Socio-Economic effects on COVID-19 Mortality Rates : A machine learning approach

Abstract

COVID-19 was one of the most impactful events of our generation and has continued to shape our lives in ever-changing ways. While the disease has run its course through many countries, questions still remain about the true impact of the disease, and how our present socio-economic factors affected the rates of mortality from it across the United States. The purpose of this study is to use machine learning models like least squares and ridge regression to form significant results and conclusions about which socio-economic factors have the greatest impact upon the rise of mortality rates from COVID-19. Along with this goal, this study will also see if these socio-economic impacts differ at a national and regional level. The results from this study demonstrate significant evidence of specific social and economic factors impact upon COVID-19 mortality.

Introduction

COVID-19 was a worldwide phenomenon that had a variety of impacts on those living in the United States and across the globe. The scale and widespread infection of the disease caused many to question what social and economic impacts had upon the rates at which mortality occurred. While many have explained how factors, like income, have had upon the rates of which mortality have occurred (Darvas, 2021), we have yet to come to understand how factors like income, play in tandem with other factors, like healthcare, in affecting the rate at with COVID-19 mortality occurred in the United States.

The focus of this study is to understand if social and economic factors like average income, poverty levels, healthcare access, and population density have any significant impact upon the rates at which mortality from COVID-19 occurred within the United States of America.

The study will be looking at the United States at a regional and national level, attempting to find if certain impacts upon mortality rate have a more regional impact, or a more national impact. To achieve the subject of this study and its goals, three questions will be asked. (1), *Do specific social and economic factors have major impacts upon mortality rates from COVID-19 in the United States*, and (2), *Do these specific social and economic factors differ between regions of the United States?*

To facilitate the answering of these questions, this study will rely on machine learning (ML) techniques like least squares regression and ridge regression. Both of these ML techniques will allow the demonstration of possible significant results, and will also allow the study to demonstrate possible significance of specific factors over others at a regional and national level.

The results and conclusions brought about by this paper will demonstrate specific predictors significance over others in the formation of COVID-19 mortality rates, and these results demonstrated below will also demonstrate the difference of significant predictors between regions of the United States.

Finally, the results and work done in this study help to contribute to research done on COVID-19 and its effects on the United States by giving insight and information to public health workers about what fields of impact seem to have the greatest effect on mortality from diseases, including those outside the realm of COVID-19.

Literature Review

COVID-19's effects on various populations from multiple perspectives has been done by various studies. (Arias, Tejada-Vera, 2023) (Raphael, Schneider, 2023) These studies have put together multiple demonstrations to the effects that COVID-19 has had on populations, and have also demonstrated how various social and economic factors attributed to the spread and morality from the disease.

Previous work by authors has demonstrated how various ethnic communities have been impacted by the disease in specific manners.These studies have brought attention to the idea that being of a certain ethnicity can affect the rate of impact from the disease. One study conducted on the Hispanic and Latino population of the United States had revealed clues to that there is significance between ethnicity and spread of COVID-19, but the study was unable to create a hypothesis testing for the finding of significant evidence supporting the claim that Hispanics and Latinos suffered more greatly from COVID-19. (Arias, Tejada-Vera, 2023) Along with this, a second study gave more evidence to the possible impacts from COVID-19 on American populations. This second study, looking at a more broad array of statistics on social and economic factors within the United States, pointed towards conclusions of disadvantage for those who came from lower economic or social standing. (Raphael, Schneider, 2023)

Along with these two studies, more information was presented demonstrating separate, long-term impacts that COVID-19 had upon various sectors of impact across both the United States and the World in whole. This study gave general conclusions to the impacts of COVID-19, and demonstrated how data is becoming more prevalent in the research and prediction of future results for both the safety of cities and their citizens. (Brannen, 2020)

These studies, while looking at a broad range of impacts from COVID-19 on the American population, are yet to bring these social and economic factors together into one study. The results and conclusions given by these studies do point to some correlation between social and economic factors and COVID-19 spread and mortality, but they still have yet to bring all various factors together into one cohesive study. These studies also propose using data to find answers to our complex scientific questions, and that data is the future for better scientific results. While these conclusions from previous studies have importance by themselves, we have yet to overlay our social and economic factors together in the formation of better, more significant results. Therefore, the goal of this study is to fill the gaps that have been displayed by the studies shown above. To fill this gap in study, ML techniques will be used to determine the weight of factors present in the United States, and determine what caused the mortality rates to rise the way they did.

Methodology

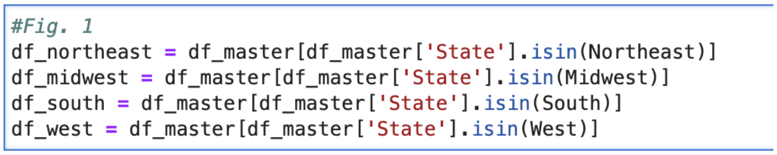
This section will outline the process for the collection of data for this study, the creation of the code for the presentation of data, and the process of getting results from the models created. All data from this project came directly from four different sources, those being the John Hopkins University, the U.S. Department of Agriculture (USDA), Forbes Magazine (who directly took all their data from the CDC) and the Federal Reserve Economic Data (FRED), which is a part of the Federal Reserve Bank of St. Louis.

To analyze and use the data effectively, the python programming language was used. To facilitate this approach, multiple packages were used like numpy, for the organization of the data, pandas, for uploading of the data and the formation of it into datasets, and sklearn, for its package which includes functions for ridge regression and least squares regression. Along with this, all suggestions for the coding and outlining the formation of the models came from the ISLP Introduction into Python textbook.

The pre-processing and cleaning for the data included the removal of unnecessary columns of data that was included in the various sources of information, and also saw the creation of a new column, which was classified by the name *Deaths Per Capita.* This column was calculated by dividing *COVID-19 Death Count* by *State Population* and was added as a column for each state.

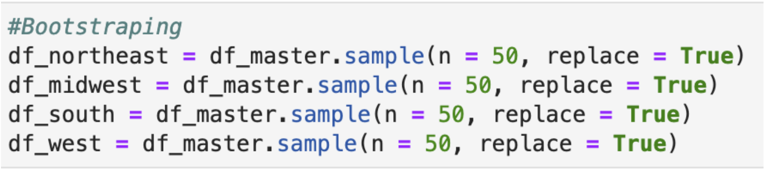
Both ridge regression and least squares regression where chosen for this study due to their ability to lower the variance-bias tradeoff, their ability to both facilitate the formation of models that can be used for prediction well, and least squares regression's ability to demonstrate the significance of each factor separately in the formation of its results. These factors and advantages caused the formation of models that demonstrated significant results, and also succeeded in predicting the outcomes of mortality with the predictors given.

To allow for the facilitation and approach using ridge regression and least squares, the data had to split into five different datasets. One master dataset was used to approach our questions from a national level, and four other datasets were used, which split the data up based on what state the data was from, since regional-based models were also being made and included in the study (Fig. 1) . Once this was complete, each dataset was then split into training and testing data for both models (Fig. 2) Since both models facilitate the use of a target variable, which is what all data is run against in order to find results, *Deaths Per Capita*, was used as our target variable.



```
#Fig. 1
df_northeast = df_master[df_master['State'].isin(Northeast)]
df_midwest = df_master[df_master['State'].isin(Midwest)]
df_south = df_master[df_master['State'].isin(South)]
df_west = df_master[df_master['State'].isin(West)]
```

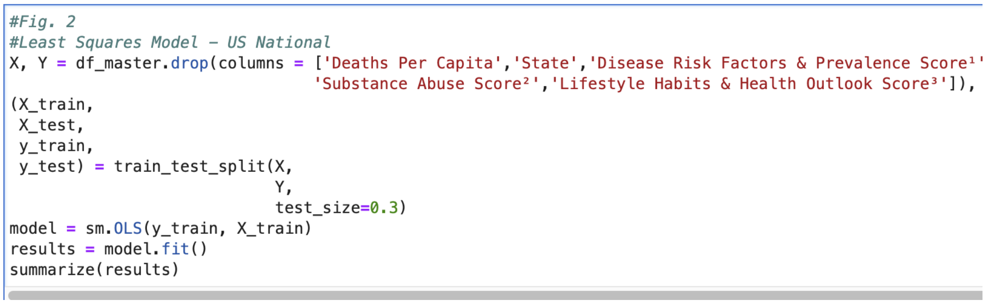
For our least squares model, since our sample sizes for our regional data frames were <= 20, a bootstrap was run before the model was created, where the resampling size will be n = 50. This allowed all assumptions to be fulfilled, and allowed the formation



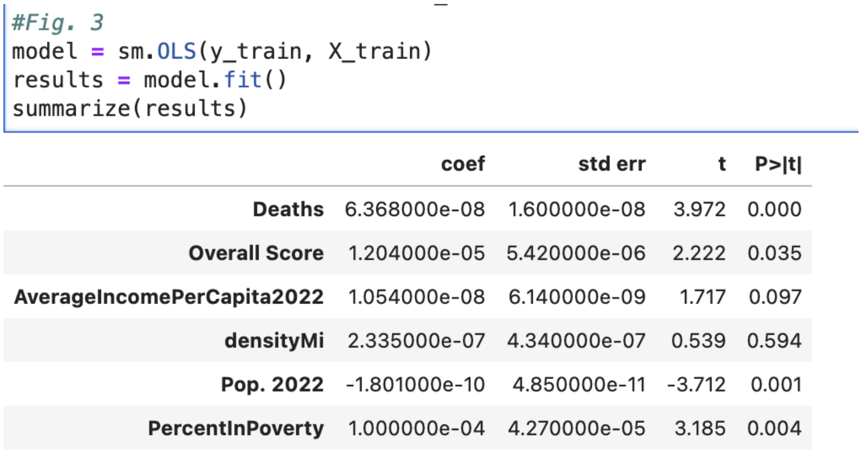
```
#Bootstraping
df_northeast = df_master.sample(n = 50, replace = True)
df_midwest = df_master.sample(n = 50, replace = True)
df_south = df_master.sample(n = 50, replace = True)
df_west = df_master.sample(n = 50, replace = True)
```

of a significant model. Once the model was run, we were given an output for both the p-values of all predictors in comparison with our target variable, and a r-squared value as well. (Fig. 3, Fig. 4) These values were then used in the building of our conclusion and results, which can be found below.

After each ridge regression was run, a secondary calculation was done to find the best alpha value for each regression. The best alpha value allows us to classify the lowest MSE possible within our regression, and allows us to also receive the best r-squared value possible after the program is run. Once this step was completed, a r-squared value was given for each regression which demonstrated the prediction capabilities of the models made.

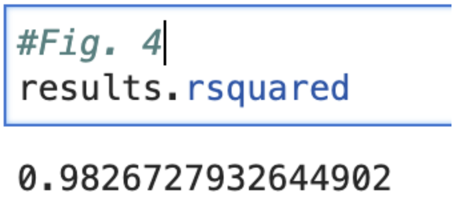


```
#Fig. 2
#Least Squares Model - US National
X, Y = df_master.drop(columns = ['Deaths Per Capita','State','Disease Risk Factors & Prevalence Score¹'
                                'Substance Abuse Score²','Lifestyle Habits & Health Outlook Score³']),
(X_train,
 X_test,
 y_train,
 y_test) = train_test_split(X,
                            Y,
                            test_size=0.3)
model = sm.OLS(y_train, X_train)
results = model.fit()
summarize(results)
```



```
#Fig. 3
model = sm.OLS(y_train, X_train)
results = model.fit()
summarize(results)
```

	coef	std err	t	P>|t|
Deaths	6.368000e-08	1.600000e-08	3.972	0.000
Overall Score	1.204000e-05	5.420000e-06	2.222	0.035
AverageIncomePerCapita2022	1.054000e-08	6.140000e-09	1.717	0.097
densityMi	2.335000e-07	4.340000e-07	0.539	0.594
Pop. 2022	-1.801000e-10	4.850000e-11	-3.712	0.001
PercentInPoverty	1.000000e-04	4.270000e-05	3.185	0.004

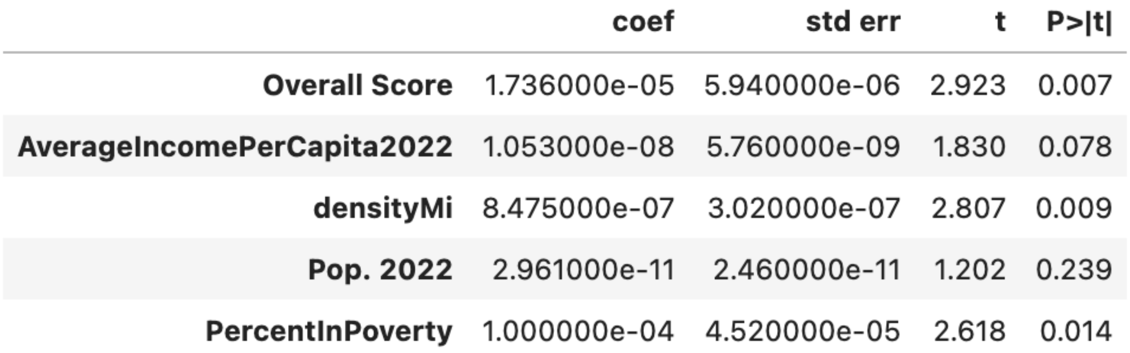


```
#Fig. 4
results.rsquared
```

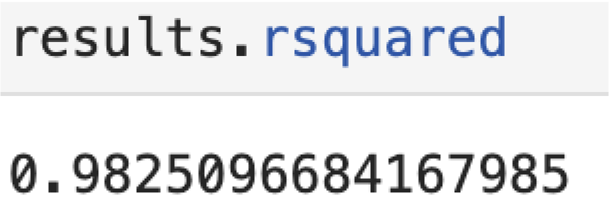
```
0.9826727932644902
```

Results/Discussion

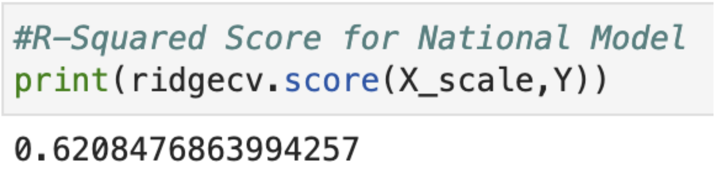
The national results of the study have allowed various conclusions to be made about the information present after the formation of both least squares and ridge regression. Looking at our national results, we can conclude (based off the p-values at a .05 significance level) that the overall health score of a state, along with its density and percentage of population below the poverty line, seem to have the greatest impacts upon the rate of mortality from COVID-19 across the United States. Along with this, the least squares regression model was also able to produce a r-squared value of 0.9825, allowing us to conclude that the results demonstrated above can be used effectively to predict and interpret the variance proposed by our dependent variable, which was *Deaths Per Capita*.



	coef	std err	t	P>|t|
Overall Score	1.736000e-05	5.940000e-06	2.923	0.007
AverageIncomePerCapita2022	1.053000e-08	5.760000e-09	1.830	0.078
densityMi	8.475000e-07	3.020000e-07	2.807	0.009
Pop. 2022	2.961000e-11	2.460000e-11	1.202	0.239
PercentInPoverty	1.000000e-04	4.520000e-05	2.618	0.014

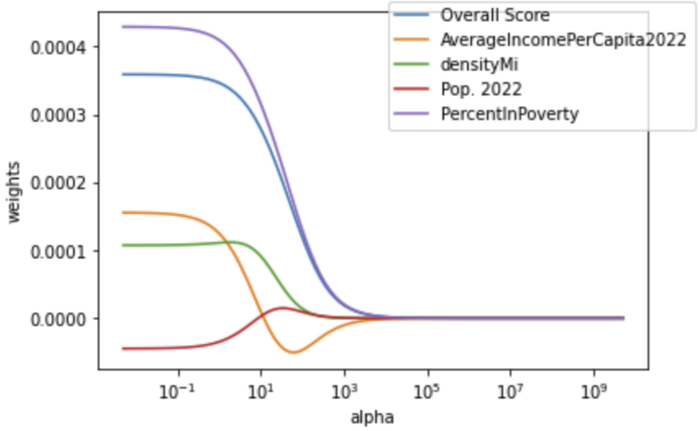


```
results.rsquared

0.9825096684167985
```



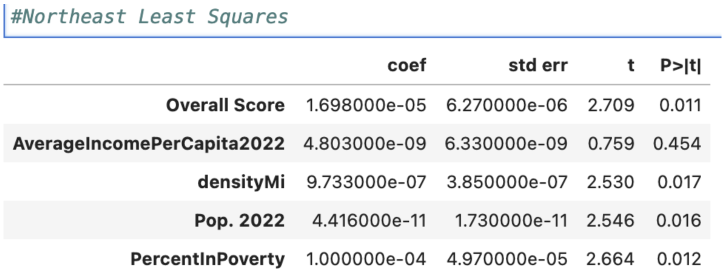
```
#R-Squared Score for National Model
print(ridgecv.score(X_scale,Y))

0.6208476863994257
```

Besides the least squares regression, the formation of a ridge regression and the calculation of a r-squared value at its best MSE was unable to improve the significance demonstrated in our least squares model, and was only able to calculate a r-squared value of 0.621 (rounded). Along with this, the tuning parameter (alpha) for ridge regression allows us to demonstrate the formation of its best and lowest MSE value, which was used to calculate the r-squared value above. While this was unable to improve our model when compared against least squares regression, it demonstrates how the ridge regression worked and attempted to form a stronger model by dampening the significance of various predictors to more modest and normalized values.



These results demonstrated above gave us significant insight into the effect of COVID-19 upon the United States at a national level, allowing us to answer the first question posed in our introduction. Along with this, results below will allow us to understand and interpret the different impacts of each variable upon the four major regions of the country.

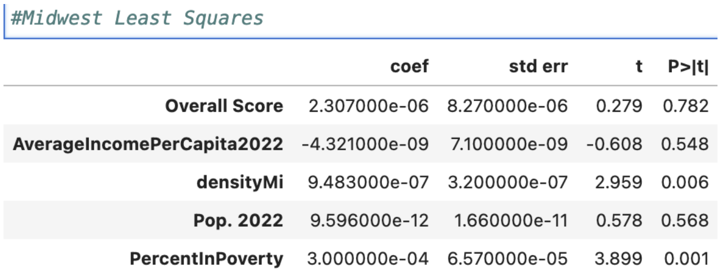
Moving onto the second question posed in our introduction, a comparison will be made between the least squares regression results of the national level, and the results of the least squares regressions from the four regions of the United States. Looking at the results from the four regions of the United States, we can make various conclusions on their difference from the national US and each other.



```
#Northeast Least Squares
```

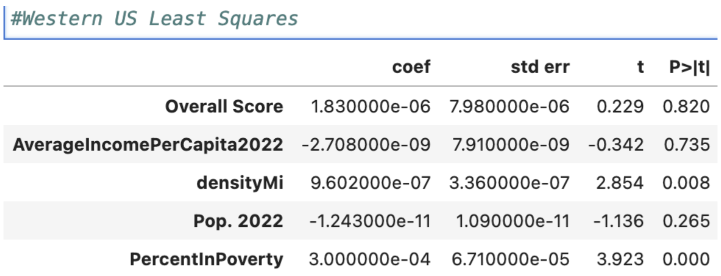
	coef	std err	t	P>|t|
Overall Score	1.698000e-05	6.270000e-06	2.709	0.011
AverageIncomePerCapita2022	4.803000e-09	6.330000e-09	0.759	0.454
densityMi	9.733000e-07	3.850000e-07	2.530	0.017
Pop. 2022	4.416000e-11	1.730000e-11	2.546	0.016
PercentInPoverty	1.000000e-04	4.970000e-05	2.664	0.012

When comparing regions of the United States and their p-values (at a .05 significance level) for predictors to the national level, the difference in significant p-values demonstrated similarities between the US and its regions. Looking at the Northeast US, this region shared all the same predictors of significance with the national level, those being the overall health score, the state population density, and the percentage below the poverty live, but this Northeastern United States also showed significance for its population levels, and that this predictor also had an impact upon the mortality rate from COVID-19 to a significant degree. Looking at the Midwest United States in comparison to the national level, both shared significance in the population density, and the percentage in poverty. Besides that, the health score rating of the midwest was found not to be significant, implying this predictor had less effect upon rates of mortality in the Midwestern United States.



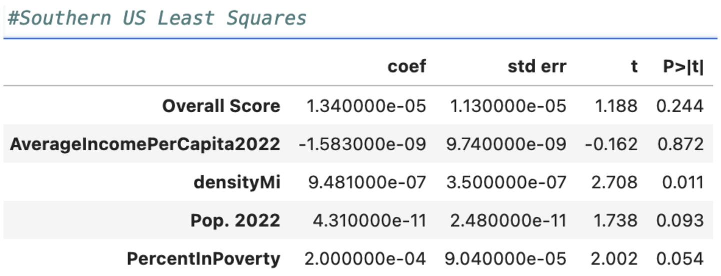
#Midwest Least Squares

	coef	std err	t	P>|t|
Overall Score	2.307000e-06	8.270000e-06	0.279	0.782
AverageIncomePerCapita2022	-4.321000e-09	7.100000e-09	-0.608	0.548
densityMi	9.483000e-07	3.200000e-07	2.959	0.006
Pop. 2022	9.596000e-12	1.660000e-11	0.578	0.568
PercentInPoverty	3.000000e-04	6.570000e-05	3.899	0.001



#Western US Least Squares

	coef	std err	t	P>|t|
Overall Score	1.830000e-06	7.980000e-06	0.229	0.820
AverageIncomePerCapita2022	-2.708000e-09	7.910000e-09	-0.342	0.735
densityMi	9.602000e-07	3.360000e-07	2.854	0.008
Pop. 2022	-1.243000e-11	1.090000e-11	-1.136	0.265
PercentInPoverty	3.000000e-04	6.710000e-05	3.923	0.000

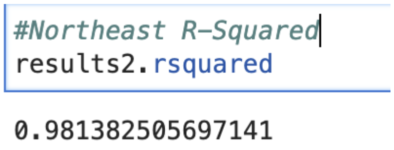


#Southern US Least Squares

	coef	std err	t	P>|t|
Overall Score	1.340000e-05	1.130000e-05	1.188	0.244
AverageIncomePerCapita2022	-1.583000e-09	9.740000e-09	-0.162	0.872
densityMi	9.481000e-07	3.500000e-07	2.708	0.011
Pop. 2022	4.310000e-11	2.480000e-11	1.738	0.093
PercentInPoverty	2.000000e-04	9.040000e-05	2.002	0.054

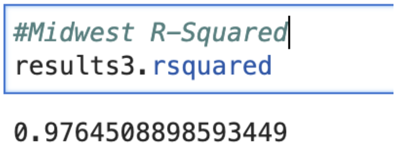
Moving onto the Western United States, this region of the country had the same outcome as our midwestern region. Both of these regions shared the same outcome in terms of significant predictors, those being population density and percentage in poverty, and the overall health score seemed to have even less significance upon the mortality rate in the Western United States. Finally, looking at the southern United States, the only significant predictor was the population density, which also was the same from the national data. Besides that, no other points of significance were found in any of our predictors in the Southern United States, and did not line up with the results from the national level least squares regression.

Along with these observations about the data from our least squares model, these results can be used with accuracy to make predictions due to the r-squared values outputted for each region of the United States. All four r-squared values were very near one, meaning they are able to effectively predict and understand the variance of our data after the models where made.



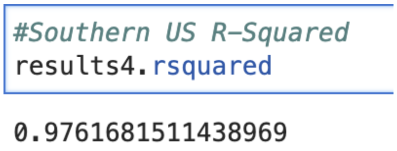
```
#Northeast R-Squared
results2.rsquared
```

```
0.981382505697141
```



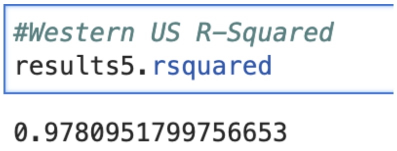
```
#Midwest R-Squared
results3.rsquared
```

```
0.9764508898593449
```



```
#Southern US R-Squared
results4.rsquared
```

```
0.9761681511438969
```



```
#Western US R-Squared
results5.rsquared
```

```
0.9780951799756653
```

When it came to the ridge regressions that were present for the regional models of the United States, the same issues posed above when discussing the national model ridge regression occurred. These models, and their shrinkage that was done, were unable to have an improvement over the least squares regression models, and therefore the least squares regression models were used for all results, as demonstrated above.

The results demonstrated above give insight into the various differences between regions of the United States. While our least squares regressions were able to show that the United States, as a whole, does share many similarities with its regions, there are also differences between those regions and the United States as a whole. One of the most interesting developments from this study was the impact factors on mortality in the Southern United States.

While many would have thought to see that the overall health score would have a significant impact upon the formation of the least squares model, it surprisingly did not have any impact. Information and conclusions like these make these studies matter, and allow us to understand that what we seem to be true never truly occurs sometimes.

Conclusion

While this study was able to demonstrate significant results and give some insight into how social and economic factors impacted rates of mortality from COVID-19 across the United States, this study still has many limitations that need to be explored. One of the major limitations this study had was its access to new information, since the economic and poverty information came from both 2022 and 2023, this data is not exactly up to date with all new and prevalent information. With the housing crisis that is occurring today within the United States, I can imagine that the percentages of those who live below the poverty line has possibly gone up, which could have affected the outcomes of the study. Additionally, with our health score calculation coming from an outside non-governmental source, the formation of its rankings could have possible bias and skew to it. While the data that was given from the Forbes magazine was significant and seemed to be outlined well, the limited understanding of where all their aggregation of data came from has yet to be understood well.

Besides these limitations of the study, the results posed and demonstrated above still go to give insight into two key questions asked within our introduction. There seems to be some correlation between the factors that were chosen for this study and the outcomes that were given, and this study was also able to demonstrate that living in various regions of the country do have an impact upon the mortality rate from COVID-19.

Works Cited

Arias, Elizabeth, and Betzaida Tejada-Vera. “Differential Impact of the COVID-19 Pandemic on Excess Mortality and Life Expectancy Loss within the Hispanic Population.” *Demographic Research*, vol. 48, 2023, pp. 339–52. *JSTOR*, https://www.jstor.org/stable/48728207. Accessed 12 June 2024.

Brannen, Samuel, et al. *Covid-19 Reshapes the Future*. Center for Strategic and International Studies (CSIS), 2020. *JSTOR*, http://www.jstor.org/stable/resrep25198. Accessed 28 June 2024.

DARVAS, ZSOLT. *THE UNEQUAL INEQUALITY IMPACT OF THE COVID-19 PANDEMIC*. Bruegel, 2021. *JSTOR*, http://www.jstor.org/stable/resrep32250. Accessed 6 June 2024.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An Introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

RAPHAEL STEVEN, and DANIEL SCHNEIDER. “Introduction: The Socioeconomic Impacts of COVID-19.” *RSF: The Russell Sage Foundation Journal of the Social Sciences*, vol. 9, no. 3, 2023, pp. 1–30. *JSTOR*, https://www.jstor.org/stable/48724527. Accessed 12 June 2024.