PSTAT 100 Final Project

Medicaid and Infant Mortality in the United States

Bella Genolio and Marcus Ortiz

Introduction

Healthcare in the United States has a complicated and divisive history. Unlike most other industrialized nations, Americans must navigate a complex system dominated by private markets but heavily subsidized by public programs. Critics of American health insurance claim either economic deficiency or adverse health outcomes unique to the U.S., but whether or not our healthcare policies are actually linked to negative impacts is a complicated question with no immediately apparent answer. One particular category of health with important consequences is childbirth. Perhaps the most emotionally charged area of medicine, inequitable access to proper birthing conditions can have devastating effects on both the physical and mental health of mothers and children across the country. In an attempt to fill the gaps left by private health insurance, Medicaid was signed into law in the 1960's to provide insurance to low-income Americans. This wildly popular program has helped countless Americans gain access to healthcare that they may have otherwise been forced to go without. Despite this fact, many lawmakers continue to push for significant cuts or even outright abolition. Critical to such a discussion is an understanding of the practical impacts of Medicaid on everyday citizens.

Dataset Description

For this project we are combining two separate datasets from the CDC called the fertility dataset and the biannual infant mortality dataset. Upon initial inspection, the fertility dataset has 2,499 observations, each recorded every two months from 2016 to 2024 (with only a handful of observations in 2024) for all 50 states and Washington DC. There are 41 columns in this initial dataset, though we will not be using all of them. These columns fall under a couple of different categories - education, age, marital status, insurance status, race/ethnicity, and population statistics for each of the categories. The infant mortality dataset has a very similar layout to the fertility dataset, for each observation counts of births and deaths per race (as well as congenital, non-congenital, and neonatal deaths) are recorded. These observations happen twice a year rather than every two months and were recorded from 2003 to 2023, though they also include all 50 states and Washington DC.

Primary Question

In our analysis, we will attempt to answer the question: *is there evidence that Medicaid coverage lowers infant mortality rates?* This may seem like a straightforward question, but there are many possible confounding factors. We must take into account certain factors like race, marriage status, and location. It is possible that Medicaid coverage improves birth outcomes only for certain populations, which implies that changes must be made to the program itself. We hypothesize that medicaid reduces infant mortality rates and that any increases in mortality rate are more associated with race/ethnicity than with medicaid coverage.

Data Analysis

Exploratory Data Analysis

Before proceeding with rigorous analysis, some simple observations and plots will be useful in understanding basic relationships present in our data set. First, we can point out that the scatterplot of medicaid coverage rate against infant mortality rate appears to produce a linear relationship such that higher rates of medicaid coverage are associated with higher rates of infant mortality (see Fig 1). When fitting a simple linear regression on the data, only taking into account medicaid coverage rates, the coefficient is highly statistically significant (see appendix). From a public policy perspective, this is problematic since it suggests that medicaid leads to higher rates of infant mortality. Of course, it is much too early to make such a strong conclusion and our rigorous analysis will yield a more accurate answer to this question. This also serves as motivation for further analysis, since infant mortality does not seem to naturally follow from financial aid to poor birthing mothers.

Checking the correlation between the significant variables in our dataset confirms our suspicion that there is something else going on with the data. Medicaid covered births are negatively correlated with the rate of non-hispanic white births. Furthermore, it is positively associated with being black and hispanic, and has a very weak positive correlation with the 'other' category for race and ethnicity. Previous studies have already shown that infant mortality rates are particularly high in communities of color (Jang & Lee, 2022). Infant mortality may have less to do with medicaid coverage and more to do with race and/or ethnicity. The correlation plot also supports this positive relationship that we would expect between mortality rate and minorities. High infant mortality rates also seem to be associated with younger mothers and lower levels of education, which are also related with high rates of medicaid coverage.

The amount of births in each age bracket appears to be relatively stable over time. There is no evidence to suggest that medicaid coverage for births has changed significantly over time either (see appendix). The same appears to be true of infant mortality according to race (see

Fig 1. Simple Linear Regression of Infant Mortality Accordin

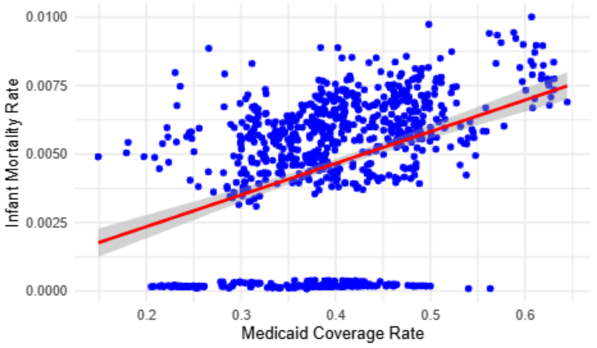


Fig 2. Correlation Plot

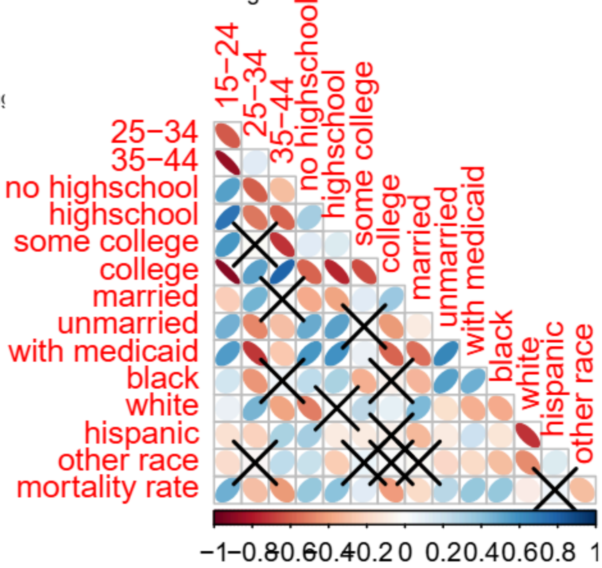


Fig 3). This suggests that the proper approach for further analysis does not include time series data analysis. Interestingly, there seems to be the highest count of infant mortality for non-hispanic white children. While this may be true when taking a simple count, it seems reasonable when considering the fact that non-hispanic white people constitute a majority in the United States. This supports our transformation of the data to proportions where each 'births' column represents the birth rate rather than a count. Therefore, our upcoming analysis will be on proportions rather than raw counts to make sure our outcomes aren't biased. Additionally, there is no clear relationship between region and medicaid covered births. While there is some variation throughout the country, it is relatively weak and does not suggest any significant disparities (see Fig 4).

Fig 3. Deaths by Race Over Time & Medicaid Births

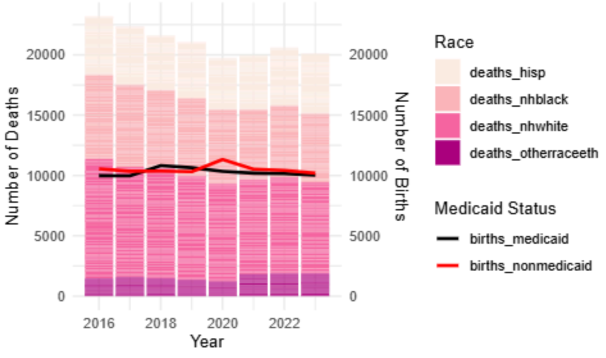
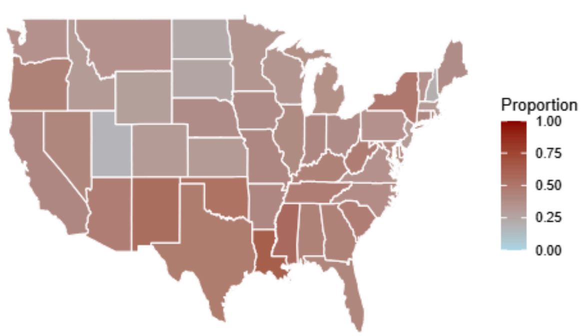


Fig 4. Heatmap of Medicaid−Covered Births Across U.S. States, 202



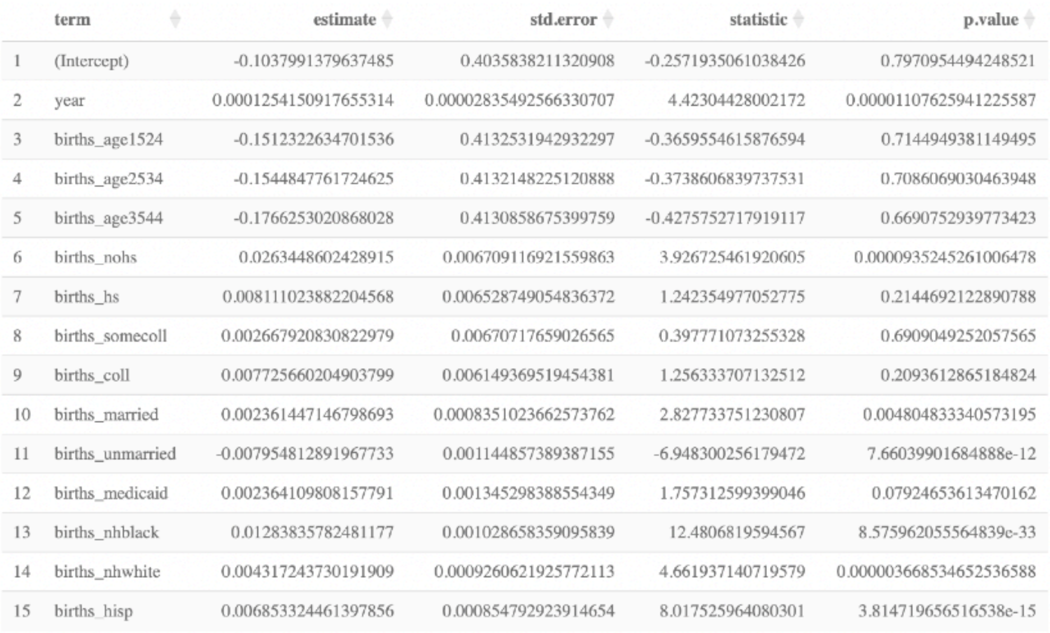
Linear Regression

While our EDA seems to suggest a positive correlation between medicaid coverage and infant mortality rates, a more complex model might reveal the real nature of the relationship when accounting for other possible confounders. In order to explore this relationship, we employed a multiple linear regression. All of the input variables are numeric and continuous on the scale (0,1) so we don't need to re-level any factors. The explanatory variables included in our multiple linear regression are the birth rates by different age brackets, race/ethnicity, marriage status, education status, and (most importantly) medicaid coverage. This means that our linear model is defined as follows:

$$\begin{aligned}\mathrm{IMR}_i = \beta_0 + \beta_1\mathbf{year}_i + \underbrace{\beta_2\mathbf{(15\text{-}24)}_i + \cdots + \beta_4\mathbf{(35\text{-}44)}_i}_{\text{age cohort}} + \underbrace{\beta_5\mathbf{no\_hs}_i + \cdots + \beta_8\mathbf{college}_i}_{\text{education}} + \\ \underbrace{\beta_9\mathbf{black}_i + \cdots + \beta_{11}\mathbf{other}_i}_{\text{ethnicity}} + \underbrace{\beta_{12}\mathbf{married}_i + \beta_{13}\mathbf{single}_i}_{\text{marriage}} + \beta_{14}\mathbf{medicaid}_i + \epsilon_i\end{aligned}$$

Recall that each of these variables represents a rate in each biannual time period in a given state. Each beta coefficient represents the marginal difference in that rate for every increase of 1 in the rate variable. Because the variables are all proportions, we have to scale down the coefficients in our interpretation. Dividing by 100 gives us a better idea of how much change occurs with change in a given rate.

Table 1.



	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.1037991379637485	0.4035838211320908	-0.2571935061038426	0.7970954494248521
2	year	0.0001254150917655314	0.00002835492566330707	4.42304428002172	0.00001107625941225587
3	births_age1524	-0.1512322634701536	0.4132531942932297	-0.3659554615876594	0.7144949381149495
4	births_age2534	-0.1544847761724625	0.4132148225120888	-0.3738606839737531	0.7086069030463948
5	births_age3544	-0.1766253020868028	0.4130858675399759	-0.4275752717919117	0.6690752939773423
6	births_nohs	0.0263448602428915	0.006709116921559863	3.926725461920605	0.0000935245261006478
7	births_hs	0.008111023882204568	0.006528749054836372	1.242354977052775	0.2144692122890788
8	births_somecoll	0.002667920830822979	0.00670717659026565	0.397771073255328	0.6909049252057565
9	births_coll	0.007725660204903799	0.006149369519454381	1.256333707132512	0.2093612865184824
10	births_married	0.002361447146798693	0.0008351023662573762	2.827733751230807	0.004804833340573195
11	births_unmarried	-0.007954812891967733	0.001144857389387155	-6.948300256179472	7.66039901684888e-12
12	births_medicaid	0.002364109808157791	0.001345298388554349	1.757312599399046	0.07924653613470162
13	births_nhblack	0.01283835782481177	0.001028658359095839	12.4806819594567	8.575962055564839e-33
14	births_nhwhite	0.004317243730191909	0.0009260621925772113	4.661937140719579	0.000003668534652536588
15	births_hisp	0.006853324461397856	0.000854792923914654	8.017525964080301	3.814719656516538e-15

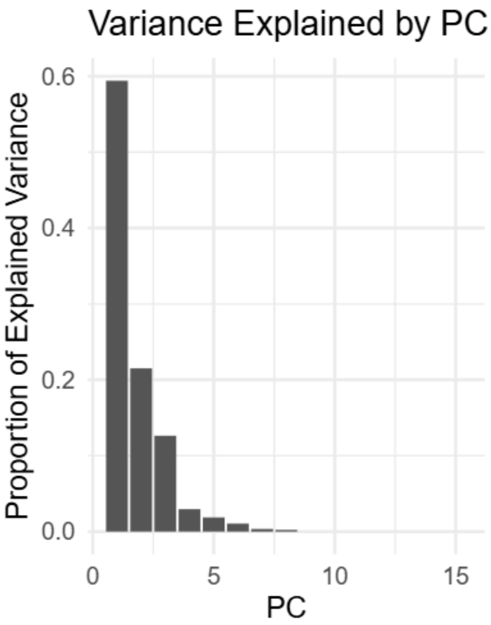
Since this is a multiple linear regression, our model helps account for confounders. We can get a sense of the complex relationships between certain birth rate categories and the infant mortality rate using the numbers in table 1. Most of the predictors are not statistically significant or lie within >2 standard errors from 0. Notably, the p-value for the medicaid coefficient is >0.05 and therefore insignificant. While in our previous simple linear regression of mortality according to medicaid had a highly significant positive linear relationship, when accounting for things like race and education, this relationship is not statistically significant. However, certain coefficients like for the rate of being hispanic, black, and having no high school education are significantly associated with increases in the infant mortality rate. Taking just one example, for every 0.1 increase in the rate of non-hispanic black births, the infant mortality rate increases by ~0.1%. As we expected, the infant mortality rate has more to do with race (and, seemingly, education) than it does with medicaid. Our multiple linear regression suggests that there is no relationship between medicaid coverage and infant mortality.

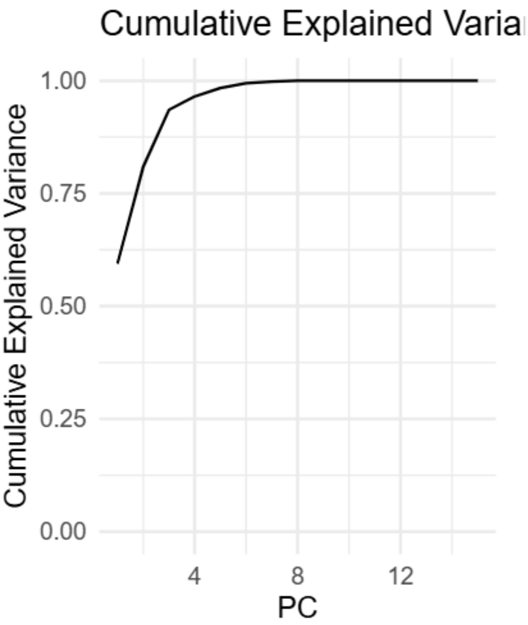
Principal Component Analysis

Principal component analysis is a useful tool for data analysis as it can be used in many different applications. For this dataset, the desired outcome is to reduce dimensionality while retaining important variation in the data. Prior to beginning principal component analysis, we examined the correlations between the medicaid variable and the other variables to see what had the greatest effect on the medicaid births. From the correlation plot (see Fig 2. above), we can see that the births where mothers had medicaid had a positive correlation with mothers aged 15-24, mothers with no schooling, highschool, and some college, unmarried mothers, black mothers, and hispanic mothers. This makes sense since medicaid is a government resource provided mainly to those with lower income and fewer resources, and these demographics represented typically have a higher percentage of people in this category than the other groups in the dataset. The medicaid births variable is negatively correlated with mothers aged 25-34 and 35-44, those who have college degrees, those who are married, white mothers and mothers of race and ethnicity different from black, white, or hispanic.

We then created a dataset that computed all of the principal components with their loadings and scores, centered and scaled. Since our initial data frame included 15 principal components, we needed to reduce this number by examining how much of the total variance could be explained by the least amount of variables possible.

Fig 5.





Looking at the graph tells us that over 75% of the cumulative variance is explained by two of the principal components and the variance is fully explained by the time eight principal components are included. To get an exact percentage of the variance explained by each of the principal components, we made a data frame that reflected this: the first principal component explained 60%, the second 23.4%, the third 8%, and so on. Since the first two principal components explained almost 84% of the variance, we decided to just focus on these two.

In order to visualize the first two principal components, we created a histogram and a scatter plot that display the loadings for each of the variables in the dataset (see Bar plots for PC1 and PC2 and Loadings plot in appendix). For the first principal component we can see that the largest loadings come from medicaid (negative) and non medicaid (positive). Some other positive loadings include white mothers, unmarried mothers, married mothers, white infant deaths, and mothers aged 25-34 while some negative loadings include hispanic mothers, black mothers, black infant deaths, and mothers aged 15-24. It seems as though the first principal component could represent mothers without medicaid, especially based on the correlation plot we saw above. It is very interesting that this accounts for the variance of 60% of the variables in the data frame as it tells us that this principal component does a decent job of providing explanations for many of the variables. However, looking at the loadings and bar plots in the appendix, we can see that our hypothesis that medicaid does not have the strongest connection to infant mortality. We can see that only the black infant deaths have a mid sized negative loading quantity and the rest of the infant deaths have a mid sized positive loading quantity. These middle of the road values do not demonstrate the correlation between medicaid and infant deaths we were expecting.

Looking at the plot for the second principal component tells us that its largest loadings are white mothers (positive), mothers aged 15 to 24 (positive), hispanic mothers (negative), and hispanic infant deaths (negative). We can pretty quickly see that this principal component could represent young white mothers, though let's take a look at what the other large loadings are. Married and unmarried mothers are equally positive, black mothers and black infant deaths are both slightly positive, and age 35 to 44 is very negative. Interestingly, medicaid and non medicaid are both very small, this could be because the younger, unmarried population of white mothers typically have less resources and therefore balance out the portion of the more well off white population of mothers that do not need medicaid.

To further visualize how these principal components impact the dataset, we created a projection of the first two principal components and put them on two scatterplots, one colored by year (see Scatterplot of Principal Components Colored by Year in appendix) and the other by state (see Scatterplot of Principal Components Colored by State in appendix). The years were pretty similar, most of them showed up in all of the primary clusters and were evenly dispersed. The states were less so, California had the most negative values for both principal components and was firmly in the bottom left corner, Hawaii slightly positive for PC 1 but very negative for the second, New York was clustered slightly below the main cluster but still negative for both, and Maine and Mississippi were clustered together in the top left with negative first principal component and positive second. The rest of the states clustered in the center of the graph with some variation in every direction.

Our principal component analysis effectively reduced the dimensionality of the dataset while preserving key variations. The first two principal components explained nearly 84% of the total variance, with the first principal component primarily capturing differences related to Medicaid and non-Medicaid births, and the second principal component reflecting variations among racial and age groups of mothers. The projection of these components showed clear clustering of certain states, suggesting regional differences in birth and death patterns. These insights highlight the strong relationship between socioeconomic factors and maternal health outcomes, offering a valuable perspective for further policy analysis.

Summary of Findings/Discussion

Throughout this project, the main question we were attempting to answer was: does medicaid coverage lower infant mortality rate across America? Given the critical role that medicaid plays in providing healthcare access to low-income families, we assumed the link between these two variables would be very clear from a couple of graphs and data tables. However, our analysis revealed a more complex relationship.

In order to visualize this relationship, we explored the data using seven different graphs to map the relation between medicaid, infant deaths, and the other variables included in the dataset. This initial exploratory data analysis suggested a positive correlation between medicaid coverage and infant mortality. Though this exploration gave us the impression that these factors were very highly correlated, we soon found that this correlation was not as strong as we previously hypothesized. A simple linear regression model showed a statistically significant positive relationship, implying that states with higher medicaid coverage also had higher infant mortality rates. However, this initial finding did not account for key confounding factors, such as race, education, and marital status.

Once these factors were included in a multiple linear regression model, medicaid coverage was no longer a significant predictor of infant mortality. Instead, the results indicated that racial and socioeconomic disparities played a much larger role in determining infant mortality rates. Specifically, higher proportions of black and hispanic births were significantly associated with increased infant mortality, as was lower maternal education. These findings align with prior research indicating that systemic inequalities—rather than medicaid coverage alone—are primary drivers of disparities in infant mortality rates. The principal component analysis further emphasised the lack of relation between infant mortality and medicaid. The first and most influential principal component, that describes 60% of the total variance, did not yield strong loadings for any of the infant deaths.

It is clear that the relationship we were anticipating from this project was disproven throughout our data analysis. This lack of a direct relationship between Medicaid coverage and infant mortality suggests that healthcare access alone does not fully account for disparities in birth outcomes.

Although we disproved our initial primary question, this analysis provided us with a stronger understanding of the dataset and the role that medicaid plays in specific demographic areas of America's mothers and infants. Though medicaid has some effect on infant mortality in America, other factors including racial disparities, quality of healthcare facilities, education access, and family planning facilities have a greater effect on infant mortality nationwide.

From this conclusion, we can see that policies targeting racial and educational inequities may be more effective in reducing infant mortality than medicaid expansion alone. All of this is not to say that medicaid should be cut or reduced, but if policy makers are specifically looking to decrease infant deaths, an increase in resources for lower income, non-white, uneducated mothers could be more beneficial.

A central limitation that we had was the lack of individual observations. Because the observations were state-wide counts, we did not have data on maternal health conditions, prenatal care access, and out of hospital births. We also were unable to see social and environmental factors that could impact infant mortality rates. Housing conditions, food security, and general safety of the mother before birth are important factors that contribute to higher infant mortality rates that were not included in the dataset. Another limitation of this analysis was that we relied on observational data, making it difficult to infer causation rather than correlation.

Adding a method of inferential statistics, bootstrapping for example, could increase the findings of the study by fitting a model in order to use the relationship between variables to predict rates of infant mortality. Additionally, analyzing medicaid policy differences as well as healthcare conditions and racial disparities from state to state would enable us to determine the importance of medicaid expansion policy from state to state. Examining state level medicaid policy difference would provide a view into the difference in infant mortality based on policy rather than population of mothers who receive medicaid.

Ultimately, our findings suggest that while medicaid plays a crucial role in providing healthcare access, addressing infant mortality requires a broader focus on racial and socioeconomic disparities. Future research should explore the intersection of medicaid policy, healthcare quality, and systemic inequalities to develop targeted interventions. By prioritizing comprehensive maternal and infant health initiatives, policymakers can more effectively reduce infant mortality and improve birth outcomes nationwide.

References

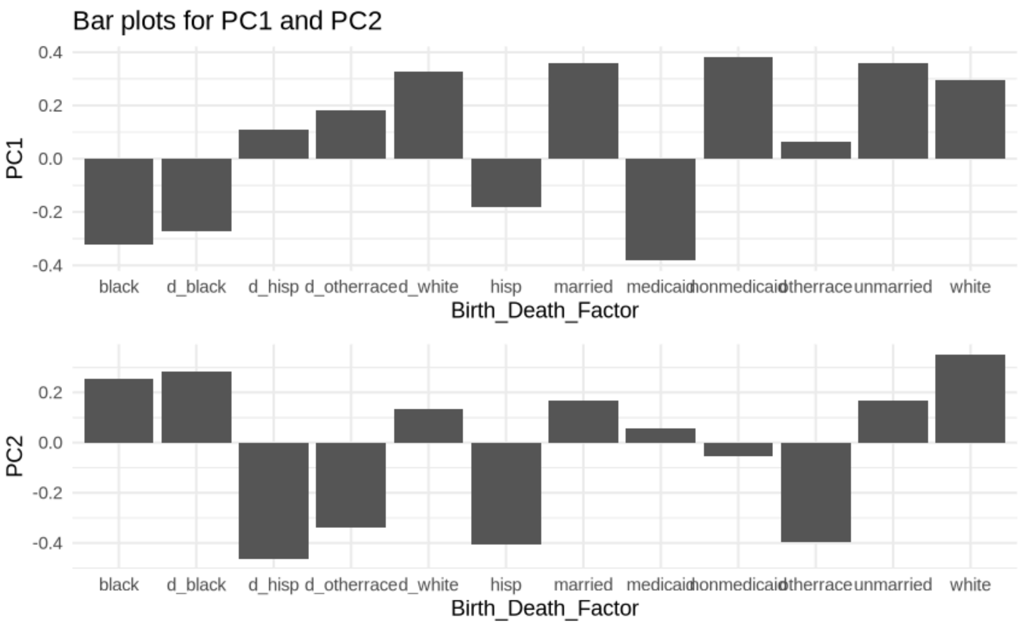
Jang, C. J., & Lee, H. C. (2022). A Review of Racial Disparities in Infant Mortality in the US. *Children*, *9*(2), 257. https://doi.org/10.3390/children9020257

Appendix

```
# Appendix plots

# Simple linear regression and associated coefficiencts/p-value
coef_tbl_slm <- tidy(deaths_simple_lm)
print(coef_tbl_slm)
```

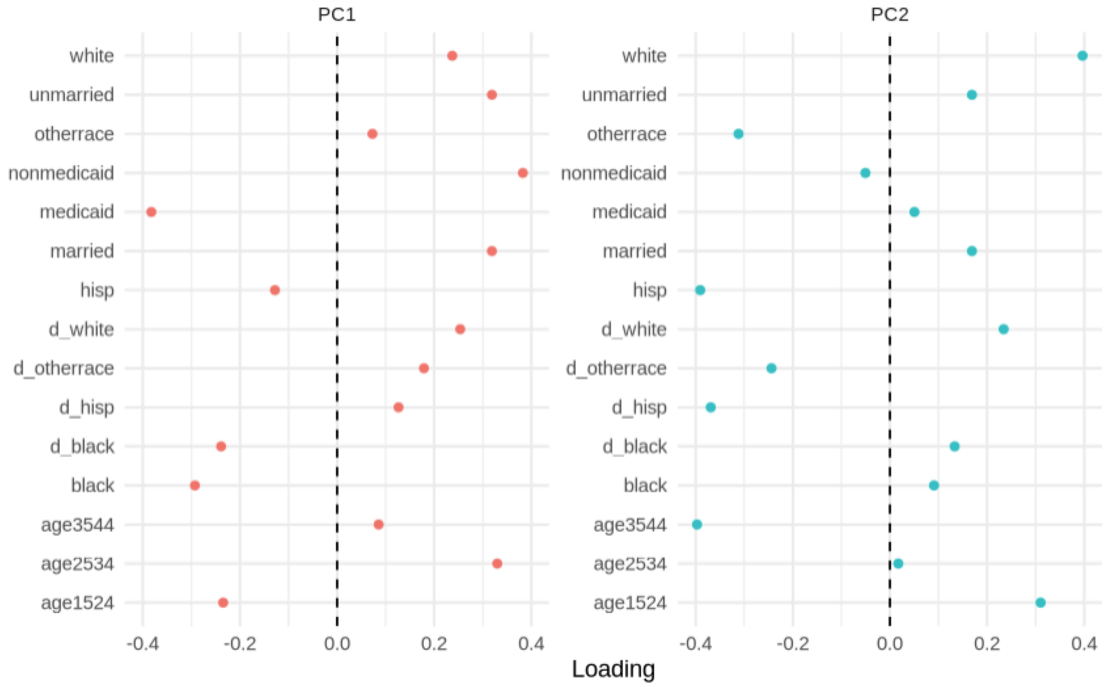
```
# A tibble: 2 x 5
  term            estimate std.error statistic   p.value
  <chr>              <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)        -7.76     0.227    -34.2  9.12e-160
2 births_medicaid     4.67     0.554      8.43 1.56e- 16
```

```
# Additional PCA Plots

library(patchwork)
# Get loadings and scores
loadings <- data.frame(data_pca$rotation[,c(1,2)])
loadings <- loadings%>%
  mutate(Birth_Death_Factor = row.names(loadings))

pc1_loadings <- ggplot(loadings, aes(x = Birth_Death_Factor, y = PC1)) +
  geom_col() +
  ggtitle('Bar plots for PC1 and PC2') +
  theme_minimal()

pc2_loadings <-  ggplot(loadings, aes(x = Birth_Death_Factor, y = PC2)) +
  geom_col() +
  theme_minimal()

pc1_loadings / pc2_loadings
```

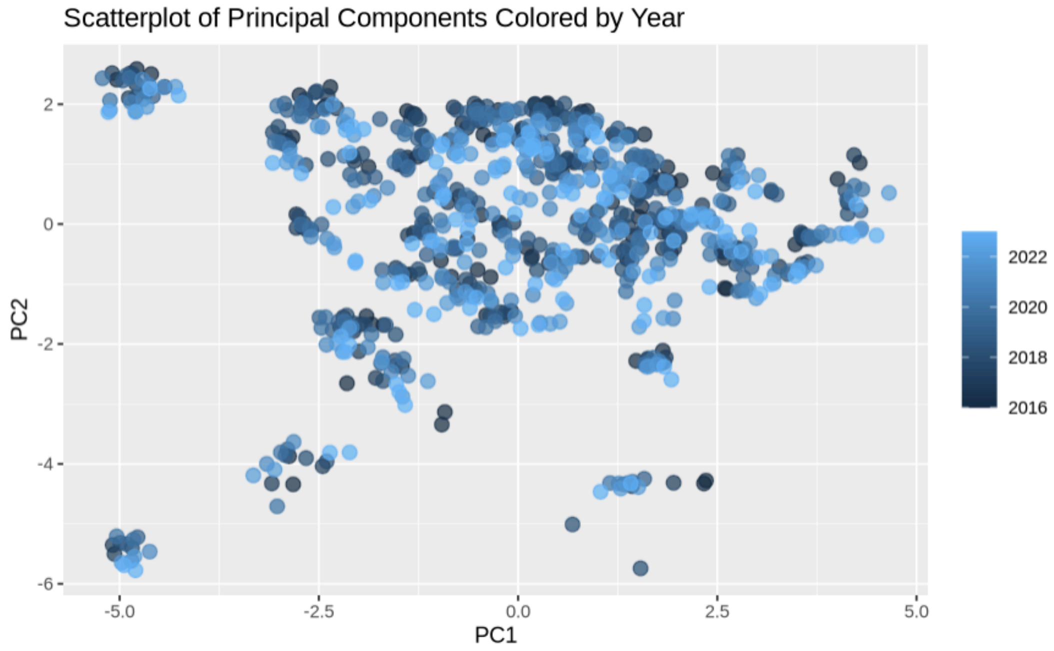


```
loadings_plot_df <- loadings %>%
  select(-Birth_Death_Factor) %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "Variable") %>%
  pivot_longer(cols = -Variable, names_to = "Principal_Component", values_to = "Loading")%>%
  mutate(zero = 0)

loadings_plot <- ggplot(loadings_plot_df, aes(x = Loading, y = Variable, color = Principal_C

  # Add lines + points for loadings
  geom_line(aes(group = Variable)) +
  geom_point() +

  # Add vertical reference line at x = 0
  geom_vline(aes(xintercept = zero), color = "black", linetype = "dashed", size = 0.5) +

  # Facet by Principal Component
  facet_wrap(~ Principal_Component, scales = "free_y") +

  # Adjust labels and theme
  labs(x = "Loading", y = "", color = "Principal Component") +
  theme_minimal() +
  theme(legend.position = "none") # Hide legend since facets already distinguish components
```

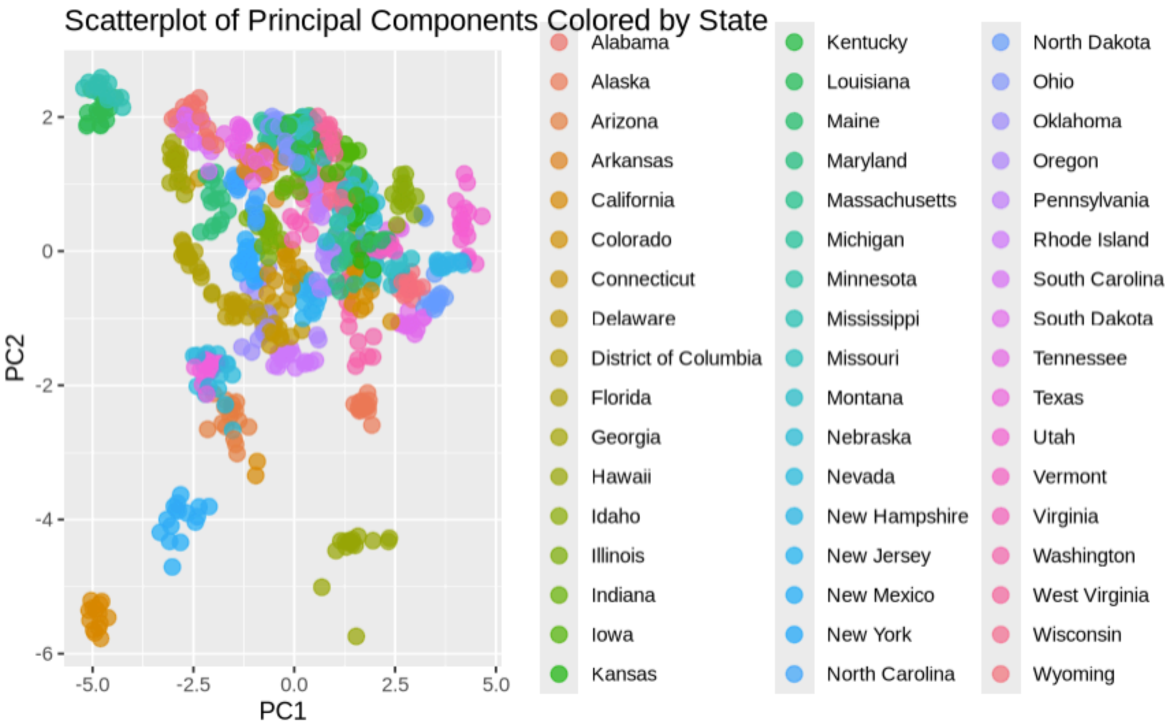
```
# Display the plot
print(loadings_plot)
```



```
# PCs by year
projected_data <- data.frame(data_pca$x[,c(1,2)]) |>
  mutate(year = data_proportions$year)

pc_scores_plot <- projected_data %>%
  ggplot(aes(x = PC1, y = PC2, color = year)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "Scatterplot of Principal Components Colored by Year",
       x = "PC1",
       y = "PC2") +
  theme(legend.title = element_blank())

pc_scores_plot
```



```
projected_data_2 <- data.frame(data_pca$x[,c(1,2)]) |>
  mutate(state = data_proportions$state)

# PCs by state
pc_scores_plot_2 <- projected_data_2 %>%
  ggplot(aes(x = PC1, y = PC2, color = state)) +
  geom_point(size = 3, alpha = 1) +
  labs(title = "Scatterplot of Principal Components Colored by State",
       x = "PC1",
       y = "PC2") +
  theme(legend.title = element_blank())

pc_scores_plot_2
```


Scatterplot of Principal Components Colored by State
PC2
PC1
2
0
-2
-4
-6
-5.0
-2.5
0.0
2.5
5.0
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming