Team 154 Final Report

Emilee Nathan, Renee Yaldoo, Songeun Lee, Jaedyn Kwon, Thomas Han, Rohan Smith

Introduction

The motivation behind Team 154's analysis is that the COVID-19 pandemic has affected everyone in the world, and thus any insight into relationships between COVID-19 and the world can be used to benefit everyone.

Problem Definition

Team 154 aims to determine if weather change and/or social determinants of health have a combined impact on COVID-19 and its effects in the United States.

Survey

Literature reviewed pertains to extreme weather, different weather conditions and social determinants of health in relation to COVID-19. Extreme climate conditions (such as hurricanes and droughts) as well as conditions created by such conditions (floods, heavy rain, wildfires, heat waves, and power outages) have created environments that go against COVID-19 prevention strategies such as social distancing and medical support supply [13, 15]. As a result of these extreme weather conditions, the movement of people and the need for public transport evacuations have created easier situations for the virus to spread amongst the people in the affected locations [14]. Some of the conducted analysis looked at meteorological weather factors including temperature, humidity, and air pollution [1, 3, 8]. Most analyses had a negative correlation between temperature and the number of COVID-19 cases [3, 8]. Other studies that used differing variables such as seasonality, air pollution, etc. demonstrated a positive correlation to the number of cases [3, 7, 8]. However, some studies concluded that a direct correlation between extreme climates and COVID-19 rates is not noticeably present, although there is a blatant indirect correlation between the two [1, 15]. Conducting such an analysis runs the risk of inconclusive direct results, and this risk applies to Team 154's planned analysis as well.

The COVID-19 pandemic has affected groups that face discrimination and historical injustices hardest [4]. Things like occupation, class, ethnicity, race, citizenship status, and gender continued to undermine health throughout the pandemic [4]. Analysis conducted determined that mortality rates among minority groups is greater than the rest of the population [5, 18]. Education level, poverty rate, and income are also associated with both COVID-19 cases and fatalities [6, 18, 16]. One literature review uses data from the United States [6] which is useful for our project as our data set is also from the United States. Some of the analyses looked at different COVID-19 dependent socio/economic factors including food insecurity, job lay off rates, delay in care, and unemployment rates in relation to educational and racial disparities [16, 17].

Many current studies fail to consider individual, social, and economic factors when considering weather versus COVID-19 conditions [9]. This will be useful for our analysis because we plan to use both socio/economic factors and weather factors in our model. Although, it is important to note that one study found that weather variables are more relevant in predicting the mortality rate when compared to the other variables such as population, age, and urbanization [7].

Proposed Method

Team 154's proposed method consists of two multiple linear regression models and an interactive Tableau dashboard that allows users to explore the dataset and visualize the regression models. This method is novel because we are combining factors related to weather and different social determinants of health to see if they have a relation to COVID-19 when combined. We are also including multiple visual components to the analysis that will allow users truly understand the data and the relationships modeled with multiple linear regression.

Before we dove into the main analysis, Team 154 did a quick exploratory data analysis to view the data in a clearer way. The correlation plot seen in Figure 1 was generated by using R and aimed to discover the top 10 coefficients out of the 199 total attributes in the data set. Another exploration we did was a simple bar chart comparing variable values by state. This visualization was made in Tableau and is fully interactive meaning the user can select different variables to compare by states. The default variables our team put was cases and deaths. Please refer to Figure 2 for this chart.

The multiple linear regression models investigate the relationship between the number of COVID-19 cases and the number of COVID-19 deaths and varying weather and socio/economic factors. Initial data preparation and cleaning was conducted to ensure each variable was in the correct format and attributes that were not directly relevant for the regression (latitude, longitude, whether a stay-at-home order was announced, etc.) were dropped. Many of the remaining independent variables in the data set contained missing values. Therefore, we decided to use KNN to impute missing data with five neighbors (k = 5) to perform a more meaningful analysis with more data points.

The "cases" and "deaths" columns in the original data set were going to be used as the response variables in our linear regression models, but these values are cumulative, meaning each value for "cases" and "deaths" in the data set depends on the values from the days before it. These cumulative sums could not be used in our models because linear regression assumes that observations are independent of one another, and by definition, cumulative sums are not. To remedy this, Team 154 grouped the data by "county" and "state" and ordered the observations by "date" to calculate the differential in COVID-19 cases and deaths for that day in that county and state. These calculated differentials were called "case_diffs" and "death_diffs".

The two models were developed, one with a response variable "case_diffs" and the other with a response variable "death_diffs". Initially, Team 154 tried using backward stepwise regression for variable selection in both our models. However, doing so took a very long time (~ 3 days) to run and there was minimal difference seen in the resulting R-squared of the models compared to their respective models that use all predictor variables. Therefore, Team 154 decided to use all the independent variables after the data cleaning/preparation and imputation in our final two models. Team 154 excluded columns such as "date", "county" and "state" from the final models. These variables were kept out because we wanted to focus on the weather and social determinants of health-related factors. These attributes were the focus of the analysis. The original "deaths" and "cases" columns were also excluded because we were using the non-cumulative data as response variables. The "case_diffs" variable in the model for COVID-19 deaths and the "death_diffs" variable in the model for COVID-19 cases were ignored to focus on the relationships between weather and socio-economic factors on COVID-19 related cases and deaths separately and not their influence on each other.

To evaluate our two models, Team 154 used k-fold cross validation with five folds (k = 5). We decided to use cross validation because it makes better use of the data, and it gives us more information about the performance of our multiple linear regression models.

To display the analysis performed by Team 154, we created multiple Tableau dashboards that contain visualizations to help the user better understand the data set, the models used and ultimately the relationship between COVID-19 and certain social determinants of health and weather-related factors. The first chart visualizes the data set as a map of the United States. It shows the values of the different predictors and response variables per state. This visualization is interactive and allows users to select a variable and see the spread and magnitude of the data values across the US. The user can manually change the month for which data is displayed or press the play button to automate moving through the months included in the data and get a time-lapse feel for the data. Please refer to Figure 3 in the Experiments/Evaluation section below for this visual.

To display the multiple linear regression models, Team 154 created four plots showing the relationship between an independent variable versus the model's response variable. These plots can be seen in Figure 4 below. There are two plots per linear regression model: one showing the actual response variable values versus the predictor variable and the other showing the predicted value from the model versus the predictor variable. The data is grouped by state and the average value of the independent attribute is displayed. This visualization is also interactive as the user can select the feature they wish to see in relation to COVID-19 cases or deaths through the corresponding multiple linear regression model. This will allow users to see how each of the factors in the model correlate to COVID-19 cases and deaths and how closely our models fit the truth data. When a user hovers their mouse over or selects one of the circles representing a state on a plot, the details for the same state will display on the other plots in the visual as well. It is important to note that the "case_diffs" and "death_diffs" are also used in these visualizations and the Tableau dashboard runs the multiple linear regression models in R through Rserve.

Experiments/Evaluation

The experiments in this analysis are aimed at answering questions related to COVID-19 cases and deaths and their association with different weather-related factors and social determinants of health. Questions such as "what weather and socio-economic factors are most highly correlated with COVID-19 cases or deaths?" and "how much do certain factors affect COVID-19 cases or deaths?". We also wanted to be able to visualize the magnitude of COVID-19 cases and deaths in addition to the different socio-economic and weather-related factors that may affect it across the United States.

The complete data set is composed of 773,676 rows and 199 columns after KNN imputation and the creation of the difference variables to use non-cumulative case and death counts.

Through the EDA, we have found examples of factor pairs that have high levels of correlation, such as average number of physically unhealthy days and percent frequency of mental distress as well as severe house cost burden and percent severe housing problems. Some insignificant correlations were rain and percent population overcrowding. Please see Figure 1 below.

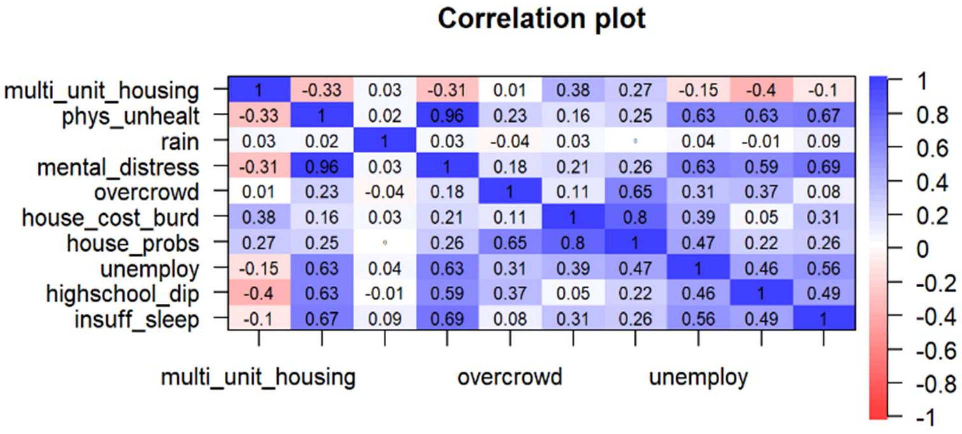


Figure 1: Correlation Plot of Top 10 Coefficients

For EDA purposes, we kept the original, cumulative “deaths” and “cases” columns and used the maximum function to compare the highest number of deaths and cases per state while adjusting the rest of the variables to average values. Please see Figure 2 below.

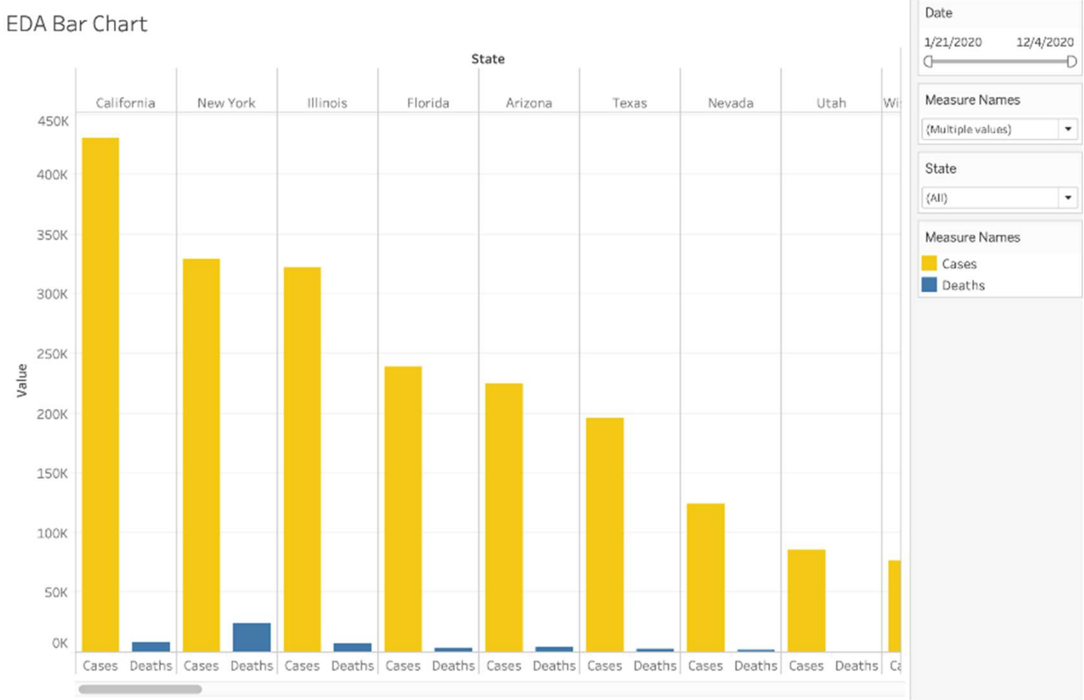


Figure 2: Bar Graph Comparing Variable Values by State

From the 5-fold cross validation results, the model based on COVID-19 cases had a final R-squared value of 0.384, meaning 38.4% of the COVID-19 cases can be explained by the 192 predictor variables provided. The model based on COVID-19 deaths had an R-squared value of 0.188 and thus 18.8% of COVID-19 deaths can be explained by the 192 predictor variables. Please refer to Figures 5 and 6 in the appendix for the output of both models.

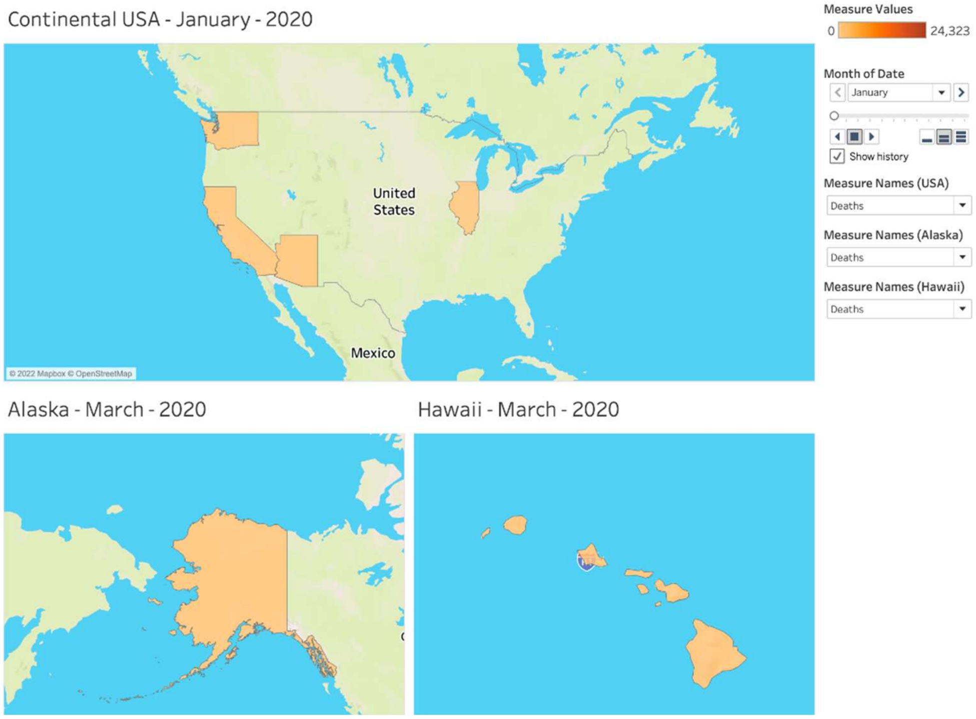
Based on the output of the model regarding COVID-19 cases, some notably significant factors include “hiv_prevalence_rate”, “percent_female”, “mean_temp” and “dewpoint”. The coefficient of “hiv_prevalence_rate” is 0.003078 which tells us that for every unit increase in the HIV prevalence rate, the number of cases increases by 0.003078 on average. Therefore, the prevalence of HIV is positively correlated with COVID-19 cases. Conversely, the coefficient of “percent_female” is -0.3806. This means that for every unit increase in the percentage of females, the number of COVID-19 cases decreases by 0.3806 on average, which tells us being female is negatively correlated with COVID-19 cases.

Regarding weather-related factors, the coefficient of “mean_temp” is -0.4577 and thus every unit increase in average temperature corresponds to a decrease in 0.4577 COVID-19 cases on average. Similarly, “dewpoint” has a coefficient of -0.1079 which tells us for every unit increase in dewpoint, the number of COVID-19 cases decreases by 0.1079 on average. Hence, average temperature and dewpoint are both inversely related to COVID-19 cases.

As for the model for COVID-19 deaths, interesting factors include “percent_physically_inactive”, “num_below_poverty” and “num_unemployed_CHR”. Most of the weather-related attributes were not statistically significant in this model. The coefficient of “percent_physically_inactive” is 0.005649 which tells us that for every percent increase in physically inactive people, the number of COVID-19 deaths increases by 0.005649. This shows the positive correlation between physical inactivity and COVID-19 deaths. Likewise, the coefficients for “num_below_poverty” and “num_unemployed_CHR” are 1.555e-05 and 8.139e-05 respectively. Both coefficients indicate an increase in individuals that are unemployed and below the poverty line are at an increased risk of death from COVID-19.

There are many other attributes from the COVID-19 cases and deaths models that are considered statistically significant at a 95% confidence level, but the aforementioned independent variables provide interesting insight to be discussed further in the conclusion.

From the map of the United States Tableau visual, the user can see the automated progression of the selected variable over the months of 2020. Team 154 noted that New York and California had the highest number of COVID-19 related deaths and cases respectively in 2020. An example of what a user will see for this visual is shown below in Figure 3.


Continental USA - January - 2020

Alaska - March - 2020

Hawaii - March - 2020

Measure Values
0 24,323

Month of Date
January

☑ Show history

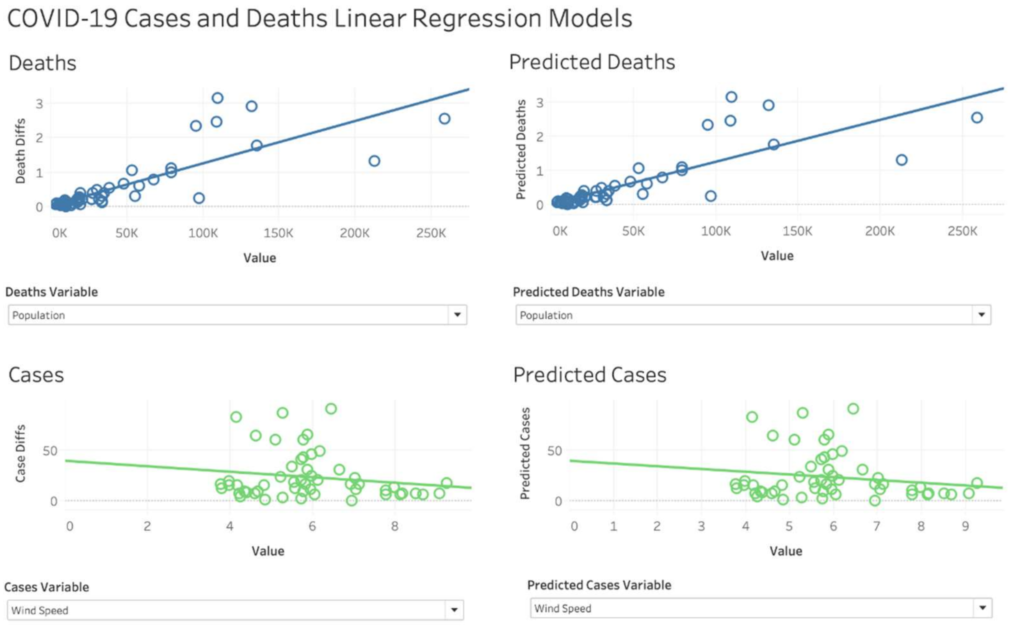
Measure Names (USA)
Deaths

Measure Names (Alaska)
Deaths

Measure Names (Hawaii)
Deaths

Figure 3: USA Map of Variables Throughout 2020

The plots in Figure 4 provide meaningful information regarding the relationship between the different weather and socio-economic factors and COVID-19. The Deaths and Predicted Deaths plots show us that as population increases, the number of COVID-19 deaths and predicted deaths, from our model, also increase on average. This positive correlation makes sense as the more people there are in each state, the more likely infection will spread, and people will die. Similarly, the Cases and Predicted Cases plots tell us that COVID-19 cases decrease on average as wind speeds are higher. The inverse relationship displayed in the plot is also relatively intuitive since all the particles in the air are moving faster and dissipating across the area thus resulting in less spread and ultimately cases of the virus.



Deaths Variable
Population

Predicted Deaths Variable
Population

Cases Variable
Wind Speed

Predicted Cases Variable
Wind Speed

Figure 4: COVID-19 Cases and Deaths Multiple Linear Regression Models

Conclusion/Discussion

The analysis performed by Team 154 lead to the conclusion of the existence of a weak relationship between COVID-19 cases and deaths and most weather-related factors and social determinants of health. This conclusion was drawn by the low R-squared values deduced from the models, however, there is still awareness to be gained from this analysis.

The positive correlation between HIV prevalence rate and COVID-19 cases discovered in our linear regression model reinforces the extra precaution necessary for those that are immunocompromised, as they are at an increased risk of catching the virus. On the other hand, the inverse relationship between females and COVID-19 cases supports existing research that males are more vulnerable to COVID-19. Regarding weather factors, our model determined that warmer weather leads to fewer cases on average and thus individuals should be extra careful in the colder months to avoid getting COVID-19. People below the poverty line and those that are physically inactive have a positive correlation with COVID-19 deaths. This is intuitive as those that are impoverished may not be able to afford the medical care necessary when severe COVID-19 cases arise, thus resulting in death. Similarly, people that are physically inactive are simply not as healthy and can die more easily from COVID-19. It is evident that everyone must do what they can to stay healthy and take the recommended precautions when it comes to the virus especially during the colder months and when the dewpoint is lower.

Even with a weak correlation between COVID-19 and the socio-economic and weather factors in the examined data set, there is insight to be had that can be used to raise awareness of the severity of COVID-19 and do what is possible for those that are at a greater risk.

Emilee and Renee implemented all the Python code with the help of Rohan for the data cleaning and preparation portion. Songeun started the Multiple Linear Regression models implementation, followed with Emilee finishing it up. Jaedyn and Thomas performed the EDA. Renee made the Tableau dashboards with help from Jaedyn, Thomas and Emilee. Songeun created the poster after Renee, Thomas and Jaedyn provided information. Emilee wrote the report with the help of Thomas, Jaedyn and Renee.

References

[1] Heibati, B., Wang, W., Ryti, N. R., Dominici, F., Ducatman, A., Zhang, Z., & Jaakkola, J. J. (2021). Weather conditions and COVID-19 incidence in a cold climate: A Time-series study in Finland. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.605128

[2] Tan, L., & Schultz, D. M. (2022). How is covid-19 affected by weather? metaregression of 158 studies and recommendations for Best Practices in Future Research. *Weather, Climate, and Society*, *14*(1), 237–255. https://doi.org/10.1175/wcas-d-21-0132.1

[3] Paraskevis, D., Kostaki, E. G., Alygizakis, N., Thomaidis, N. S., Cartalis, C., Tsiodras, S., & Dimopoulos, M. A. (2021). A review of the impact of weather and climate variables to COVID-19: In the absence of public health measures high temperatures cannot probably mitigate outbreaks. *Science of The Total Environment*, *768*, 144578. https://doi.org/10.1016/j.scitotenv.2020.144578

[4] Paremoer, L., Nandi, S., Serag, H., & Baum, F. (2021). Covid-19 pandemic and the Social Determinants of Health. *BMJ*. https://doi.org/10.1136/bmj.n129

[5] Badalov, E., Blackler, L., Scharf, A. E., Matsoukas, K., Chawla, S., Voigt, L. P., & Kuflik, A. (2022). Covid-19 double jeopardy: The overwhelming impact of the social determinants of health. *International Journal for Equity in Health*, *21*(1). https://doi.org/10.1186/s12939-022-01629-0

[6] Hawkins, R. B., Charles, E. J., & Mehaffey, J. H. (2020). Socio-economic status and covid-19–related cases and fatalities. *Public Health*, *189*, 129–134. https://doi.org/10.1016/j.puhe.2020.09.016

[7] Malki, Z., Atlam, E.-S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons & Fractals*, *138*, 110137. https://doi.org/10.1016/j.chaos.2020.110137

[8] McClymont, H., & Hu, W. (2021). Weather variability and COVID-19 transmission: A review of recent research. *International Journal of Environmental Research and Public Health*, *18*(2), 396. https://doi.org/10.3390/ijerph18020396

[9] Lin, R., Wang, X., & Huang, J. (2022). The influence of weather conditions on the COVID-19 epidemic. *Environmental Research*, *206*, 112272. https://doi.org/10.1016/j.envres.2021.112272

[10] Adiga, A., Dubhashi, D., Lewis, B., Marathe, M., Venkatramanan, S., & Vullikanti, A. (2020). Mathematical models for covid-19 pandemic: A comparative analysis. *Journal of the Indian Institute of Science*, *100*(4), 793–807. https://doi.org/10.1007/s41745-020-00200-6

[11] Murray, C. J. L. (2020). Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area Countries. *medRxiv*. https://doi.org/10.1101/2020.04.21.20074732

[12] Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S. (2020). The COVID-19 pandemic. *Crit Rev Clin Lab Sci*., 57(6), 365-388. https://doi.org/10.1080/10408363.2020.1783198

[13] Salas, R. N., Shultz, J. M., & Solomon, C. G. (2020). The climate crisis and covid-19 — a major threat to the pandemic response. *New England Journal of Medicine*, *383*(11). https://doi.org/10.1056/nejmp2022011

[14] Thalheimer, L. (2022). Compound impacts of extreme weather events and covid -19 on Climate mobilities. *Area*. https://doi.org/10.1111/area.12821

[15] Walton, D., Arrighi, J., Aalst, M. van, & Claudet, M. (2021). *The Compound Impact of Extreme Weather Events and COVID-19*.

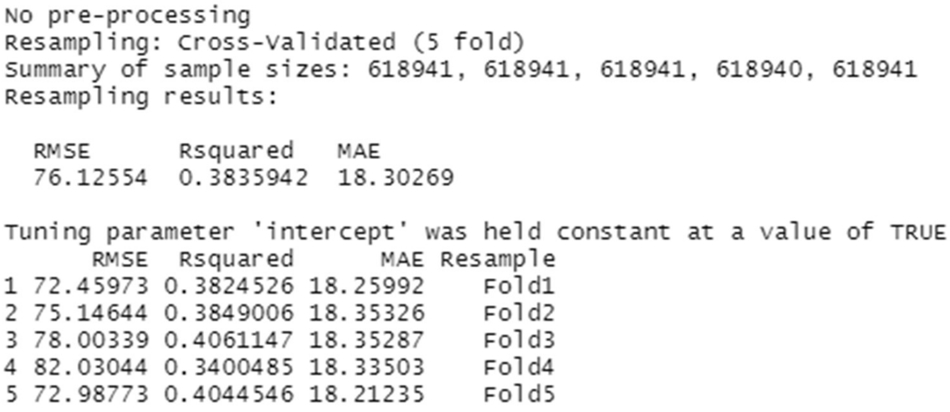
[16] Perry, B. L., Aronson, B., & Pescosolido, B. A. (2021). Pandemic precarity: Covid-19 is exposing and exacerbating inequalities in the American Heartland. *Proceedings of the National Academy of Sciences*, 118(8). https://doi.org/10.1073/pnas.2020685118

[17] Geranios, K., Kagabo, R., & Kim, J. (2022). Impact of covid-19 and socioeconomic status on delayed care and unemployment. *Health Equity*, 6(1), 91–97. https://doi.org/10.1089/heq.2021.0115

[18] Magesh S, John D, Li WT, et al. (2021). Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status: A Systematic Review and Meta-analysis. *JAMA Netw Open*, 4(11). doi:10.1001/jamanetworkopen.2021.34147

[19] Paul, A., Englert, P., & Varga, M. (2021). Socio-economic disparities and covid-19 in the USA. J*ournal of Physics: Complexity*, 2(3), 035017. https://doi.org/10.1088/2632-072x/ac0fc7

Appendix

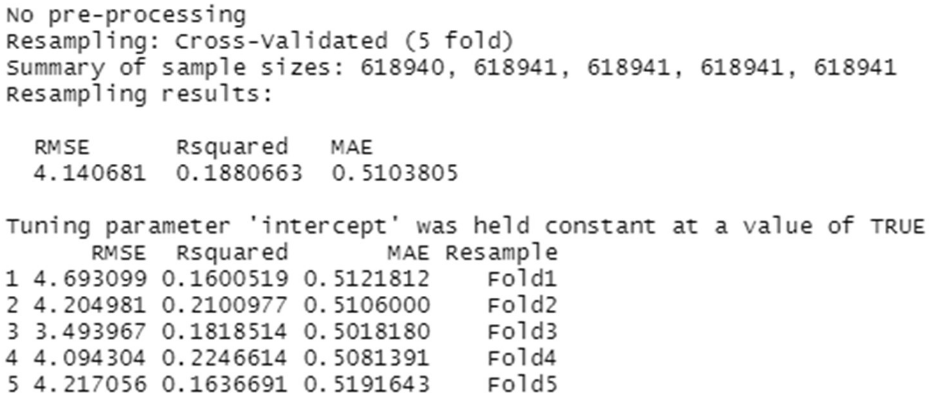


```
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 618941, 618941, 618941, 618940, 618941
Resampling results:

  RMSE      Rsquared   MAE
  76.12554  0.3835942  18.30269

Tuning parameter 'intercept' was held constant at a value of TRUE
      RMSE  Rsquared      MAE Resample
1 72.45973 0.3824526 18.25992    Fold1
2 75.14644 0.3849006 18.35326    Fold2
3 78.00339 0.4061147 18.35287    Fold3
4 82.03044 0.3400485 18.33503    Fold4
5 72.98773 0.4044546 18.21235    Fold5
```

Figure 5: Cases Multiple Linear Regression Output



```
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 618940, 618941, 618941, 618941, 618941
Resampling results:

  RMSE      Rsquared   MAE
  4.140681  0.1880663  0.5103805

Tuning parameter 'intercept' was held constant at a value of TRUE
      RMSE  Rsquared       MAE Resample
1 4.693099 0.1600519 0.5121812    Fold1
2 4.204981 0.2100977 0.5106000    Fold2
3 3.493967 0.1818514 0.5018180    Fold3
4 4.094304 0.2246614 0.5081391    Fold4
5 4.217056 0.1636691 0.5191643    Fold5
```

Figure 6: Deaths Multiple Linear Regression Output