Factors that Predict Infant Mortality in the United States

Introduction

Births provide hope for the future. They augment populations and allow cultures to continue thriving. Infant mortality rate (IMR), or the number of infant (under 1-year-old) deaths per 1,000 live births, provides insight into the overall health of a population by indicating an area's economic development, quality of and access to medical care, socioeconomic disparity and inequality, living conditions, and more. The United States has an alarmingly high IMR compared to other high-income countries, with significant disparities among marginalized populations. For this project, we chose to investigate which socioeconomic and demographic factors are important for predicting which populations in the U.S. are the most vulnerable to infant mortality.

Our research questions are: Do states with higher educational attainment, median household incomes, and number of insured citizens correlate with lower IMRs? Do political environments and/or policies (e.g., a state's abortion legislation, political affiliation, Census region) affect IMR? Do states with higher proportions of racial minorities have higher IMRs?

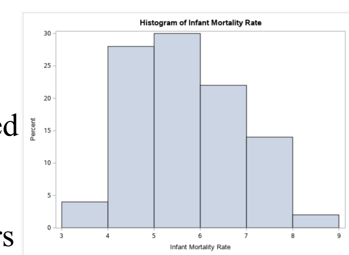
Methods/Analysis

Data Summary

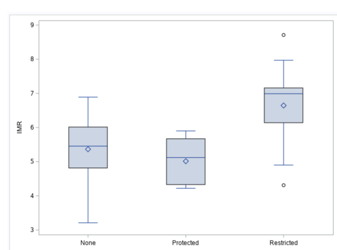
We combined data from various credible sources (e.g., the Centers for Disease Control and Prevention, U.S. Census Bureau) into one data set. Specifically, we collected data from 2019, used states as observations, and implemented IMR as our response variable. The quantitative explanatory variables we used were percentage of the population over 25 years old with at least a Bachelor's Degree (PctBachelors), the median household income (MedianIncome), the percentage of the civilian noninstitutionalized population without health insurance coverage (PctUninsured), the percentage of the population that is white (PctWhite), and the percentage of the population that is not white (PctNonWhite). Our qualitative explanatory variables were the governor's political affiliation (GovernorParty, with two levels: Republican and Democrat), the Census region (Region, with four levels: Midwest, Northeast, South, and West), and the change abortion access (AbortionChange, with three levels: none, protected, and restricted).

Exploratory Data Analysis

To begin determining if our explanatory variables were relevant to our response variable, we performed an exploratory data analysis. Our response variable, IMR, is unimodal and not concerningly skewed. Therefore, we are working with an approximately Normal distribution and can appropriately perform tests and draw conclusions with it. We generated correlation coefficients, scatter plots, box plots, and line plots in SAS for all of our explanatory variables. PctWhite and PctUninsured showed weak relationships with IMR. On the other hand, PctBachelors and MedianIncome produced correlation coefficients of approximately -0.66407 and -0.65376, respectively, indicating moderately negative relationships with IMR. Thus, we predict that the infant mortality decreases as an area's percentage of people over the age of 25 with at least a Bachelor's degree increases or an area's median household income increases. Furthermore, each of our qualitative variables (Region, GovernorParty, and AbortionChange) showed meaningful variation. From the box plots, we predict that states in the Northeast,



Histogram of IMR



Box plots for AbortionChange

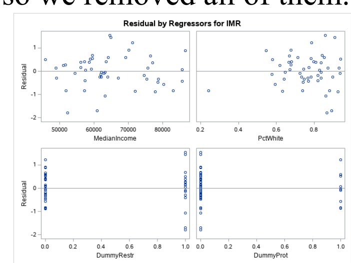
with Democratic governors, or that protect abortion access will report lower IMRs.

Model Building

The first step within the three-stage model building process was to add all of our quantitative variables to our model (PctBachelors, MedianIncome, PctUninsured, and PctWhite). We checked for multicollinearity among our quantitative variables by generating a matrix of scatter plots and the pairwise correlation coefficients. MedianIncome and PctBachelors was the only pair with a high correlation (0.83955), so we performed variable screening. We ran stepwise regression on our quantitative variables with the default values for entry and removal, then discovered that PctUninsured and PctBachelors should be removed, and PctWhite and MedianIncome should stay. We also confirmed these results using a global F-test and t-tests with a significance level of 0.05.

The next step was to add our qualitative variables. We added three dummy variables for Region, two for AbortionChange, and one for GovernorParty to our previous model. We performed a global F-test and t-tests. We removed the variables with p-values above 0.05, which included all of the dummy variables for Region and GovernorParty. For AbortionChange, the dummy variable for protecting abortion access (DummyPro) had a p-value above 0.05, but the dummy variable for restricting abortion access (DummyRes) had a p-value below 0.05. Thus, we performed a nested F-test that resulted in a p-value of 0.0059, determining that, together, both were significant enough to remain in the model.

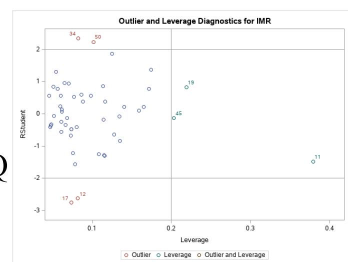
The final step was to add interactions. We added a total of six interactions between MedianIncome, PctWhite, DummyPro, and DummyRes, then ran a global F-test and t-tests with a significance level of 0.05. None of the interactions were significant, so we removed all of them. Our model now includes four variables: MedianIncome, PctWhite, DummyPro and DummyRes.



Fanning in PctWhite's Residual by Regressors Plot

Regression Assumptions

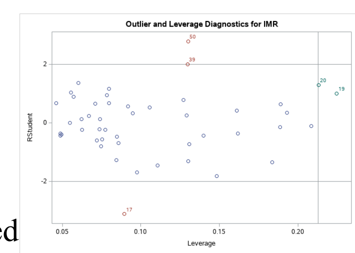
After building our model, we wanted to ensure that our data was adequate for our model. We checked the mean zero, or lack of fit, assumption by looking at the residual by regressors plots. There was slight fanning in PctWhite's, but it was not severe enough to worry about obstructing our model or to violate the mean zero assumption. We checked the equal variance, or homoscedasticity, assumption by looking at the residual by predicted plot. There was no fanning nor a clear imbalance of points, so we concluded our data had constant variance and the equal variance assumption was not violated. We then checked the normality assumption by looking at the histogram and QQ plot. The histogram was slightly skewed, but because our model is robust, it did not prompt concern. Also, the QQ plot had aligned points, proving the normality assumption was not violated. Lastly, we checked the independence assumption by performing the Durbin-Watson test. Because the p-values were above 0.05, we concluded that there was no autocorrelation and the independence assumption was not violated.



Outlier and Leverage Diagnostics Before Removal

Outliers and Influential Points

In SAS, we checked for outliers and influential points. Observations 11 (Hawaii), 12 (Idaho), and 34 (North Dakota) appeared in multiple tests, so we removed them. Although doing so generated



Outlier and Leverage Diagnostics After Removal new potential outliers and influential points, we believe that the combination of observations that we removed maintained the regression assumptions the best.

Weighted Least Squares

To further validate our model, we implemented an additional technique. We tried performing weighted least squares on our data because we considered that IMR may vary with state population sizes. By weighing our data using the total population of each state in 2019 (TotalPop) and using our model, only one aspect of our regression output changed: DummyPro's p-value went below 0.05. However, we did not believe that this change was enough to reconsider our model or regression method, so we were confident to move forward with our previous multiple linear regression (MLR) techniques.

Results

Our final model and prediction equation are as follows:

$$E(IMR) = \beta_0 + \beta_1(MedianIncome) + \beta_2(PctWhite) + \beta_3(DummyPro) + \beta_4(DummyRes)$$

$$E(IMR) = 12.72 - 0.00007(MedianIncome) - 3.78(PctWhite) - 0.39(DummyPro) + 0.69(DummyRes)$$

Based on this MLR analysis of strategically selected explanatory variables, we can conclude that the economic conditions (MedianIncome), racial composition (PctWhite), and abortion legislation (DummyPro and DummyRes) are important predictors for estimating the IMR of a U.S. state.

The REG Procedure
Model: MODEL1
Dependent Variable: IMR IMR

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	44.73330	11.18332	21.43	<.0001
Error	45	23.48805	0.52196		
Corrected Total	49	68.22135			

Root MSE	0.72247	R-Square	0.6557
Dependent Mean	5.73640	Adj R-Sq	0.6251
Coeff Var	12.59441		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.71543	1.12927	11.26	<.0001
MedianIncome	MedianIncome	1	-0.00006574	0.00001154	-5.70	<.0001
PctWhite	PctWhite	1	-3.78350	0.85436	-4.43	<.0001
DummyPro		1	-0.38525	0.29402	-1.31	0.1967
DummyRes		1	0.69343	0.25295	2.74	0.0087

Our Final Regression Output

Conclusions

We recommend that policy makers, public health strategists, and hospital systems refer to our model and prediction equation to appropriately make systematic decisions on public health and social welfare resources, especially in the context of the recent overturning of Roe v. Wade by the Supreme Court. The delegalization of abortion in some states may potentially coincide with an increase in IMR as a result of more babies being born into unsafe, unwanted, or inadequate environments. Therefore, it is imperative that IMR is surveilled and estimated correctly in order to shed light on the potential detrimental population health effects of abortion bans. We advise policy makers and public health strategists to simply substitute in the data for their state and year of interest to our model in order to estimate a prediction of IMR.

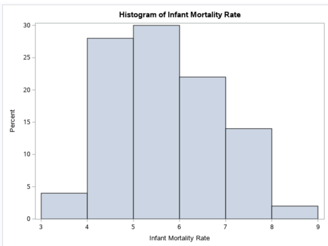
Some limitations to our grasp on the true predictors of IMR include lack of accurate data coverage of less visible, underserved populations, including undocumented immigrants, densely populated urban centers, and expansive rural communities. Because of geographical and sociopolitical circumstances, these populations are often undercounted while simultaneously being at risk for disparate health statuses and conditions as a result of poorer access to and quality of health care, or political barriers. In other words, it is more difficult to collect data on these populations, and they also may be susceptible to higher rates of IMR in the U.S. In future research, we should increase surveillance of these populations, so we can compile the most accurate and representative data and prediction model.

Future research should also collect IMR data with maternal mortality rate (MMR) data simultaneously to determine a strength of association between the two and if there are certain factors that exacerbate one rate more than the other. Furthermore, given the fact that the IMR in the U.S. is 2.5 higher compared to other high-income nations, future research should aim to identify what distinct characteristics of the U.S. healthcare system specifically give rise to this outcome.

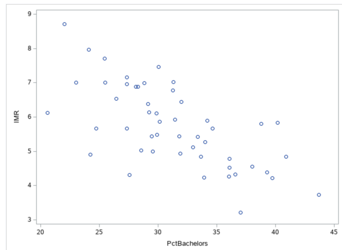
Appendix

<u>Additional Visualizations</u>

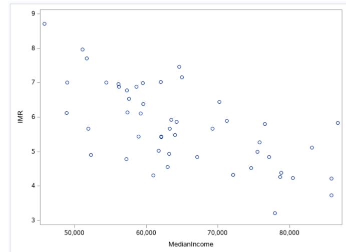
Exploratory Data Analysis



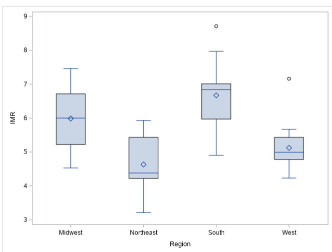
Histogram of IMR



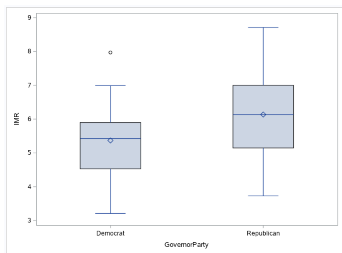
Scatter plot for PctBachelors



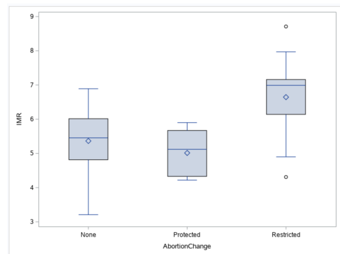
Scatter plot for MedianIncome



Box plots for Region

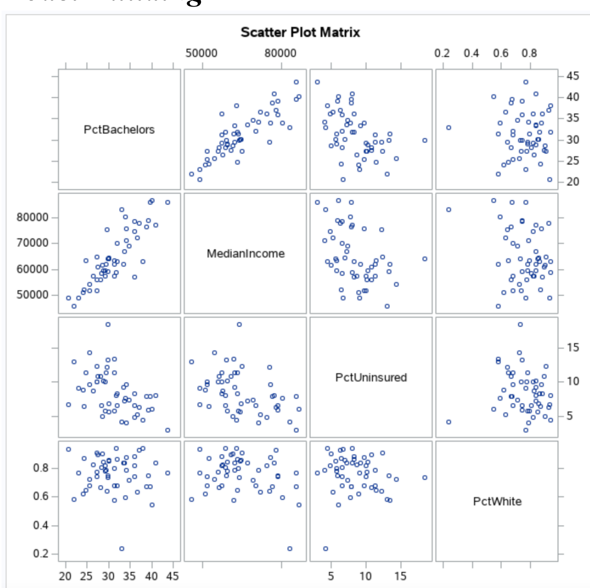


Box plots for GovernorParty



Box plots for AbortionChange

Model Building



Scatter plots for Relationships Between Quantitative Variables

The REG Procedure
Model: MODEL1
Dependent Variable: IMR IMR

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	39.52166	9.88041	15.49	<.0001
Error	45	28.69970	0.63777		
Corrected Total	49	68.22135			

Root MSE	0.79861	R-Square	0.5793
Dependent Mean	5.73640	Adj R-Sq	0.5419
Coeff Var	13.92172		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	13.18952	1.46326	9.01	<.0001	0
PctBachelors	PctBachelors	1	-0.03366	0.04314	-0.78	0.4393	3.96016
MedianIncome	MedianIncome	1	-0.00006554	0.00002210	-2.97	0.0048	4.21893
PctUninsured	PctUninsured	1	0.02873	0.04276	0.67	0.5051	1.32192
PctWhite	PctWhite	1	-3.13937	0.99930	-3.14	0.0030	1.26786

Regression Output for Our Quantitative Variables

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PctBachelors		PctBachelors	1	0.4410	0.4410	13.5342	37.87	<.0001
2	PctWhite		PctWhite	2	0.0457	0.4866	10.6724	4.18	0.0465
3	MedianIncome		MedianIncome	3	0.0885	0.5751	3.2516	9.58	0.0033
4		PctBachelors	PctBachelors	2	0.0076	0.5675	2.0606	0.82	0.3692

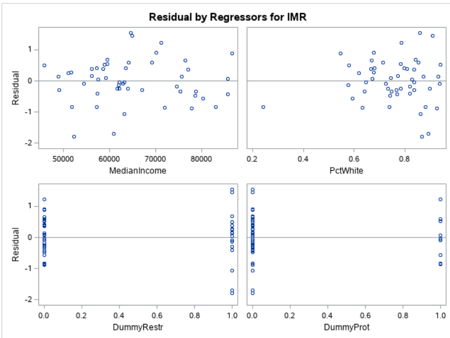
Stepwise Regression on Quantitative Variables

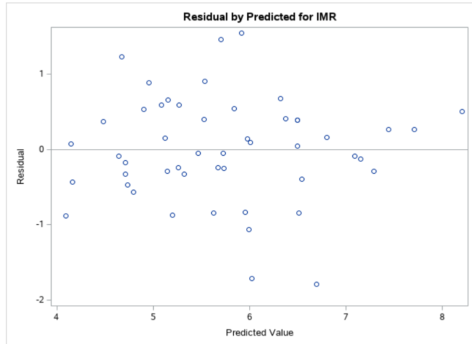
Test 1 Results for Dependent Variable IMR

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	3.00889	5.76	0.0059
Denominator	45	0.52196		

Nested F-test for DummyPro and DummyRes

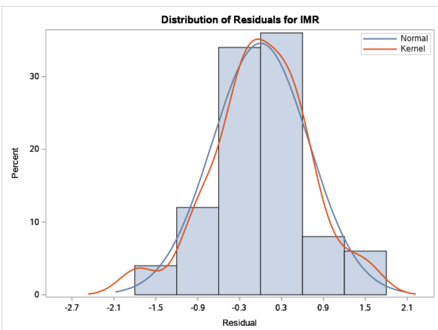
Regression Assumptions

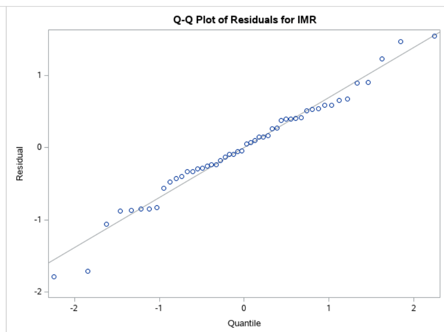




Mean Zero

Equal Variance





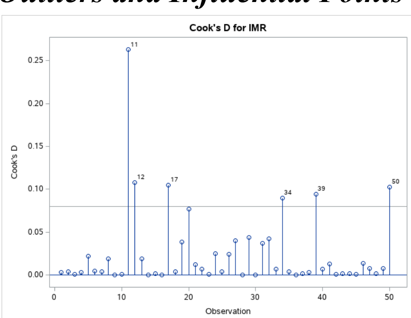
The REG Procedure
Model: MODEL1
Dependent Variable: IMR

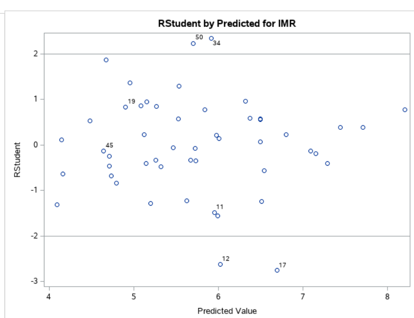
Durbin-Watson D	1.572
Pr < DW	0.0682
Pr > DW	0.9318
Number of Observations	50
1st Order Autocorrelation	0.167

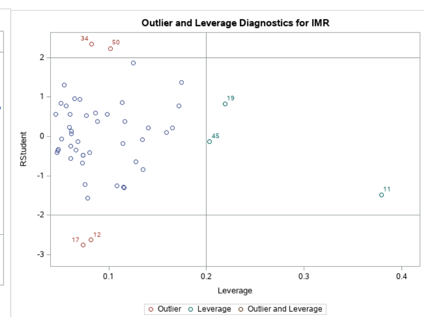
Normality

Independence

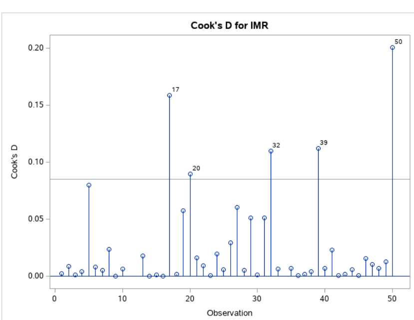
Outliers and Influential Points

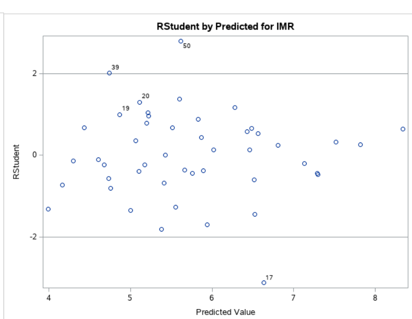


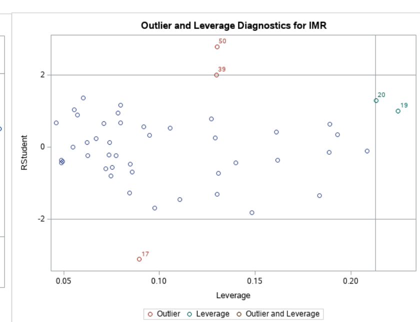




Before Removal of Observations 11, 12, and 34







After Removal of Observations 11, 12, and 34

Weighted Least Squares

The REG Procedure
Model: MODEL1
Dependent Variable: IMR IMR

Number of Observations Read	50
Number of Observations Used	50

Weight: TotalPop TotalPop

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	309316737	77329184	42.54	<.0001
Error	45	81806489	1817922		
Corrected Total	49	391123226			

Root MSE	1348.30337	R-Square	0.7908
Dependent Mean	5.56885	Adj R-Sq	0.7723
Coeff Var	24212		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	13.45397	1.02724	13.10	<.0001
MedianIncome	MedianIncome	1	-0.00007141	0.00000973	-7.34	<.0001
PctWhite	PctWhite	1	-4.16258	0.91358	-4.56	<.0001
DummyPro		1	-0.79778	0.22277	-3.58	0.0008
DummyRes		1	0.49784	0.20780	2.40	0.0208

Regression Output When Weighed by TotalPop

Results

The REG Procedure
Model: MODEL1
Dependent Variable: IMR IMR

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	44.73330	11.18332	21.43	<.0001
Error	45	23.48805	0.52196		
Corrected Total	49	68.22135			

Root MSE	0.72247	R-Square	0.6557
Dependent Mean	5.73640	Adj R-Sq	0.6251
Coeff Var	12.59441		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.71543	1.12927	11.26	<.0001
MedianIncome	MedianIncome	1	-0.00006574	0.00001154	-5.70	<.0001
PctWhite	PctWhite	1	-3.78350	0.85436	-4.43	<.0001
DummyPro		1	-0.38525	0.29402	-1.31	0.1967
DummyRes		1	0.69343	0.25295	2.74	0.0087

Our Final Regression Output

<u>Data Summary</u>

Variable Name (Type)	**Description**
State (Qualitative)	The name of the state that the collected and observed data applies to.
IMR (Quantitative)	The state's infant mortality rate (number of infant deaths for every 1,000 live births) in 2019.
PctUninsured (Quantitative)	The percentage of the state's civilian noninstitutionalized population without health insurance coverage in 2019.
MedianIncome (Quantitative)	The median household income (in 2019 dollars) of the state in 2019.
PctBachelors (Quantitative)	The percentage of the state's population over the age of 25 with a Bachelor's degree in 2019.
White (Quantitative)	The 2019 state population (number of people) that are White alone.
TotalPop (Quantitative)	The 2019 state population (number of people).
PctWhite (Quantitative)	The percentage of the 2019 state population that are White alone.
PctNonWhite (Quantitative)	The percentage of the 2019 state population that are not White alone.
Region (Qualitative)	The region that the state is located in Levels: 1. Northeast 2. South 3. Midwest 4. West
AbortionChange (Qualitative)	The change in the state's abortion access in 2019 Levels: 1. None 2. Protected 3. Restricted
GovernorParty (Qualitative)	The 2019 governor's political party affiliation Levels: 1. Democrat 2. Republican

<u>Appendix of Data</u> (First 20 Rows)

State	IMR	PctUninsured	MedianIncome	PctBachelors	White	TotalPop	PctWhite	PctNonWhite	Region	AbortionChang	GovernorParty
Alabama	7.71	9.7	51,700	25.47	3326375	4903185	0.678411074	0.321588926	South	Restricted	Republican
Alaska	4.99	12.2	75,500	29.55	469771	731545	0.64216282	0.35783718	West	None	Republican
Arizona	5.43	11.3	62,100	29.47	5701810	7278717	0.783353715	0.216646285	West	None	Republican
Arkansas	7	9.1	49,000	23.03	2315020	3017804	0.767120728	0.232879272	South	Restricted	Republican
California	4.23	7.7	80,400	33.93	23484958	39512223	0.594371974	0.405628026	West	Protected	Democrat
Colorado	4.85	8	77,100	40.91	4822379	5758736	0.83740234	0.16259766	West	None	Democrat
Connecticut	4.38	5.9	78,800	39.28	2660228	3565287	0.746146944	0.253853056	Northeast	None	Democrat
Delaware	6.44	6.6	70,200	31.99	659719	973764	0.677493725	0.322506275	South	None	Democrat
Florida	6.1	13.2	59,200	29.88	16010079	21477737	0.745426718	0.254573282	South	None	Republican
Georgia	7.02	13.4	62,000	31.32	6131699	10617423	0.577512924	0.422487076	South	Restricted	Republican
Hawaii	5.12	4.2	83,100	32.98	341211	1415872	0.240990005	0.759009995	West	Protected	Democrat
Idaho	4.31	10.8	61,000	27.57	1596918	1787065	0.893598162	0.106401838	West	Restricted	Republican
Illinois	5.67	7.4	69,200	34.65	9044059	12671821	0.713714233	0.286285767	Midwest	Protected	Democrat
Indiana	6.54	8.7	57,600	26.46	5571899	6732219	0.827646724	0.172353276	Midwest	Restricted	Republican
Iowa	5.02	5	61,700	28.57	2835536	3155070	0.898723642	0.101276358	Midwest	None	Republican
Kansas	5.42	9.2	62,100	33.38	2436902	2913314	0.836470768	0.163529232	Midwest	None	Democrat
Kentucky	4.9	6.4	52,300	24.22	3873691	4467673	0.867048909	0.132951091	South	Restricted	Democrat
Louisiana	7.97	8.9	51,100	24.11	2874784	4648794	0.618393502	0.381606498	South	Restricted	Democrat
Maine	5.43	8	58,900	31.81	1263287	1344212	0.939797443	0.060202557	Northeast	Protected	Democrat

References

Centers for Disease Control and Prevention. (2022a, June 22). *Infant mortality*. https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm.

Centers for Disease Control and Prevention. (2022b, September 30). *Stats of the states - Infant mortality*. https://www.cdc.gov/nchs/pressroom/sosmap/infant_mortality_rates/infant_Mortality.htm

Kamal, R., Hudman J., & McDermott, D. (2019, October 18). *What do we know about infant mortality in the U.S. and comparable countries?* Peterson-KFF Health System Tracker. https://www.healthsystemtracker.org/chart-collection/infant-mortality-u-s-compare-countries/#Infant%20death%20rate%20per%201,000%20live%20births,%202007-2017

Keisler-Starkey, K. & Bunch, L. N. (2020, September 15). *Health insurance coverage in the United States: 2019*. United States Census Bureau. https://www.census.gov/library/publications/2020/demo/p60-271.html

Manzanetti, Zoe. (2020, November 20). *State population by race, ethnicity data.* Governing. https://www.governing.com/now/State-Population-By-Race-Ethnicity-Data.html

Nash, E., Mohammed, L., Cappello, O., & Naide, S. (2019, December 10). *Policy trends 2019: A wave of abortion bans, but some states are fighting back.* Guttmacher Institute. https://www.guttmacher.org/article/2019/12/state-policy-trends-2019-wave-abortion-bans-some-states-are-fighting-back

National Center for Education Statistics. (n.d.). *Median household income, by state: Selected years, 1990 through 2019*. https://nces.ed.gov/programs/digest/d20/tables/dt20_102.30.asp

National Governors Association. (2022). *Governors*. https://www.nga.org/governors/

Social Explorer. (2019). *Population 25 years and over: Bachelor's degree or more* [Map]. https://www.socialexplorer.com/6147519f48/view

United States Census Bureau. (2021, October 8). *Geographic levels*. https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html

World Bank. (2022, July 5). *Infant mortality rate for the European Union*. FRED. https://fred.v stlouisfed.org/series/SPDYNIMRTINEUU