HIGH SCHOOL EDITION
Journal of Student Research

Volume 11 Issue 2 (2022)

Applying Machine Learning Techniques to Mitigate Impact of COVID-19 Pandemic

Sidarth Krishna[1], Rajagopal Appavu[#] and Jothsna Kethar[#]

[1] Acton-Boxborough Regional High School, Acton, MA, USA
[#]Advisor

ABSTRACT

Since March 2020, COVID-19 has played a very influential role in our lives. Totaling over 300 million cases and 5.5 million deaths worldwide it has been one of the most transmittal viruses humans have seen in recent generations. Even after the mass distribution of vaccines, COVID-19 shows no signs of stopping. This is because many communities that are especially struggling during this time period have not been identified and are not being helped adequately enough. By better understanding how different factors in communities such as ethnic percentages, poverty rates and much more can help us determine which communities need to be addressed to slow the spread of COVID-19. To identify the most significant of these demographic factors an in depth data analysis using machine learning models and regression analysis were carried out on various datasets. The results highlighted that for COVID-19 cases the most influential factor was Population Density. For deaths, the most significant factors were poverty rates in communities as well as education level. From this analysis and results, in order to mitigate the impact of the COVID-19 pandemic in the future it is of utmost importance to address the needs of underprivileged communities by providing access to low cost and high quality medical resources for all.

Introduction

The death toll and impact on human lives has grown exponentially, since the official declaration of COVID-19 as an international pandemic on March 11, 2020 (World Health Organization, 2022). Currently, there are over 320 million COVID-19 cases resulting in 5.5 million deaths, making SARS-CoV-2, the COVID-19 virus, one of the most deadly viruses to hit mankind in the past century (World Health Organization, 2022). Billions of dollars have been spent on research and trials to create vaccines with the hopes of finally putting an end to the devastating impacts on the world (World Health Organization, 2022). Even as our understanding keeps growing by the day, COVID-19 remains a threat as more mutations create variants, some more contagious and deadly than others (Roy & Ghosh, 2020). It is important to utilize data on the deadly virus to analyze the past as well as the future of the pandemic and the societal impact . In order to predict the future of this pandemic and the areas that will be hit the hardest, it is of utmost importance to determine the most significant and influential factors for COVID-19. Using these significant factors determined from various models and analysis of data we can predict what future trends of COVID-19 will look like in various places (Roy & Ghosh, 2020). This data analysis paper will begin with a literature review of prior research focusing on demographic factors, socioeconomic factors and comorbid diseases and conditions. Then, machine learning and analysis of COVID-19 data from the United States will be used to fill the gaps of prior research. Using the results of these models and analysis, conclusions will be drawn about how these factors can predict future trends in the United States for COVID-19 as it continues affecting people’s lives.

Understanding Significant COVID-19 Risk Factors

After over one and a half years of this novel disease, COVID-19 has shown a very diverse set of symptoms and effects. There have been growing cases of asymptomatic patients, who carry and spread this disease without knowing it since they do not show any symptoms (Monita 2021). However, millions of others have not been as lucky, because their COVID-19 symptoms can be critical and often result in the death of the individual (Monita 2021). These severe cases are often marked with many common symptoms including a massive decline in lung condition, blood oxygen saturation, respiratory frequency decreases and long term pulmonary and lung diseases (Gao et. al, 2020). There has been many studies and research over the past year trying to understand what determines the severity of COVID-19 cases in different individuals. These studies have pointed to many important factors that are the most influential in the COVID-19 cases leading to deaths in patients. Some of these important factors, which will be discussed in this paper are age, geographical region, gender and race (Figure 1) (Gao et. al, 2020). Furthermore, this paper will look at what role comorbidity factors play in the development of severe COVID-19 cases.

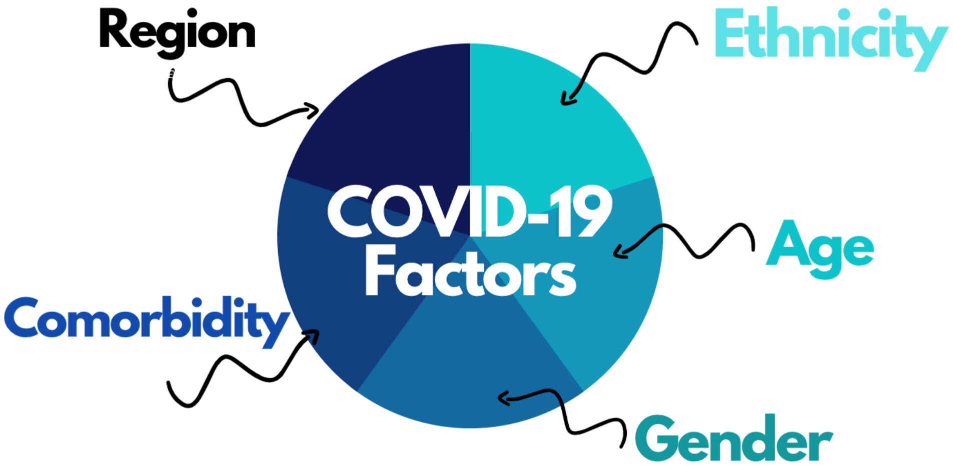


Figure 1. Diagram of the Most Important COVID-19 Factors from Research created by Sidarth K.

Demographic Factors

Demographic factors have been commonly cited in research papers as heavily influencing COVID-19 cases and Demographic factors have been commonly cited in research papers as heavily influencing COVID-19 cases and deaths. Demographic factors are characteristics that define the population and society that one is in (Gao et. al, 2020). This includes characteristics such as population numbers, gender distribution and other socioeconomic factors (Gao et. al, 2020). The most influential of these demographic factors from this research included gender, age distribution of especially the senior population, ethnicity and finally geographical region (Gao et. al, 2020; Kopel et. al, 2020; Magesh, 2021).

Gender

The male gender is most strongly correlated to COVID-19 cases as well as developing severe symptoms. Firstly, males have a higher association with smoking compared to females (Gao et. al, 2020). Smoking, a lifestyle choice, is strongly shown to have long term impacts on the health and respiratory system of an individual (Gao et. al, 2020). This has been one reason why there is such a large discrepancy in the gender populations and percentages that have severe hospitalization COVID-19 cases. Biologically, the differences in the hormones between males and females makes females more resistant to the COVID-19 infections than men (Gao et. al, 2020). This is due to the differences in sex hormones with a high expression of coronavirus receptors (ACE 2) in men as well as a difference in the levels of TMPRSS2 (Gao et. al, 2020; Kopel et. al, 2020). This results in males having a higher likelihood of attracting the SARS-CoV-2 infection in their bloodstream and developing more severe symptoms (Gao et. al, 2020). In a prior US study, over 83 percent of the patients who had to be treated with mechanical ventilation were male (Gao et. al, 2020). Furthermore, in a study from Wuhan, China 56 percent of the patients with COVID-19 were males and 67 percent of the fatalities were male (Kopel, 2020). Through biological differences and lifestyle differences, populations with a higher percentage of males are more likely to face a higher death rate after developing COVID-19 (Gao et. al, 2020; Kopel et. al, 2020).

Age Distribution

At this point it is widely known that the older population is at a much higher risk of developing serious COVID-19 symptoms compared to the younger population. One study stated that the median age of those receiving intensive care was 66 years old, which was much higher than the median age of those with less severe COVID-19 cases at 51 years (Gao et. al, 2020. Furthermore, for people who tested positive for the virus, 19.8% to 49% of adults had severe COVID-19 cases, whereas only 2.2% of the pediatric cohort had severe cases (Gao et. al, 2020). From data analysis derived from tens of thousands of cases in the world, those who were over 59 years of age were 8.5 times more likely to die from COVID-19 symptoms compared to those under 30 years (Gao et. al, 2020). This huge discrepancy between the effects in the pediatric population versus the geriatric population is caused because of the physiological changes of the immune system that come from aging as well as the higher percentage of comorbidity factors present in the geriatric population (Ludvigsson, 2020; O’Driscoll, 2020). Some of these comorbidity factors include diabetes, kidney and lung disease and hypertension (O’Driscoll, 2020). Each of these pre-existing medical conditions increases the risk heavily for patients developing severe cases of COVID-19 (Ludvigsson, 2020; O’Driscoll, 2020).

Ethnicity

Many studies have shown that the minority populations have a higher mortality risk compared to majority populations nationwide (Ford et. al, 2020; Magesh 8; Monod et. al, 2021). In a recent study analyzing ethnicity and COVID-19, per 100k people the black ethnicity had 178 deaths caused by COVID-19, followed by 172 deaths per 100k for Native Americans and 154 deaths per 100k for Hispanics (Figure 2) (Monod et. al, 2021). Minority races are over 1.4 times likely to die from COVID-19 than the white population (Monod et. al, 2021). The two largest minority populations facing the greatest consequences of COVID-19 are the African Americans and the Hispanics (Ford et. al, 2020). The African American distribution in the US states especially centers around the southeast of the United States (Figure 3). Many of these states have especially been hit hard with high COVID-19 cases and deaths. Hispanics, on the other hand, are mainly distributed in the southwest of the US, where places like Texas, Arizona and California have had a very high number of COVID-19 case rate and death rate (Figure 4). These discrepancies in the minority ethnic populations of the US could be caused by the fact that they do not contain access to the same level of medical resources and health practices as other populations (Ford et. al, 2020; Kopel et. al, 2020; Monod et. al, 2021). This can lead to a much higher percentage of individuals not getting the proper medical treatment they need to prevent themselves from developing severe cases (Monod et. al, 2021).

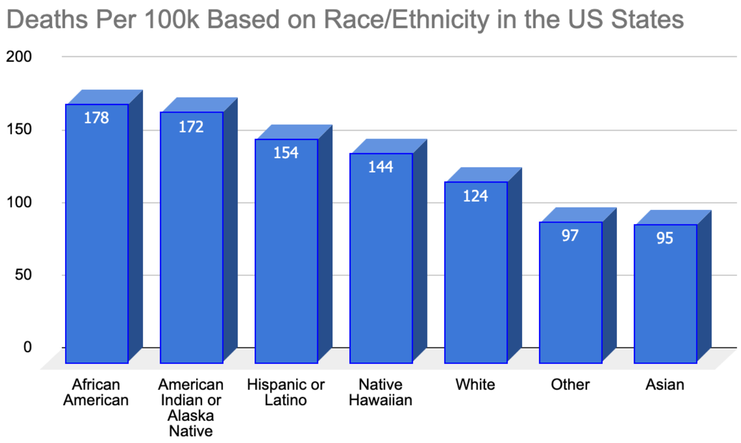


Figure 2. Bar Graph Showing Deaths Per 100k by COVID-19 based on Ethnicity in the US created by Sidarth K.

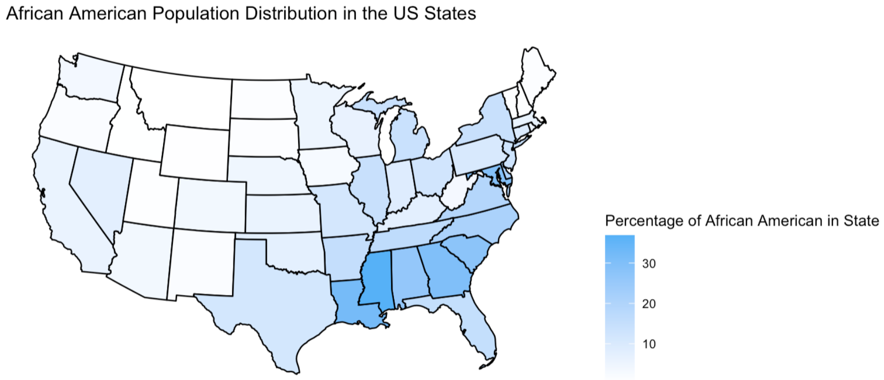

African American Population Distribution in the US States
Percentage of African American in State
30
20
10

Figure 3. Heatmap of African American Population Distribution in the US States created by Sidarth K.

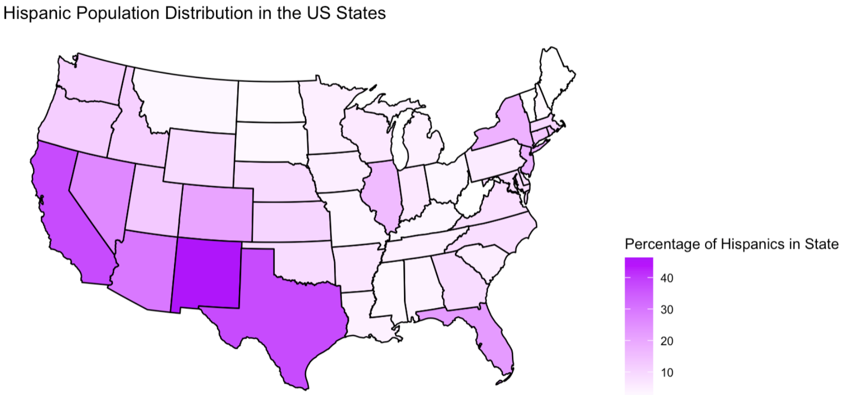

Hispanic Population Distribution in the US States
Percentage of Hispanics in State
40
30
20
10

Figure 4. Heatmap of the Hispanic Population Distribution in the US States created by Sidarth K.

Comorbidity Factors

After analyzing demographic factors in a society, the risks of COVID-19 for each individual in a society should be understood. After developing COVID-19 though the most important risk factors are comorbidity factors, which are the underlying conditions of an individual (Gao et. al, 2020. These comorbidity factors play a very influential role in determining how likely an individual is to get symptoms from the COVID-19 infection and also how severe their case will be (Gao et. al, 2020. Based on many studies, the most influential of the comorbidity factors that this paper will cover are diabetes and hypertension (Gao et. al, 2020).

Diabetes

Diabetes is a group of diseases that result in too much sugar in the blood or a high blood glucose (Gao et. al, 2020). There are different types of diabetes including Prediabetes, gestational diabetes, type 1 diabetes and type 2 diabetes. Gestational diabetes is high blood sugar affecting pregnancy. Prediabetes is where the blood sugar is very high but is not yet type 2 diabetes (Gao et. al, 2020). With type 1 diabetes, the pancreas is unable to produce enough insulin to reduce the blood glucose levels (Gao et. al, 2020). Type 2 diabetes affects the ability of the body to process blood sugar (Gao et. al, 2020). Diabetes can cause fatigue, weight loss, blurred vision and can affect blood pressure (Gao et. al, 2020). Any stage or type of diabetes can play a large role in the severity of developing COVID-19 symptoms (Gao et. al, 2020). Furthermore, type 2 diabetes which is developed over time in the body, increases the expression of ACE2 in the body which is the entry receptor of SARS-CoV-2 (Gao et. al, 2020). This increases the activation of the cells to allow the COVID-19 virus to enter the lungs and other tissues which has been shown to lead to more severe COVID-19 responses in patients (Gao et. al, 2020).

Hypertension

Hypertension is observed to differ significantly between COVID-19 patients having severe symptoms versus patients having nonsevere symptoms (Gao et. al, 2020). Fifty percent of hospitalized patients had hypertension (Kopel et. al, 2020). In another study, patients requiring ICU (Intensive care unit) over 58 percent had hypertension (Gao et. al, 2020). This shows a strong positive correlation between hypertension and the probability of developing a very severe case of COVID-19 (Gao et. al, 2020; Kopel et. al, 2020). Hypertension also has a correlation with the older population which is one of the reasons that the geriatric population has been seen to have an increased risk of COVID-19 (O'Driscoll, 2020).

Methods

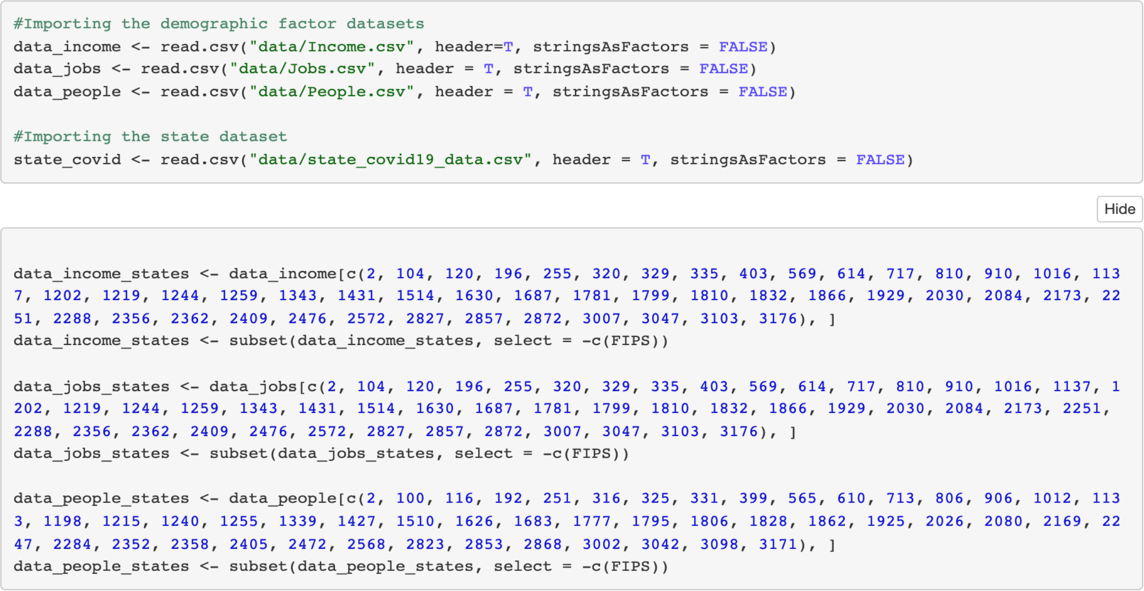
Data Selection

For my data analysis, I used two datasets. The first dataset was taken from Kaggle. This dataset contained each of the 50 US states and their total COVID-19 cases and deaths. This dataset was selected because it contained the most up to date information out of all of the datasets I could find. Furthermore, it used COVID-19 data from Johns Hopkins, a very reputable source. For the second part of data that I needed on the socioeconomic factors and demographic factors for each of the US states I selected data from the USDA (US Department of Agriculture) . This USDA data had only been updated in June of 2021, but the data was from the US government which is a very reputable source . Furthermore, the data contained in this dataset would not have changed drastically over the past 6 months or enough to create a large enough difference to cause problems when the regression model is executed.

Data Cleaning

The first dataset from Kaggle contained the 50 US states as well as a set of COVID-19 variables including total cases, total deaths, total cases per million and total deaths per million. Since my data analysis paper was comparing different states to each other to determine what the most significant factor in determining COVID-19 severe cases were, I cleaned this dataset to only include the cases and deaths per million, standardizing the data for all states. As the regression model used in this paper only focused on the continental US states I spliced out both Alaska and Hawaii (Figure 5).

The data from USDA came from 3 different csv files. Each of them consisted of different demographic factors of society including factors for people, jobs and income. Furthermore, each of the different entries for these three datasets were at the county level instead of the state level. In order to get them to the state level, I worked on splicing out an entry for each of the 50 states that took the average of each of the country variables and combined them into a single entry for the state. Once I had cleaned each of the 3 different csv files I combined them into a new csv file for the 50 states called country_data. This dataset contained over 175 variables from the culmination of the three csv files. After looking through each of these 175 variables, I removed the ones that were not relevant to my study such as netmigrationrate0010 (net migration rate from 2000-2010) since many of them were outdated. After careful assessment and removal of variables that are either irrelevant or do not contribute significantly, the country_data was left with 48 variables including the states covid-19 deaths and cases per million.



```
#Importing the demographic factor datasets
data_income <- read.csv("data/Income.csv", header=T, stringsAsFactors = FALSE)
data_jobs <- read.csv("data/Jobs.csv", header = T, stringsAsFactors = FALSE)
data_people <- read.csv("data/People.csv", header = T, stringsAsFactors = FALSE)

#Importing the state dataset
state_covid <- read.csv("data/state_covid19_data.csv", header = T, stringsAsFactors = FALSE)
```

```
data_income_states <- data_income[c(2, 104, 120, 196, 255, 320, 329, 335, 403, 569, 614, 717, 810, 910, 1016, 113
7, 1202, 1219, 1244, 1259, 1343, 1431, 1514, 1630, 1687, 1781, 1799, 1810, 1832, 1866, 1929, 2030, 2084, 2173, 22
51, 2288, 2356, 2362, 2409, 2476, 2572, 2827, 2857, 2872, 3007, 3047, 3103, 3176), ]
data_income_states <- subset(data_income_states, select = -c(FIPS))

data_jobs_states <- data_jobs[c(2, 104, 120, 196, 255, 320, 329, 335, 403, 569, 614, 717, 810, 910, 1016, 1137, 1
202, 1219, 1244, 1259, 1343, 1431, 1514, 1630, 1687, 1781, 1799, 1810, 1832, 1866, 1929, 2030, 2084, 2173, 2251,
2288, 2356, 2362, 2409, 2476, 2572, 2827, 2857, 2872, 3007, 3047, 3103, 3176), ]
data_jobs_states <- subset(data_jobs_states, select = -c(FIPS))

data_people_states <- data_people[c(2, 100, 116, 192, 251, 316, 325, 331, 399, 565, 610, 713, 806, 906, 1012, 113
3, 1198, 1215, 1240, 1255, 1339, 1427, 1510, 1626, 1683, 1777, 1795, 1806, 1828, 1862, 1925, 2026, 2080, 2169, 22
47, 2284, 2352, 2358, 2405, 2472, 2568, 2823, 2853, 2868, 3002, 3042, 3098, 3171), ]
data_people_states <- subset(data_people_states, select = -c(FIPS))
```

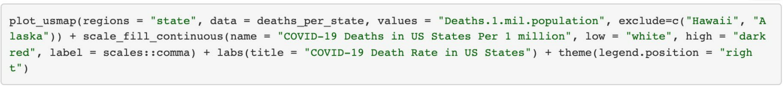
Figure 5. RStudio Code to Import and Clean Data

Analysis

Exploratory Data Analysis

Visualization of Cases in US States per million and Deaths in US States per 1 million in a heatmap are present in both figure 6 and 7. Through just viewing these heatmaps North Dakota, a state with a smaller population, has the highest number of cases per million out of all US states. For the deaths in US States, Arizona, New York, Mississippi and Alabama all appear to have the highest number of deaths in US states per 1 million. However those states did not have the highest number of cases. The comparison between the two maps shows that there is no definite correlation between COVID-19 cases per million and COVID-19 deaths per million.

For states such as Washington and Oregon with a very low COVID-19 cases rate there is a correlation to it also having the lowest death rate among the US states. Through this simple data visualization of the cases and deaths in the US states that there is more research that needs to be done to understand the complexity of what factors influence the COVID-19 mortality rate.



```
plot_usmap(regions = "state", data = deaths_per_state, values = "Deaths.1.mil.population", exclude=c("Hawaii", "A
laska")) + scale_fill_continuous(name = "COVID-19 Deaths in US States Per 1 million", low = "white", high = "dark
red", label = scales::comma) + labs(title = "COVID-19 Death Rate in US States") + theme(legend.position = "righ
t")
```

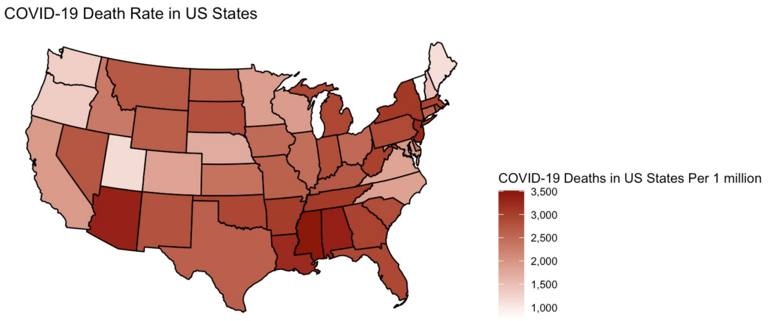
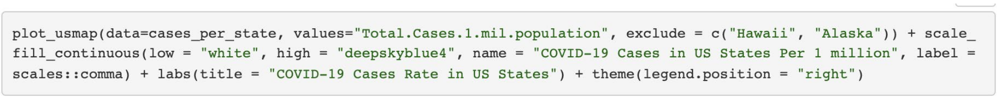

COVID-19 Death Rate in US States
COVID-19 Deaths in US States Per 1 million
3,500
3,000
2,500
2,000
1,500
1,000

Figure 6. RStudio Code and Heatmap of COVID-19 Deaths per 1 Million People In Each US State created by Sidarth K.



```
plot_usmap(data=cases_per_state, values="Total.Cases.1.mil.population", exclude = c("Hawaii", "Alaska")) + scale_
fill_continuous(low = "white", high = "deepskyblue4", name = "COVID-19 Cases in US States Per 1 million", label =
scales::comma) + labs(title = "COVID-19 Cases Rate in US States") + theme(legend.position = "right")
```

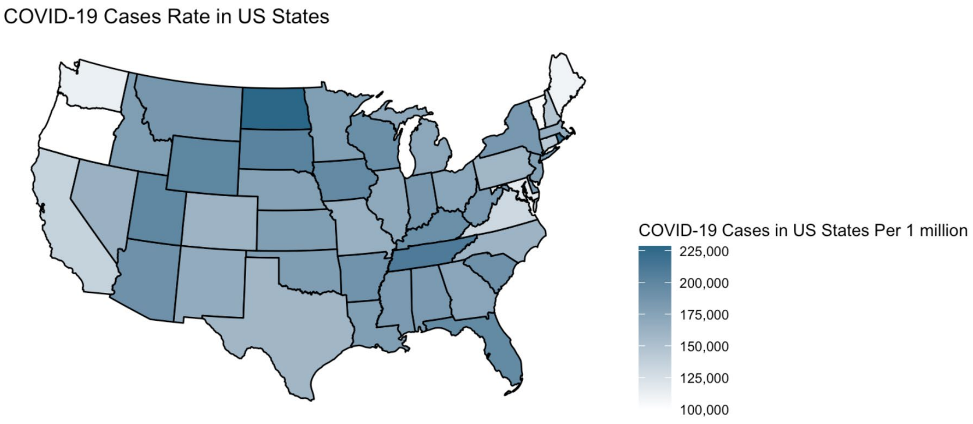
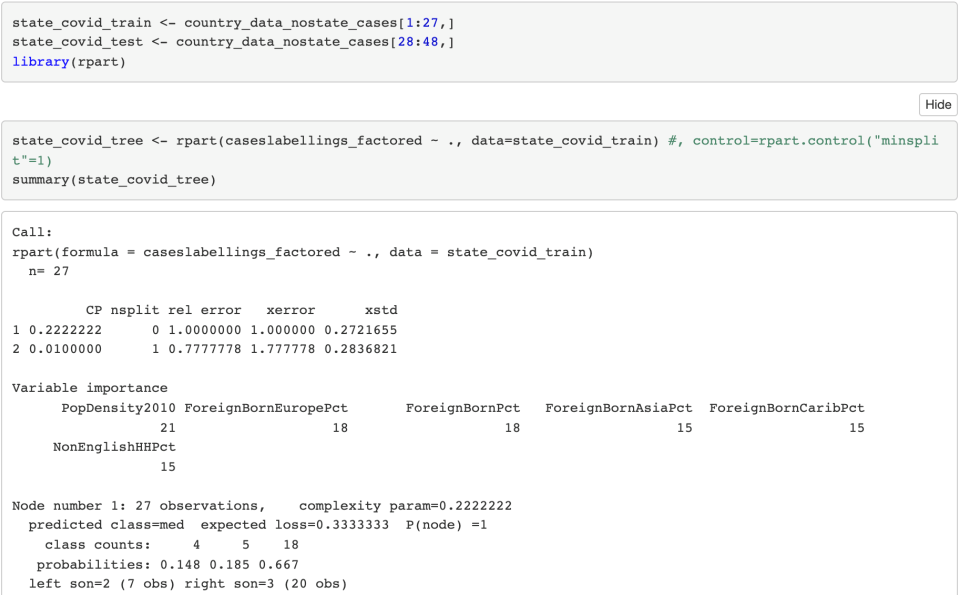

COVID-19 Cases Rate in US States
COVID-19 Cases in US States Per 1 million
225,000
200,000
175,000
150,000
125,000
100,000

Figure 7. RStudio Code and Heatmap of COVID-19 Cases per 1 Million People In Each US State created by Sidarth K.

Machine Learning and Regression Model

To analyze this data, two statistical methods were chosen. The first one chosen was a machine learning model using training data to figure out the most significant factors influencing COVID-19 cases and COVID-19 deaths in each of the states. The second one chosen was multiple regression to find a quantitative relationship between the most significant factor(s) and COVID-19 cases and deaths (Figure 9). Based on my literature search I determined that the most reliable model for the purposes of my research is the rpart regression model. The rpart programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. The R programming language was used to implement the rpart function to create CART (classification regression trees) which is a machine learning tool to find the most significant factors in a set of several possible variables to correctly determine the resulting case and death levels (Figure 8). R was also used to create scatterplots to find variables that were linearly correlated with cases and deaths and the most significant variables as determined from the CART analysis were used to create a multiple regression model.

The dataset contained total COVID-19 cases per million people, COVID-19 deaths per million people and values for demographic factors in each state of the United States. These demographic factors included percent of people who only have a high school diploma, percent of foreign born people, percent of African Americans, percent of households run by women and others. To run this model the COVID-19 cases and deaths were broken into low, medium and high where medium represented the range from quartile 1 to 3, low represented the number of cases/deaths below quartile 1 and high represented the number of cases/deaths above quartile 3.



```
state_covid_train <- country_data_nostate_cases[1:27,]
state_covid_test <- country_data_nostate_cases[28:48,]
library(rpart)
```

```
state_covid_tree <- rpart(caseslabellings_factored ~ ., data=state_covid_train) #, control=rpart.control("minspli
t"=1)
summary(state_covid_tree)
```

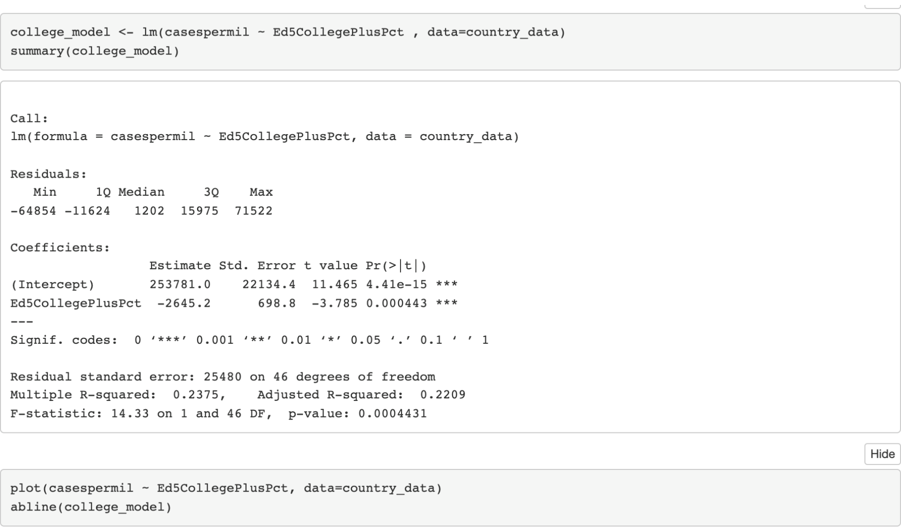
```
Call:
rpart(formula = caseslabellings_factored ~ ., data = state_covid_train)
  n= 27 

         CP nsplit rel error   xerror      xstd
1 0.2222222      0 1.0000000 1.000000 0.2721655
2 0.0100000      1 0.7777778 1.777778 0.2836821

Variable importance
      PopDensity2010 ForeignBornEuropePct       ForeignBornPct   ForeignBornAsiaPct  ForeignBornCaribPct 
                  21                   18                   18                   15                   15 
     NonEnglishHHPct 
                  15 

Node number 1: 27 observations,    complexity param=0.2222222
  predicted class=med  expected loss=0.3333333  P(node) =1
    class counts:     4     5    18
   probabilities: 0.148 0.185 0.667 
  left son=2 (7 obs) right son=3 (20 obs)
```

Figure 8. RStudio Code for Machine Learning Model



```
college_model <- lm(casespermil ~ Ed5CollegePlusPct , data=country_data)
summary(college_model)
```

```
Call:
lm(formula = casespermil ~ Ed5CollegePlusPct, data = country_data)

Residuals:
   Min     1Q Median     3Q    Max 
-64854 -11624   1202  15975  71522 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       253781.0    22134.4  11.465 4.41e-15 ***
Ed5CollegePlusPct  -2645.2      698.8  -3.785 0.000443 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25480 on 46 degrees of freedom
Multiple R-squared:  0.2375,	Adjusted R-squared:  0.2209 
F-statistic: 14.33 on 1 and 46 DF,  p-value: 0.0004431
```

```
plot(casespermil ~ Ed5CollegePlusPct, data=country_data)
abline(college_model)
```

Figure 9. RStudio Code for Regression Analysis

COVID-19 Cases Model

After running the model for cases-per-million, I determined the most significant variables and their respective variable importance values. Figure 10 below includes the 5 most significant variables.

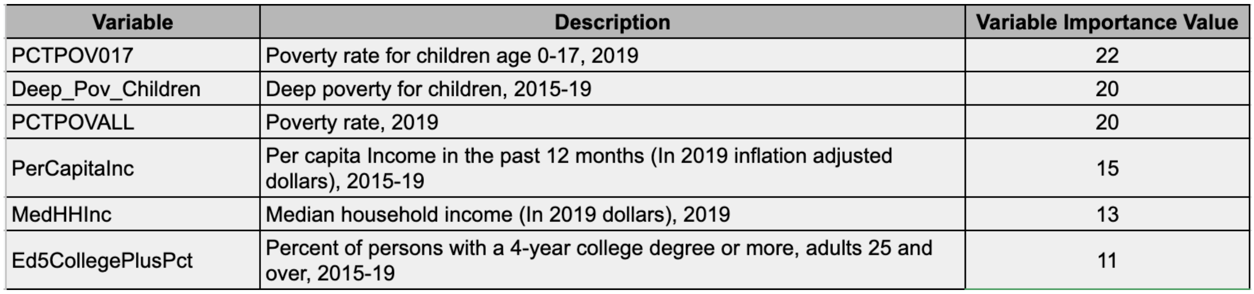
Variable	Description	Variable Importance Value
PopDensity2010	Population density, 2010	21
ForeignBornEuropePct	Percent of persons born in Europe, 2015-19	18
ForeignBornAsiaPct	Percent of persons born in Asia, 2015-19	15
ForeignBornCaribNum	Number of persons born in the Caribbean, 2015-19	15
NonEnglishHHPct	Percent of non-English speaking households of total households, 2015-19	15

Figure 10. Most Significant Variables Determined by Model for COVID-19 Cases created by Sidarth K.

Based on the classification tree for cases (Figure 14), the population density in 2010 (Figure 12) greater than 214 indicates low cases with a probability of 0.57. A population density in 2010 less than or equal to 214 indicated medium cases with a probability of 0.80. Although higher population density might intuitively indicate higher rate of spreading of COVID-19, the higher population in certain states may be a reflection of other factors that could lower the spread of COVID-19, for example more factories to create a higher distribution of supplies for hospitals, or more people because of more jobs in the hospitals leading to more awareness and lower spread rates. The other four significant factors determined were all about immigration percentage in the states. This can highlight that increased traveling in a society could be a large factor in COVID-19 cases because of how easy this disease has spread throughout the world and is being transferred and brought to the US through other countries.

COVID-19 Deaths Model

After the model for the COVID-19 deaths the most significant factors and their corresponding significance levels. Figure 9 below includes the 6 most significant variables.



Variable	Description	Variable Importance Value
PCTPOV017	Poverty rate for children age 0-17, 2019	22
Deep_Pov_Children	Deep poverty for children, 2015-19	20
PCTPOVALL	Poverty rate, 2019	20
PerCapitaInc	Per capita Income in the past 12 months (In 2019 inflation adjusted dollars), 2015-19	15
MedHHInc	Median household income (In 2019 dollars), 2019	13
Ed5CollegePlusPct	Percent of persons with a 4-year college degree or more, adults 25 and over, 2015-19	11

Figure 11. Most Significant Variables Determined by Model for COVID-19 Deaths created by Sidarth K.

The classification model for deaths (Figure 13) showed that percent of population in poverty (PCTPOVALL) was the most significant factor for fatal COVID-19 cases relative to the other demographic factors present in this dataset. With a percent of population in poverty greater than or equal to 13% the model indicated a high death rate with a probability of .37 and less than 13% of the population in poverty indicates a high death rate with a probability of .63. This may be because higher poverty rates are correlated with a lower access to medical supplies that are needed for severe COVID-19 symptoms (Gao et. al, 2020.   This results in the positive correlation between increasing poverty rates and increasing COVID-19 deaths per million (Figure 15). This finding makes sense to documented findings in literature (Link et. al, n.d.) that families from lower economic backgrounds suffer from more illnesses associated with high mortality, such as diabetes, heart disease, and pulmonary issues. The next 4 factors determined by the model are all related to poverty and household income. The last factor though is about education level with it having a negative correlation with COVID-19 deaths.

Discussion

Clearly, demographic factors have a strong influence on COVID-19 cases and deaths.

In order to quantify some of the significant factors, simple regression models were made, since multiple regression models were determined to not be necessary. The relationships determined showed for every increase in 1 percent of total poverty, 109.25 more deaths per 1 million people due to COVID-19 occurred, which was significant (P-value <.002) (Figure 13). Additionally, for every increase of 1 percent of people with some college experience, the cases per million decreased by 2645.2, which was also significant (P-value < .0005) (Figure 16). This shows that a higher education rate in a community helps with decreasing the number of COVID-19 cases and deaths.

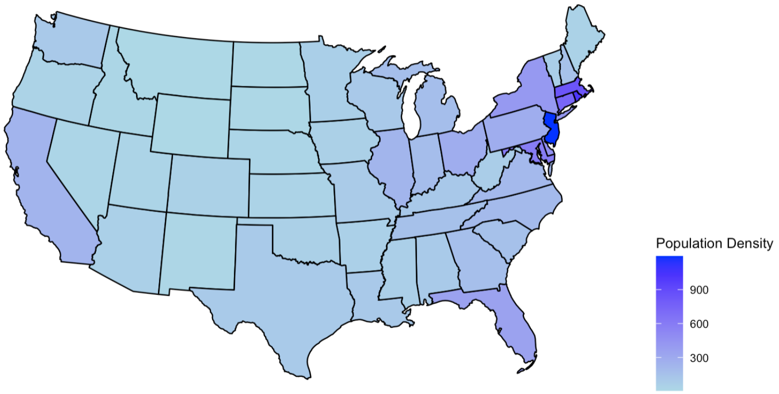


Figure 12. Heatmap of Population Density in Units of People Per Square Mile In Each US State created by Sidarth K.

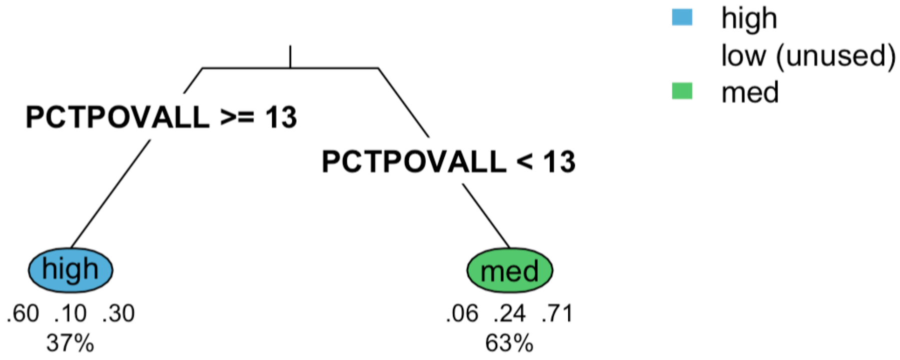


Figure 13. Classification Tree to Determine Level of Deaths Based on Percentage of Population in Poverty (PCTPOVALL) in a US State created by Sidarth K.

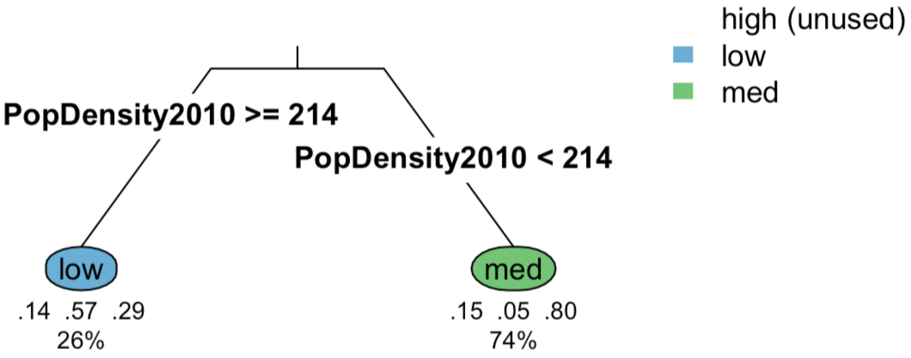


Figure 14. Classification Tree to Determine Level of Cases Based on Population Density (Number of People per Square Mile) in each US State created by Sidarth K.

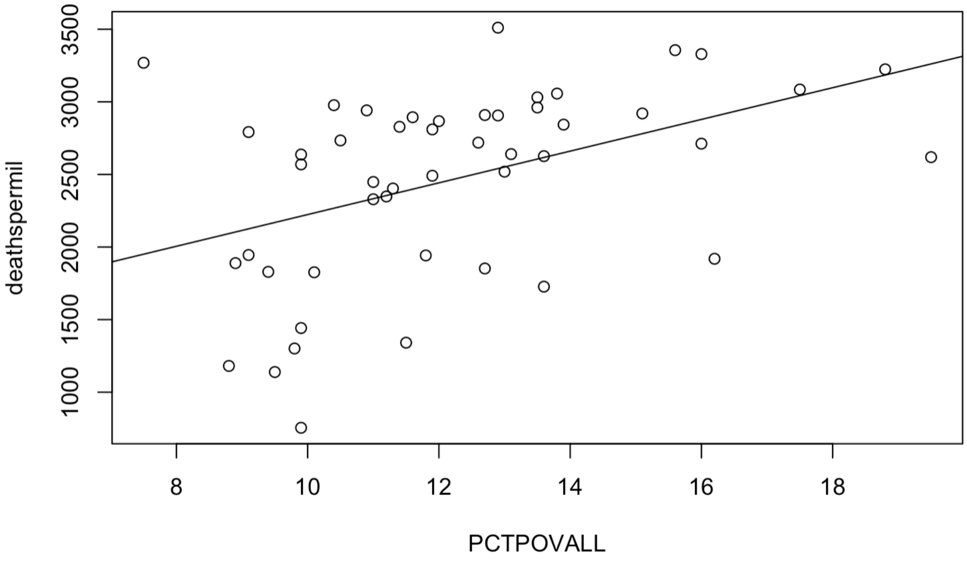


Figure 15. Simple Regression Model of Percent Poverty All (PCTPOVALL) versus Deaths per Million created by Sidarth K.

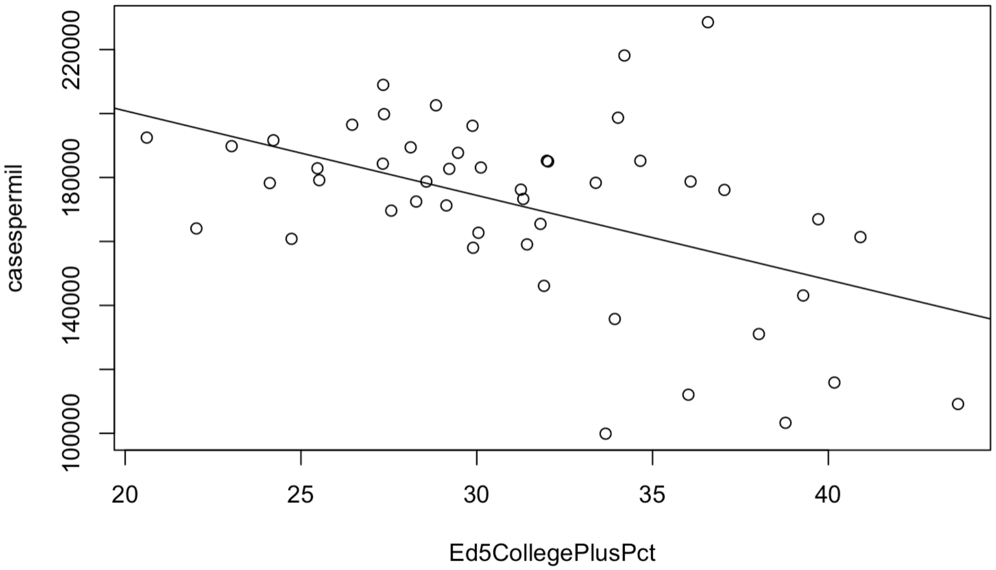


Figure 16. Simple Regression Model of Percent with College 4 plus years (Ed5CollegePlusPct) versus Cases per Million created by Sidarth K.

Conclusion

This paper leverages machine learning techniques to analyze COVID-19 data from US states based on reliable statistical methods and determine key demographic factors that had significantly higher probability to predict deaths due to COVID-19. The poverty rates in the states, the education levels in communities and the primary environmental factors that influence likelihood of spread of any epidemic such as population density were found to be most significant. The results of the data analysis included in this paper highlight that the published findings of research work on the impact of COVID-19 have to be applied selectively to address different sections of the society appropriately. This research along with the other factors highlighted in other studies can be used to project how COVID-19 statistics will trend in the near future. In order to limit additional harm to society it is of utmost importance to address the inequalities present in society, especially from an economic perspective. Additionally, by increasing education levels in societies around the US and decreasing poverty rates it is possible to correct the current trend in COVID-19 cases and deaths that the US is headed towards. Most importantly, though, providing high quality medical treatment to all people for free or for a low cost will substantially increase the chances of overcoming the havoc caused by the pandemic in many communities around the US.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

US Department of Agriculture Data. (n.d.). Retrieved from https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/

Ford, T. N., Reber, S., & Reeves, R. V. (2020, June 17). Race gaps in COVID-19 deaths are even bigger than they appear. Retrieved from https://www.brookings.edu/blog/up-front/2020/06/16/race-gaps-in-covid-19-deaths-are-even-bigger-than-they-appear/amp/

Gao, Y., Ding, M., Dong, X., Zhang, J., Azkur, A. K., Azkur, D., . . . Akdis, C. A. (2020, December 04). Risk factors for severe and critically ill COVID 19 patients: A review. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/all.14657

Ludvigsson, Jonas. (2020, June). Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32202343/

Kopel, J., Perisetti, A., Roghani, A., Aziz, M., Gajendran, M., & Goyal, H. (2020, July 28). Racial and Gender-Based Differences in COVID-19. Retrieved from https://www.frontiersin.org/articles/10.3389/fpubh.2020.00418/full

Magesh, S. (2021, November 11). Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic
Status. Retrieved from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2785980

Monita Karmakar, P. (2021, January 29). Association of Sociodemographic Factors With COVID-19 Incidence and
Death Rates in the US. Retrieved from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2775732

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A., Azman, A. S., Paireau, J., . . . Salje, H. (2020,
November 02). Age-specific mortality and immunity patterns of SARS-CoV-2. Retrieved from https://www.nature.com/articles/s41586-020-2918-0

Roy, S., & Ghosh, P. (n.d.). Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. Retrieved from
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241165

World Health Organization. (n.d.). WHO Coronavirus (COVID-19) Dashboard. World Health Organization. Retrieved from https://covid19.who.int/

Monod, Melodie, Blenkinsop, Alexandra, Xi, X., Hebert, D., Bershan, S., Tietze, S., . . . Ratmann, Oliver*, (2021, February 02). Age groups that sustain resurging COVID-19 epidemics in the United States. Retrieved from https://www.science.org/doi/10.1126/science.abe8372

T.M. Therneau. A short introduction to recursive partitioning. Orion Technical Report 21, Stanford University, Department of Statistics, 1983. Retrieved from https://www.mayo.edu/research/documents/biostat-61pdf/doc-10026699

T.M Therneau and E.J Atkinson. An introduction to recursive partitioning using the rpart routines. Division of Biostatistics 61, Mayo Clinic, 1997.Retrieved from https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Link, B. G., and Phelan, J. Social conditions as fundamental causes of disease. J. Health Sociol. Behav. 80–94. doi: 10.2307/2626958, 1995. Retrieved from https://www.jstor.org/stable/2626958?origin=crossref