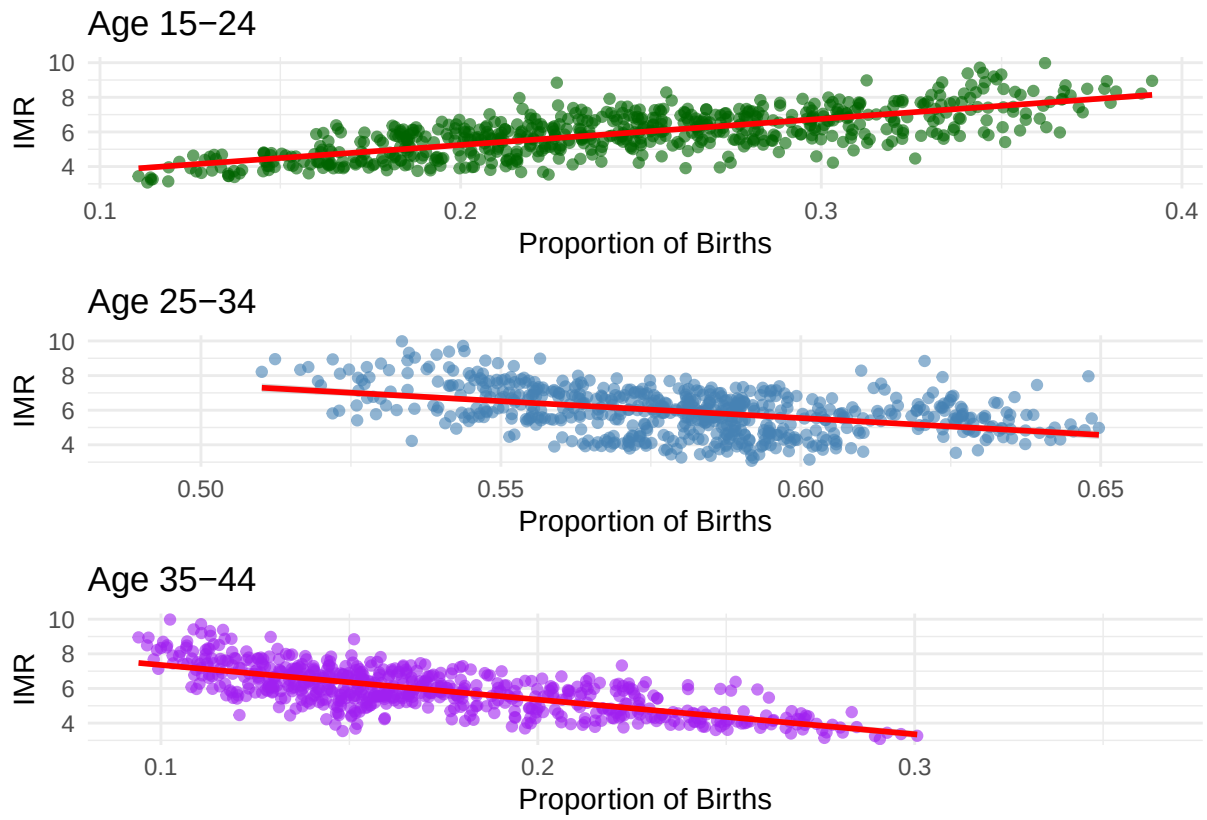
When we first look at the data and compare within different ethnic groups we see that generally Non-Hispanic Black mothers have a higher infant mortality rate compared to the other ethnic groups. The other ethnicities have a generally lower IMF.

The pairwise t test shows that there is a statistically significant difference between all of the ethnic groups except Non-Hispanic White and Other Races/Ethnicities.

Ethnic Group	Mean IMR	Standard Deviation IMR
Hispanic	5.80	1.41
Non-Hispanic Black	11.97	3.24
Non-Hispanic White	4.72	1.16
Other Races/Ethnicity	4.47	2.65

When looking through the documentation, it states that the mean IMR for the states is around 5.6. Based on that we see that the Hispanic population is only slightly over, the Non-Hispanic White and other races are a good amount below. However we can see that in the numbers, Non-Hispanic Blacks have 11.97 which is significantly higher than the national average but has the highest standard deviation implying that their are extreme ends. This does not mean just because you have a Non-Hispanic black mother, the infant has a higher chance of mortality. This might show that non-Hispanic Black mothers are being effected by other factors such as education, healthcare access, etc that could be making this number higher. Now we will look into these possible factors.

Proportion of Births by Different Maternal Age Groups vs IMR




```
##                   Age.Group Correlation..r. Regression.Slope
## pct_births_age1524     15-24           0.733           15.087
## pct_births_age2534     25-34          -0.428          -19.495
## pct_births_age3544     35-44          -0.719          -20.005
```

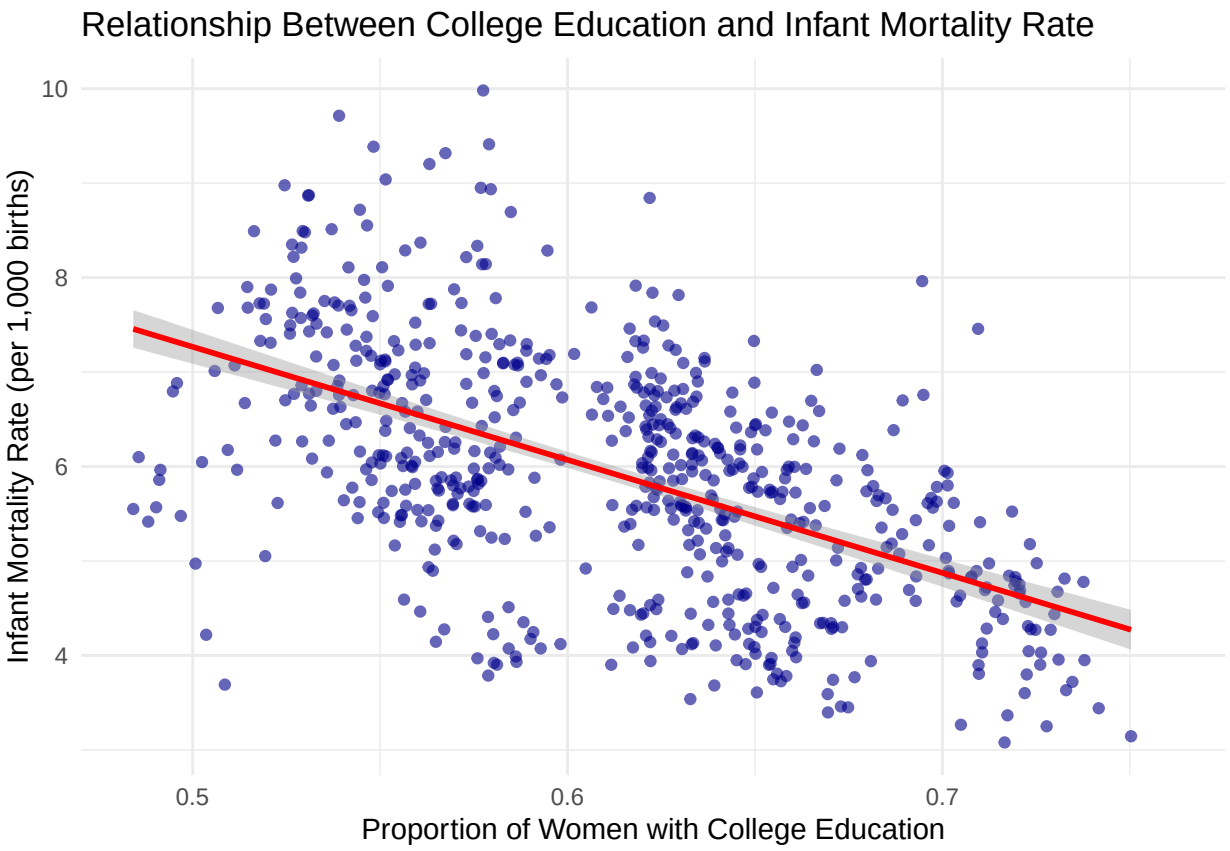
From the graph above we can see that for mothers that are on the younger side, 15-24, has a positive correlation with IMR (Infant Mortality Rate). This suggests that states with higher proportions of births

to younger mothers tend to experience higher infant mortality rates. While mothers that are aged 25-34 and 35-44 both have a negative correlation to IMR, indicating that a higher proportion of births in these older maternal age groups is associated with lower infant mortality rates. This now leads us to the question, what are the factors between older and younger woman that affects the infant mortality rate? This can be attributed to the 15-24 age group having struggles accessing healthcare services, possibly a lower socioeconomic status, higher rates of unintended pregnancies, and potential lack of education pertaining to an infants health and practices.

Age Group	Correlation R	Regression Slope
15-24	0.733	15.087
25-34	-0.428	-19.495
35-44	-0.719	-20.005

We calculate two variables, correlation R which measures the strength and direction of the linear relationship between the variables and the regression slope which tells us the rate of change in the response variable for every unit change in our predictor variable (proportion of births in an age group). As can see from the table above, we see that the age group 15-24 has a positive correlation of 0.733. While the age group 25-34 has a negative correlation of -0.428 and the age group for 35-44 also has a negative correlation of -0.719. The regression slope for 15-24 has the value of 15.087 while for age groups 25-34 and 35-44 they have a negative regression slope of -19.495 and -20.005. This helps confirm our beliefs of how each age group behaves when compared with IMF.

Proportion of Births by College-Educated Mothers vs IMR



Another major factor that we investigated was education among woman has a significant correlation with the infant mortality rate. As shown in the graph above, we see that there is a negative correlation between IMF and proportion of women with college education. This graph is showing us that as the proportion of women with some or completed college education increases, the IMF decreases. This is not meaning that woman who did not go to college are causing higher rates, it could be more that those woman who are in positions to be receiving a college education tend to have lower IMR.

Predictor	Coefficient	Standard Error	P-Value
(Intercept)	13.2503	0.4413	$<$ 2e-16
prop_college	-11.9663	0.7167	$<$ 2e-16

The regression table above provides evidence supporting the negative relationship between the proportion of college-educated women and the infant mortality rate. The intercept is 13.2503 represents the predicted IMR (per 1,000 births) in a hypothetical situation where the proportion of women with college education is 0. This intercept serves as the baseline for our model. The predictor, prop_college, has a coefficient of -11.9663, indicating that for every one unit increase in the proportion of college-educated women, the IMR is expected to decrease by about 12 deaths per 1,000 births. The standard error of 0.7167 for this coefficient measures the variability of the estimate. The p-value being so low shows us that this relationship is significant.

Proportion of Births with Medicaid Birth Coverage vs IMR

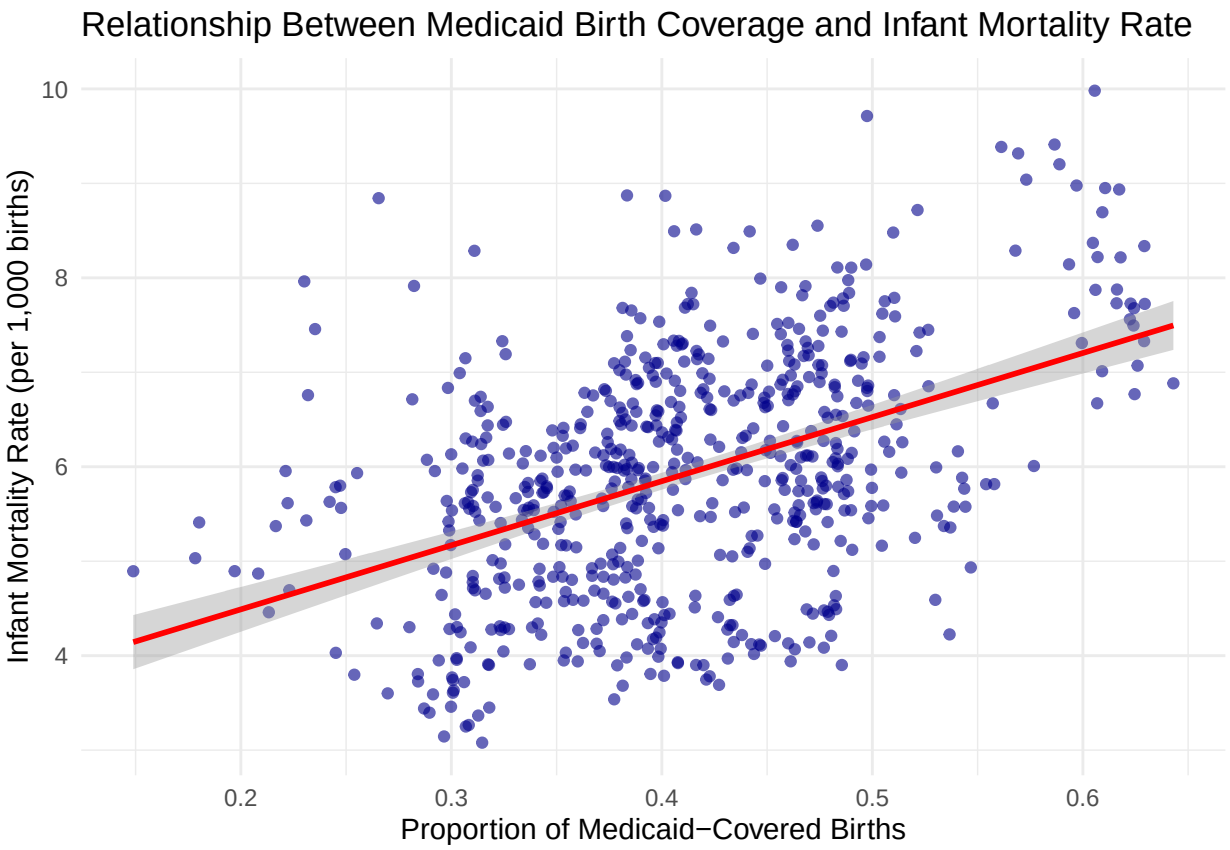


Table 4: Regression Model Coefficients

	Predictor	Coefficient	P.Value
(Intercept)	(Intercept)	3.133	2.48e-39
prop_medicaid	prop_medicaid	6.785	2.25e-33

The graph illustrates a strong positive correlation between the proportion of Medicaid-covered births and infant mortality rates (IMR) across states. The linear regression model indicates that for each 1% increase in Medicaid-covered births, the IMR increases by 0.06785 per 1,000 births. This relationship is highly statistically significant, with a p-value of 2.25e-33, confirming that Medicaid coverage is a key predictor of IMR.

At first glance, one might expect higher Medicaid coverage to be associated with lower infant mortality rates, given that Medicaid provides access to healthcare for low-income individuals. However, it is important to recognize that Medicaid primarily serves economically disadvantaged populations, who may already face higher health risks due to limited prenatal care access, higher rates of preterm births, and other socioeconomic stressors. Thus, rather than Medicaid causing higher IMR, the correlation likely reflects underlying disparities in maternal and infant health outcomes among lower-income groups.

Summary of Findings and Discussion

This study explored the relationship between key socioeconomic and healthcare-related factors and infant mortality rates (IMR) across U.S. states between 2003 and 2023. Using two large datasets covering fertility and infant mortality statistics, we investigated how ethnicity, maternal age, educational attainment, and Medicaid coverage are associated with variations in IMR. Our goal was to identify demographic and structural predictors of infant mortality in order to better understand disparities in maternal and child health outcomes.

One of the most significant findings is the strong association between maternal education and infant mortality. Our analysis demonstrated a clear negative correlation between the proportion of women with some or completed college education and IMR. Specifically, the linear regression model revealed a slope of -11.97, indicating that for every percent increase of college-educated women, the IMR decreased by 0.1197 deaths per 1,000 births. The statistical significance of this result, reflected in a p-value of less than 2e-16, provides compelling evidence that education plays a crucial role in improving infant health outcomes. However, it is important to emphasize that this relationship likely reflects underlying differences in access to healthcare, health literacy, and socioeconomic status. Women with higher education levels may have better access to prenatal care, healthier living conditions, and stronger social support systems, all of which can contribute to reduced infant mortality.

We also analyzed the role of maternal age distribution in relation to IMR. Our findings show that births to younger mothers (ages 15-24) are associated with higher IMR, while births to mothers aged 25-34 and 35-44 are associated with lower IMR. The correlation for the 15-24 age group was positive ($r = 0.733$), with a regression slope of 15.087, suggesting that as the proportion of births to younger mothers increases, the IMR also rises significantly. In contrast, the 25-34 and 35-44 age groups exhibited negative correlations ($r = -0.428$ and -0.719, respectively), with corresponding negative regression slopes. These results are consistent with existing literature that identifies young maternal age as a risk factor for adverse infant health outcomes. Younger mothers may have less financial stability, limited education, and reduced access to healthcare resources, which can increase the risk of complications during pregnancy and infancy. Older mothers, particularly those between 25 and 34 years old, are generally at an optimal age for childbirth, often benefitting from greater financial and social stability as well as more comprehensive access to healthcare services.

Another major finding of this study concerns Medicaid coverage. Our regression analysis identified a statistically significant positive relationship between the proportion of births covered by Medicaid and IMR. The

model estimated that each percent increase in Medicaid-covered births corresponds to an increase of approximately 0.06785 deaths per 1,000 live births in IMR, with a highly significant p-value of 2.25e-33. At first glance, this relationship may appear counterintuitive, as Medicaid is intended to provide healthcare access to economically disadvantaged groups, potentially improving outcomes. However, it is critical to interpret this finding in context. Medicaid coverage is often a proxy for lower socioeconomic status, which correlates with higher health risks, including limited access to quality prenatal care, higher rates of chronic conditions, and increased exposure to environmental and social stressors. Thus, the observed relationship may reflect systemic inequalities rather than the direct impact of Medicaid itself. Medicaid provides vital support to populations who may otherwise lack healthcare access, but structural barriers and disparities continue to drive differences in infant mortality outcomes among Medicaid recipients.

In summary, our findings suggest that higher maternal education, optimal maternal age (25-34 years), and addressing socioeconomic disparities are critical factors in reducing infant mortality rates. The positive association between Medicaid coverage and IMR underscores the need for comprehensive policy solutions that address the broader social determinants of health, rather than relying solely on healthcare coverage as a solution. Policymakers should prioritize expanding educational opportunities for women, improving access to high-quality prenatal and postnatal care, and addressing socioeconomic inequalities to reduce infant mortality rates across all demographic groups.

While this analysis provides valuable insights, it is not without limitations. Our study relies on aggregated state-level data, which may obscure important within-state and community-level disparities. Additionally, the datasets lacked certain individual-level information, such as detailed income data, health conditions, and access to specific healthcare services, which could further illuminate the underlying causes of disparities in IMR. Future research should incorporate more granular data and explore additional factors, including healthcare quality and social support networks, to build a more comprehensive understanding of the determinants of infant mortality.