

Assignment 4 Writeup

Abstract and Introduction (2%):

While drafting ideas for this project and the presentation for assignment 3, my first thought went to datasets involving COVID-19. Since the pandemic began back in 2020, the subject has been of great interest to me due to the fact that despite there being many studies on the topic and time period, people still disagree on COVID-19 and its effects on people and greater society. Back home, it felt as though almost no one believed that COVID-19 was nearly as dangerous as the news claimed. Since I had never met anyone who suffered extreme symptoms or death due to COVID-19, I believed this to be true as well. However, after coming to RPI, I found that the answer was not as straightforward as I once thought. Regardless, the one thing that has been clear to me since the beginning is that the pandemic caused widespread panic and isolation through the subsequent quarantine policies put in place. This led me to my hypothesis: U.S. states with higher COVID-19 mortality rates in a given week saw an increase in mental health care. Additionally, severity of mental health care can be predicted based on the local COVID-19 mortality rate.

Data Description and preliminary analysis (3%)

To prove this hypothesis, I turned to the place I thought best to find data and other statistics about COVID-19 and other diseases, the Center for Disease Control (CDC). After scouring the CDC website for suitable datasets, I stumbled upon the COVID-19 dataset (Provisional Covid-19 Death Counts by Week Ending Date and State). This dataset was key for my research as it separated COVID-19 deaths from those caused by other respiratory diseases, such as Influenza and Pneumonia. Additionally, it separated this information down to the week and categorized it by state. Next, I found the Mental Health dataset (Mental Health Care in the Last 4 Weeks), which took mental health statistics from the same time period and categorized them by week and state. These datasets would be compatible since they could be connected through the state, start date, and end date columns, which detailed the state in which the data was collected from and the start and end dates of the week it was collected in.

After stepping away from the project for about a month to focus on other projects, I came back and took a deeper look at the 2 datasets. First, I found that the start and end dates did not match up between datasets. In fact, they didn't match up *within* datasets, as some timeframes were as little as 1 week and sometimes as large as 3. Furthermore, many of these time frames were not full weeks, which meant that there would need to be significant data cleaning and organization performed to combine the datasets. To find a solution, I looked at the Household Pulse Survey (Household Pulse Survey Data Tables), which is a 20 minute long online survey published by the U.S. Census Bureau that measures how emergent social and economic issues impacted households across the country. Being the source of the data from the Mental Health dataset, I tried to find alternate time tables or an explanation for why the intervals were uneven, but found nothing. Instead, I found that many of the features involved were useless for my endeavors. This meant that the only reliable features from the dataset were the "Indicator" [for mental health] and death totals. Preliminary analysis showed that the "Group" and "Subgroup" features had a large variety of different possible points. However, they turned out to be useless as about 2/3 of all rows had "By State" as the value for the "Group" value and the state name as

the “Subgroup” value. For these rows, “Group” and “Subgroup” would act as noise since the state is already found in the “State” feature. Furthermore, the other 1/3 states, regardless of their “Group” or “Subgroup” values, listed “United States” as the “State” value, meaning that the death totals would be heavily inflated compared to where the survey participants actually lived.

| | Group | Subgroup | State | count |
|----|---|---|----------------------|-------|
| 0 | By Age | 18 - 29 years | United States | 152 |
| 1 | By Age | 30 - 39 years | United States | 152 |
| 2 | By Age | 40 - 49 years | United States | 152 |
| 3 | By Age | 50 - 59 years | United States | 152 |
| 4 | By Age | 60 - 69 years | United States | 152 |
| 5 | By Age | 70 - 79 years | United States | 152 |
| 6 | By Age | 80 years and above | United States | 152 |
| 7 | By Disability status | With disability | United States | 84 |
| 8 | By Disability status | Without disability | United States | 84 |
| 9 | By Education | Bachelor's degree or higher | United States | 152 |
| 10 | By Education | High school diploma or GED | United States | 152 |
| 11 | By Education | Less than a high school diploma | United States | 152 |
| 12 | By Education | Some college/Associate's degree | United States | 152 |
| 13 | By Gender identity | Cis-gender female | United States | 52 |
| 14 | By Gender identity | Cis-gender male | United States | 52 |
| 15 | By Gender identity | Transgender | United States | 52 |
| 16 | By Presence of Symptoms of Anxiety/Depression | Did not experience symptoms of anxiety/depression in the past 4 weeks | United States | 152 |
| 17 | By Presence of Symptoms of Anxiety/Depression | Experienced symptoms of anxiety/depression in past 4 weeks | United States | 152 |
| 18 | By Race/Hispanic ethnicity | Hispanic or latino | United States | 152 |
| 19 | By Race/Hispanic ethnicity | Non-Hispanic Asian, single race | United States | 152 |
| 20 | By Race/Hispanic ethnicity | Non-Hispanic Black, single race | United States | 152 |
| 21 | By Race/Hispanic ethnicity | Non-Hispanic White, single race | United States | 152 |
| 22 | By Race/Hispanic ethnicity | Non-Hispanic, other races and multiple races | United States | 152 |
| 23 | By Sex | Female | United States | 152 |
| 24 | By Sex | Male | United States | 152 |
| 25 | By Sexual orientation | Bisexual | United States | 52 |
| 26 | By Sexual orientation | Gay or lesbian | United States | 52 |
| 27 | By Sexual orientation | Straight | United States | 52 |
| 28 | By State | Alabama | Alabama | 132 |
| 29 | By State | Alaska | Alaska | 132 |
| 30 | By State | Arizona | Arizona | 132 |
| 31 | By State | Arkansas | Arkansas | 132 |
| 32 | By State | California | California | 132 |
| 33 | By State | Colorado | Colorado | 132 |
| 34 | By State | Connecticut | Connecticut | 132 |
| 35 | By State | Delaware | Delaware | 132 |
| 36 | By State | District of Columbia | District of Columbia | 132 |
| 37 | By State | Florida | Florida | 132 |
| 38 | By State | Georgia | Georgia | 132 |
| 39 | By State | Hawaii | Hawaii | 132 |
| 40 | By State | Idaho | Idaho | 132 |
| 41 | By State | Illinois | Illinois | 132 |
| 42 | By State | Indiana | Indiana | 132 |
| 43 | By State | Iowa | Iowa | 132 |
| 44 | By State | Kansas | Kansas | 132 |
| 45 | By State | Kentucky | Kentucky | 132 |
| 46 | By State | Louisiana | Louisiana | 132 |
| 47 | By State | Maine | Maine | 132 |
| 48 | By State | Maryland | Maryland | 132 |
| 49 | By State | Massachusetts | Massachusetts | 132 |
| 50 | By State | Michigan | Michigan | 132 |
| 51 | By State | Minnesota | Minnesota | 132 |
| 52 | By State | Mississippi | Mississippi | 132 |
| 53 | By State | Missouri | Missouri | 132 |
| 54 | By State | Montana | Montana | 132 |
| 55 | By State | Nebraska | Nebraska | 132 |
| 56 | By State | Nevada | Nevada | 132 |
| 57 | By State | New Hampshire | New Hampshire | 132 |
| 58 | By State | New Jersey | New Jersey | 132 |
| 59 | By State | New Mexico | New Mexico | 132 |
| 60 | By State | New York | New York | 132 |
| 61 | By State | North Carolina | North Carolina | 132 |
| 62 | By State | North Dakota | North Dakota | 132 |
| 63 | By State | Ohio | Ohio | 132 |
| 64 | By State | Oklahoma | Oklahoma | 132 |
| 65 | By State | Oregon | Oregon | 132 |
| 66 | By State | Pennsylvania | Pennsylvania | 132 |
| 67 | By State | Rhode Island | Rhode Island | 132 |
| 68 | By State | South Carolina | South Carolina | 132 |
| 69 | By State | South Dakota | South Dakota | 132 |
| 70 | By State | Tennessee | Tennessee | 132 |
| 71 | By State | Texas | Texas | 132 |
| 72 | By State | Utah | Utah | 132 |
| 73 | By State | Vermont | Vermont | 132 |
| 74 | By State | Virginia | Virginia | 132 |
| 75 | By State | Washington | Washington | 132 |
| 76 | By State | West Virginia | West Virginia | 132 |
| 77 | By State | Wisconsin | Wisconsin | 132 |
| 78 | By State | Wyoming | Wyoming | 132 |
| 79 | National Estimate | United States | United States | 152 |

Figure 1. Unique combinations of “Group” and “Subgroup” values placed alongside “State” values and the number of times the combinations occurred in the data

Exploratory Analysis (5%):

Before attempting to combine the datasets, I needed to clean and organize the data to ensure that it would be compatible and only have useful information. First, I removed all unnecessary columns from the original datasets, ensuring that the only columns remaining were “Indicator”, “State”, “Time Period Start Date”, “Time Period End Date” for the Mental Health dataset and “Start Date”, “End Date”, “State”, “COVID-19 Deaths”, and “Total Deaths” for the COVID-19 dataset. “Indicator” refers to the mental health treatment that was or wasn’t taken by the participant, the start and end dates refer to the dates between which the survey was taken, “State” refers to the state that the participant lived in, “COVID-19 Deaths” refers to the number of deaths attributed to COVID-19 occurred between the start and end dates, and “Total Deaths” refers to the number of deaths attributed to COVID-19 and other respiratory diseases during that time frame. Next, I removed all rows with NA values within the data to ensure that the dataset was complete. Finally, I converted all dates to datetime objects to ensure that they would be compatible regardless of the date taking notation used.

Next, I created a new dailyCovid dataset which stores information from the COVID-19 dataset but on a day to day basis. For each row in the COVID-19 dataset, I found the inclusive range between the start and end dates and took the mean number of “COVID-19 Deaths” and “Total Deaths” in said range. Then, I populated the dailyCovid dataset with values for the “State”, “Date”, “COVID-19 Deaths” and “Total Deaths” columns. The state from the COVID-19 dataset would be stored in “State”, each date from a date range found in the COVID-19 dataset would be stored in “Date”, and the mean number of “COVID-19 Deaths” and “Total Deaths” between dates in the date range would be stored in similarly named columns.

```
Total rows in dailyCovidData before duplicate removal: 411659  
Total rows in dailyCovidData after duplicate removal: 113022
```

Figure 2. Total rows present in the dailyCovid dataset before and after duplicate removal

This created 411,659 rows. However, since participants from the same state could have filled out the survey during time periods with overlapping dates. Due to recording discrepancies, death totals could be different between these entries. Thus, for every set of rows that had the same “State” and “Date” values, I replaced them with a single row containing the same “State” and “Date” value along with the mean “COVID-19 Deaths” and “Total Deaths” values for each of the rows in the same grouping. This significantly reduced noise in the dataset as the final result was 113,022 rows.

I finished up the data organization by creating a finalData dataset, which took the “State”, “Start Date”, “End Date”, and “Indicator” values from each row in the Mental Health Dataset. Then, I placed the sum of all “COVID-19 Deaths” from rows in the dailyCovid dataset with dates between the “Start Date” and “End Date” values of the Mental Health dataset. This process was also done with “Total Deaths”. Then, I created the “Cumulative COVID-19 Deaths” and “Cumulative Total Deaths” columns, which were populated with the cumulative amount of deaths that occurred from the first date in the dailyCovid dataset until the “End Date” in the Mental Health dataset. Both the Mental Health and finalData (herein referred to as “the data”) datasets had 10,404 rows, which means that the data organization was performed correctly.

```
Total rows in finalData: 10404  
Total rows mentalHealthData: 10404
```

Figure 3. Total rows present in the finalData dataset and Mental Health dataset

| State | |
|----------------------|------|
| United States | 3672 |
| Alabama | 132 |
| Alaska | 132 |
| Arizona | 132 |
| Arkansas | 132 |
| California | 132 |
| Colorado | 132 |
| Connecticut | 132 |
| Delaware | 132 |
| District of Columbia | 132 |
| Florida | 132 |
| Georgia | 132 |
| Hawaii | 132 |
| Idaho | 132 |
| Illinois | 132 |
| Indiana | 132 |
| Iowa | 132 |
| Kansas | 132 |
| Kentucky | 132 |
| Louisiana | 132 |
| Maine | 132 |
| Maryland | 132 |
| Massachusetts | 132 |
| Michigan | 132 |
| Minnesota | 132 |
| Mississippi | 132 |
| Missouri | 132 |
| Montana | 132 |
| Nebraska | 132 |
| Nevada | 132 |
| New Hampshire | 132 |
| New Jersey | 132 |
| New Mexico | 132 |
| New York | 132 |
| North Carolina | 132 |
| North Dakota | 132 |
| Ohio | 132 |
| Oklahoma | 132 |
| Oregon | 132 |
| Pennsylvania | 132 |
| Rhode Island | 132 |
| South Carolina | 132 |
| South Dakota | 132 |
| Tennessee | 132 |
| Texas | 132 |
| Utah | 132 |
| Vermont | 132 |
| Virginia | 132 |
| Washington | 132 |
| West Virginia | 132 |
| Wisconsin | 132 |
| Wyoming | 132 |

Figure 4. Distribution of values in “States” Column of the data

Further indication that the data organization was performed correctly can be found in the breakdown of “State” in the data (Figure 4). “United States” appears 3672 times and all other states appear 132 times, which matches the counts from the Mental Health dataset (Figure 1).

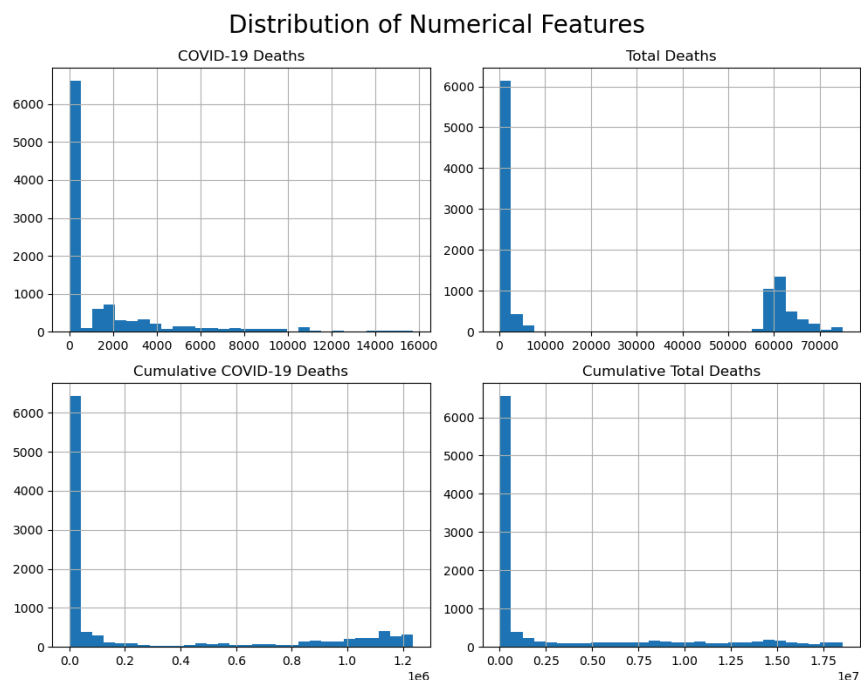


Figure 5. Distribution of Numerical Features in the data via histograms

All distributions of numerical features (Figure 5) follow the same trend of experiencing a high volume of instances with relatively low death counts and a tiny volume of instances for the steadily increasing death totals. Since the data maps late 2019 to late 2025, it can be assumed that the large number of low death counts is due to the amount of survey responses retrieved before and after the COVID-19 pandemic.

The boxplot of the numerical data (Figure 6) shows only “COVID-19 Deaths” and “Cumulative Total Deaths” experienced a large number of high valued outliers. This can be attributed to the large spikes in COVID-19 deaths in 2020 - 2022. Likewise, “Cumulative Total Deaths” would experience outliers in the same time period since COVID-19 deaths are included in these totals. It can be assumed that “Cumulative COVID-19 Deaths” didn’t experience outliers since the number was always steadily increasing or plateauing. Similarly, “Total Deaths” likely experienced similar death counts in smaller time frames over the years.

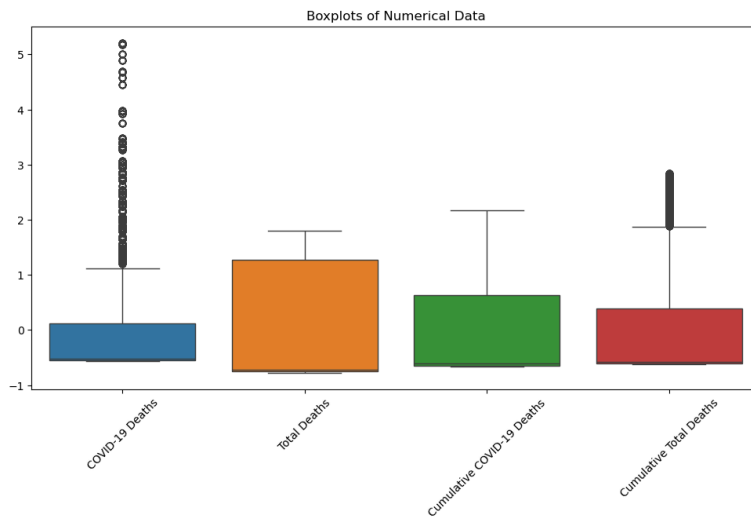


Figure 6. Boxplot of Numerical Features in the data

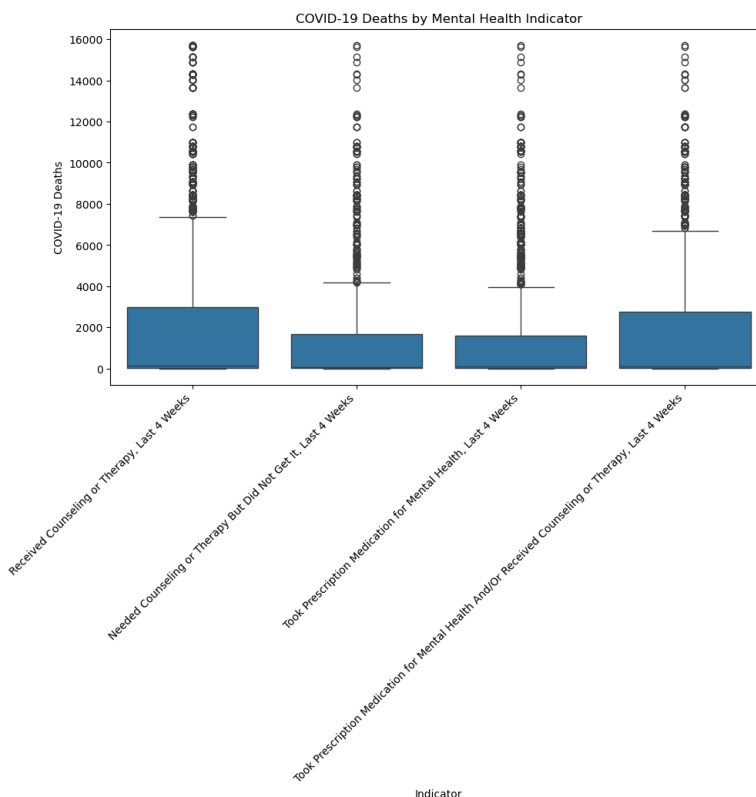


Figure 7. Boxplot of COVID-19 Deaths by Mental Health Indicator in the data

Figure 7 features a boxplot depicting COVID-19 Deaths by Mental Health Indicator found within the data, and it is clear that as the number of COVID-19 deaths rise so does the number of outliers present in mental health indicators.

Figures 8-15 show the timeline of each death statistic over the last 5 years organized by state. To display these figures, rows with "United States" as a state were removed because it vastly overshadowed the statistics of even the largest individual states. Figures 9, 11, 13, and 15 depict the same information as Figures 8, 10, 12, and 14 respectively, except they only display 5 states. Four of the states represent different quantiles (25%, 50%, 75%, and maximum) for the given statistic, and the last state represented is my home state, New Hampshire. I included my home state because while not statistically significant, it illustrates why I originally had such a gap in knowledge about the severity of COVID-19 prior to coming to RPI, since there are very few deaths related to COVID-19 compared to other states. Overall, it can be seen that the states which suffered the most deaths in any category were also those with some of the largest state populations, eluding to the fact that these numbers likely represent a proportional increase rather than any missteps by the state.

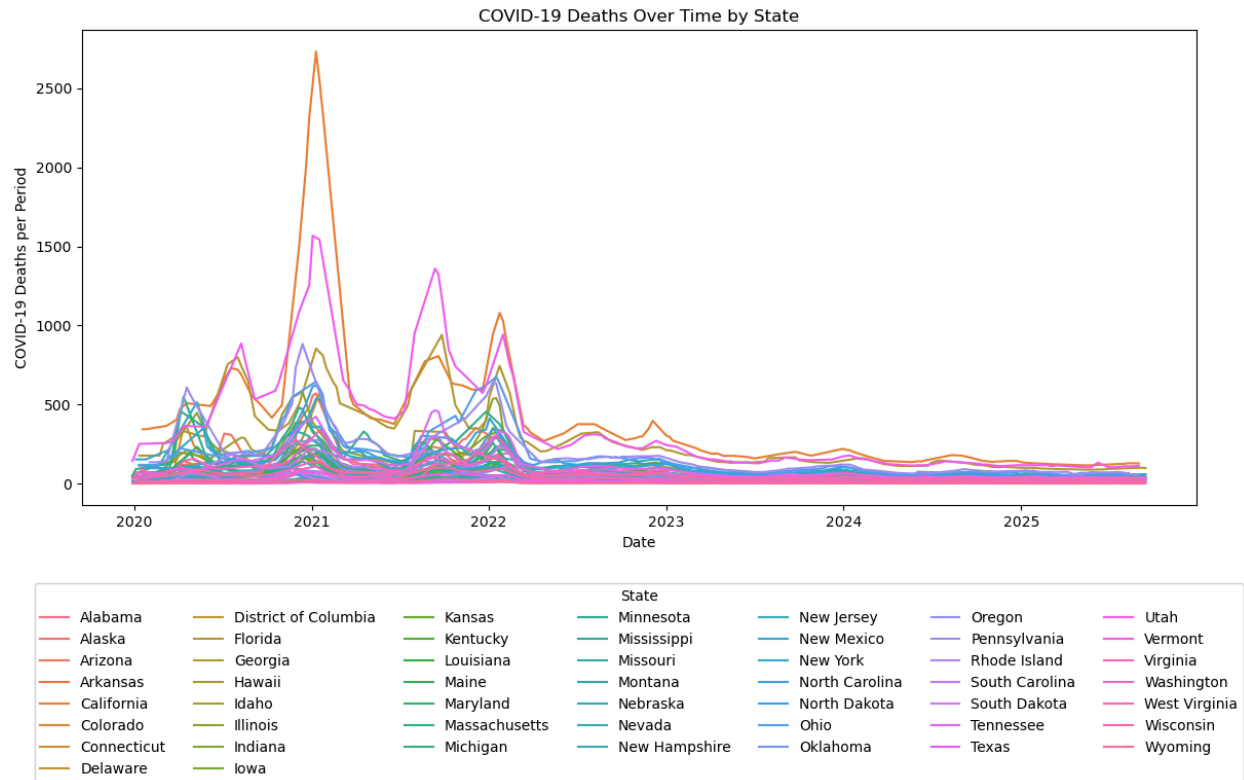


Figure 8. Scatterplot of COVID-19 Deaths over time by state

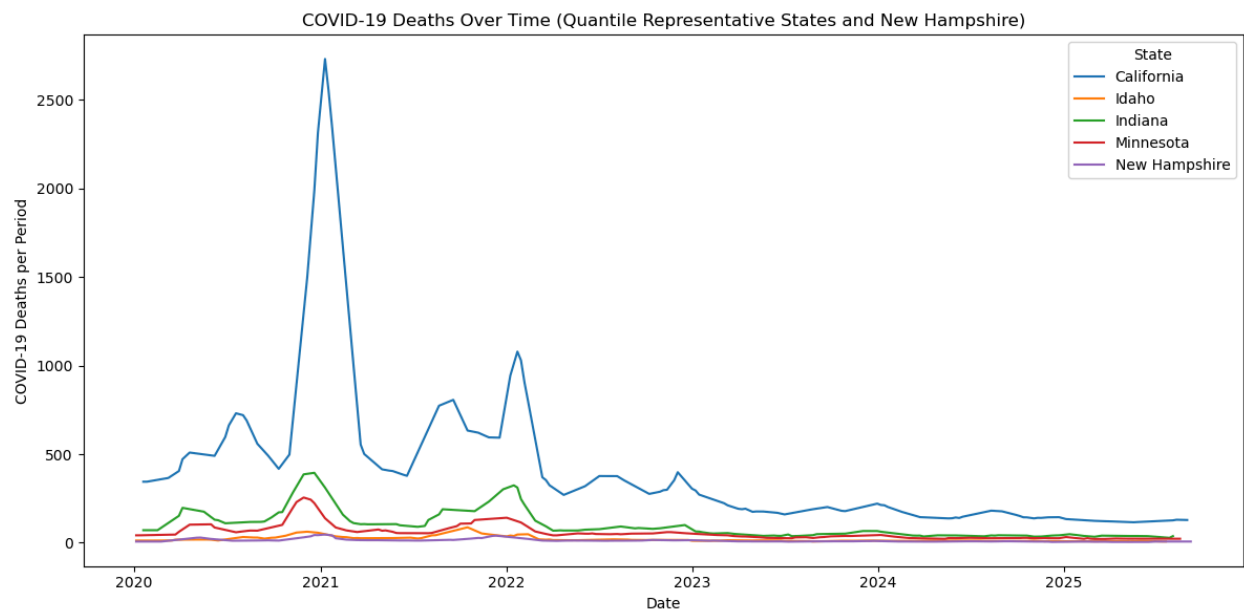


Figure 9. Scatterplot of COVID-19 Deaths over time by quantile representative states and New Hampshire

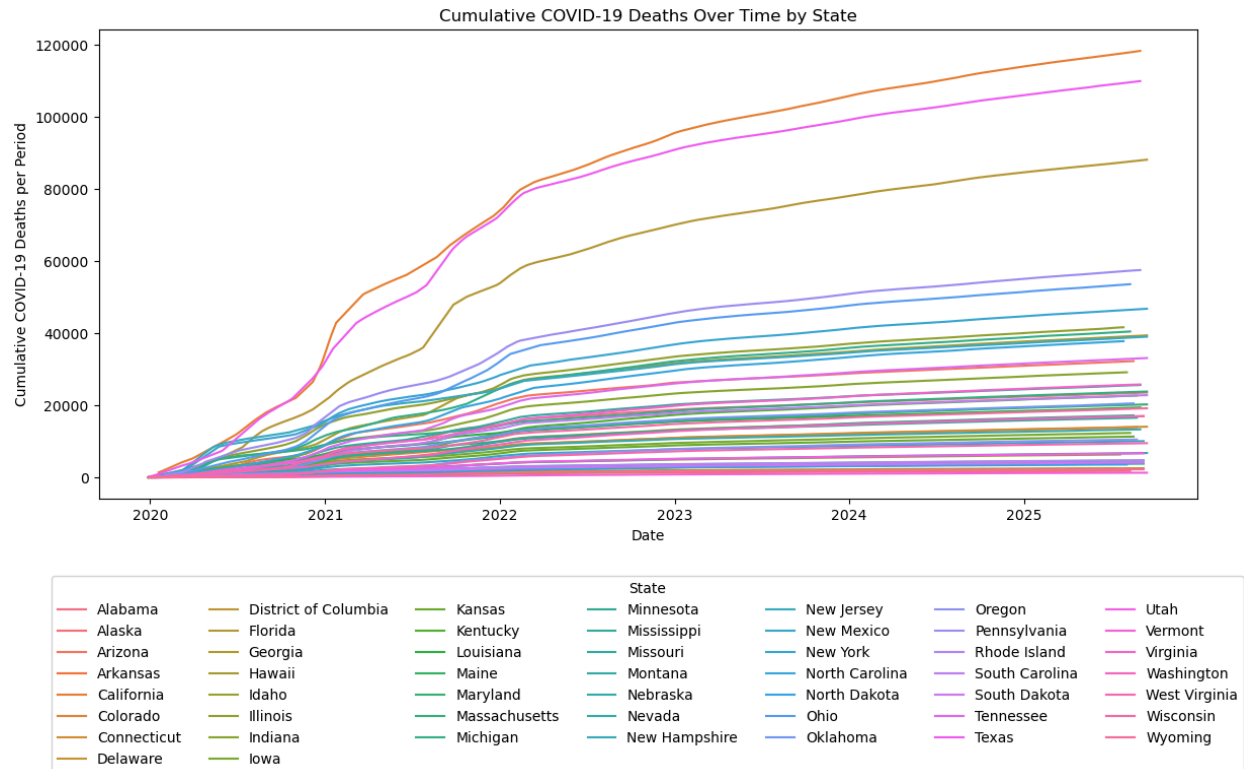


Figure 10. Scatterplot of Cumulative COVID-19 Deaths over time by state

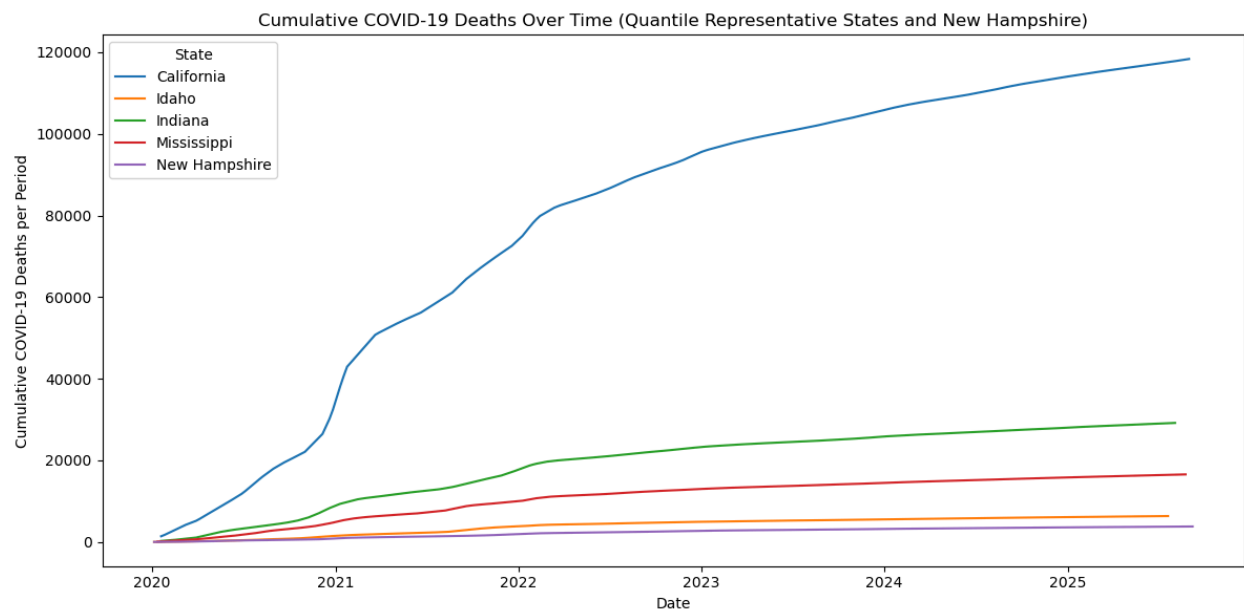


Figure 11. Scatterplot of Cumulative COVID-19 Deaths over time by quantile representative states and New Hampshire

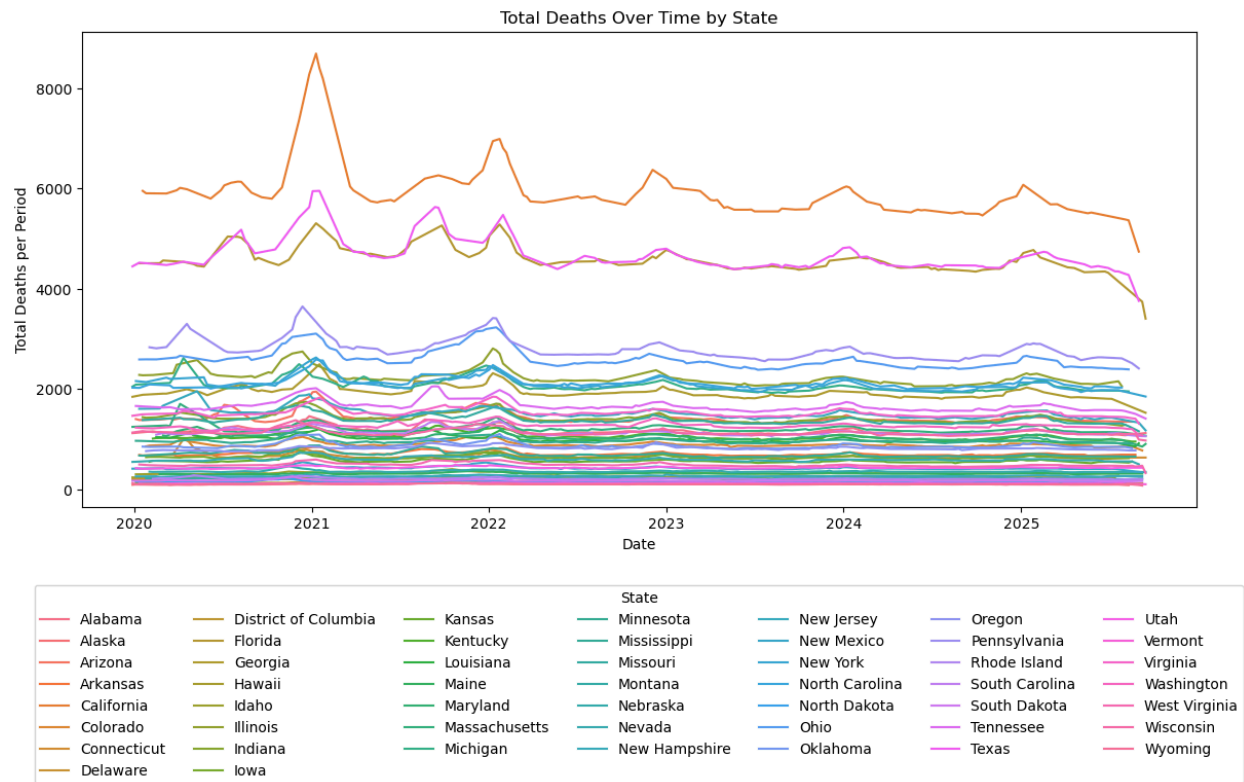


Figure 11. Scatterplot of Total Deaths over time by state

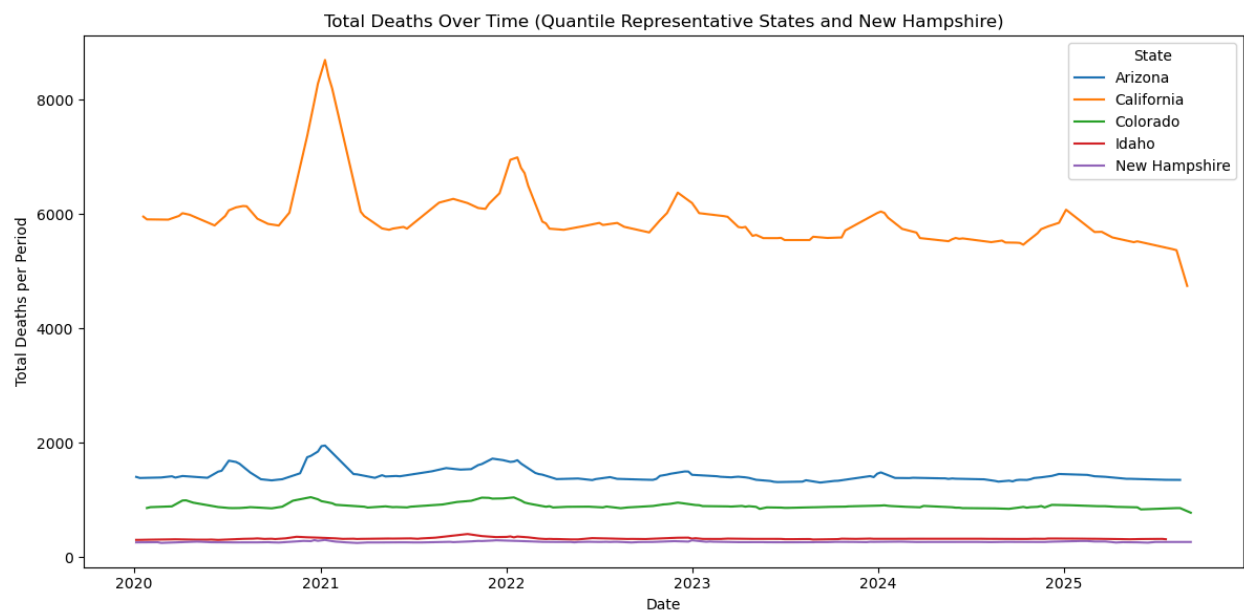


Figure 12. Scatterplot of Total Deaths over time by quantile representative states and New Hampshire

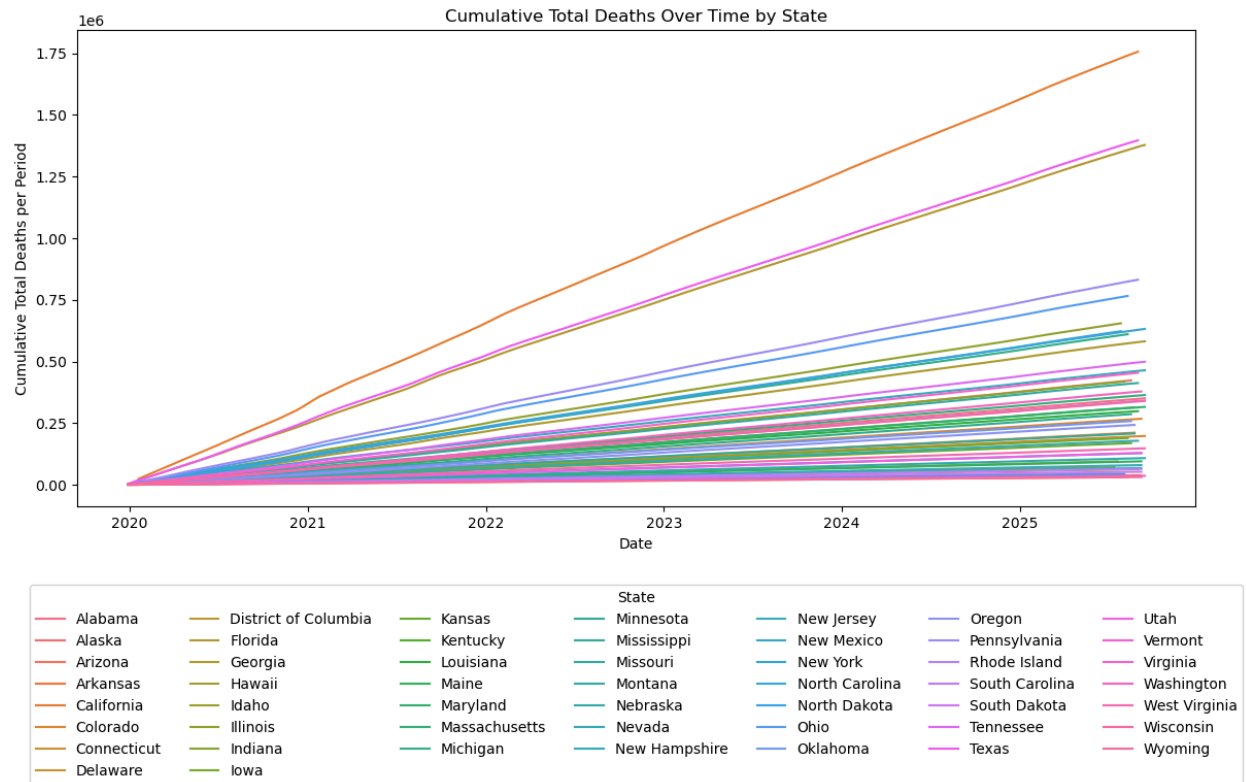


Figure 13. Scatterplot of Cumulative Total Deaths over time by state

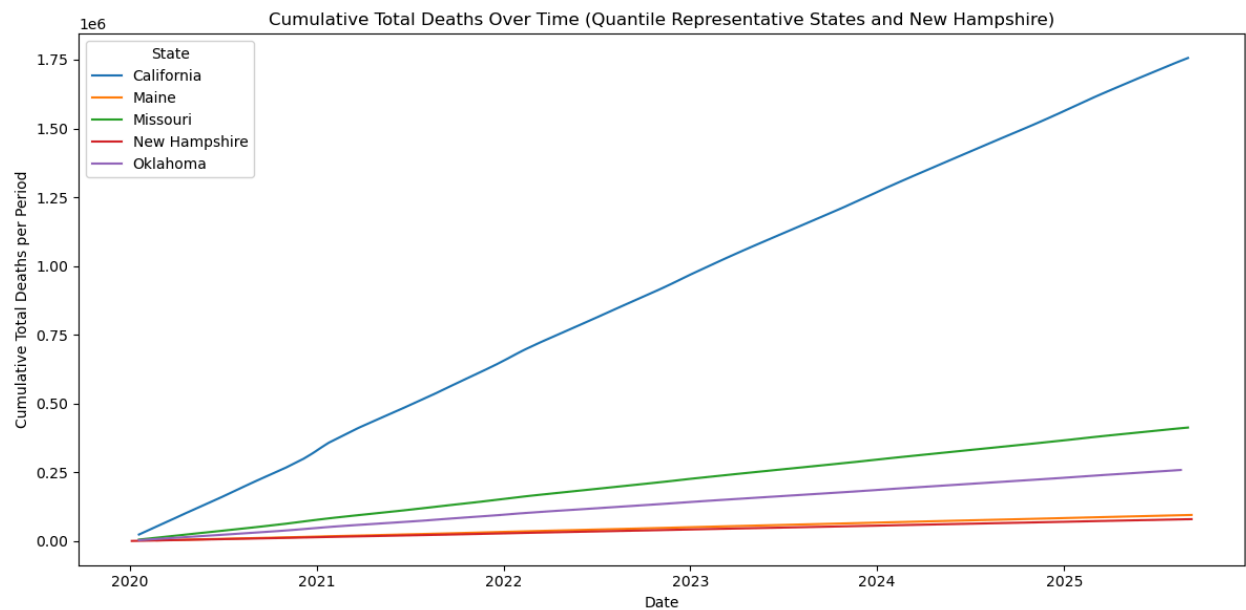


Figure 14. Scatterplot of Cumulative Total Deaths over time by quantile representative states and New Hampshire

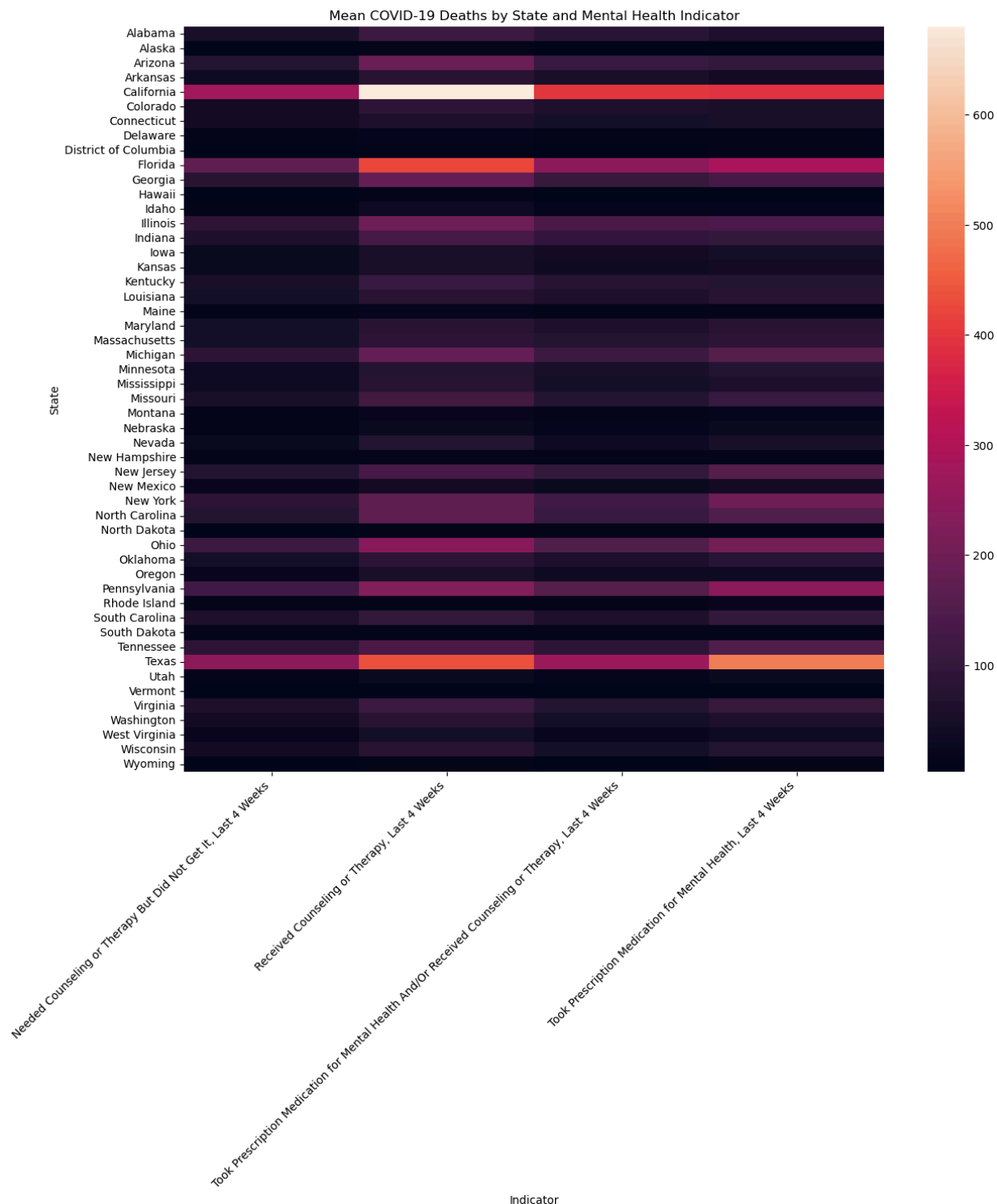


Figure 15. Heatmap of Mean COVID-19 Deaths by State and Mental Health Indicator

Model Development and Application of model(s) (12%):

I chose to implement 5 different types of models. The first four models, Logistic Regression, Random Forest, Classifier, K-Nearest-Neighbors (KNN), and Support Vector Machine (SVM) with a polynomial kernel, are all supervised classifier models with the target variable of "Indicator" [for mental health]. They also use standardized numeric death counts and one-hot encoded categorical variables with an 80/20 stratified train/test split. The fifth model is K-Means Clustering with Principal Component Analysis (PCA) Visualization. For K-Means I used silhouette score and elbow method analysis to find natural groupings and find whether clusters align with the different "Indicator" types. Classification performance was compared using accuracy, precision, recall, and F1 score. Cluster summaries were used to show meaningful distributions of death totals and indicator types within the clusters.

The numeric features include "COVID-19 Deaths", "Total Deaths", "Cumulative COVID-19 Deaths", and "Cumulative Total Deaths" and the categorical features include "State". Anything that could utilize a seed was provided a seed of 42 to ensure reproducibility upon subsequent runs. StandardScaler() was used for numeric features and OneHotEncoder() was used for the categorical features prior to deployment of all models. ColumnTransformer() and Pipeline() were used to ensure that an identical pipeline was created for preprocessing prior to all model deployment. I chose these models because they were all ones that were taught in class, and the kernel used in SVM was chosen due to having the best performance compared to the other kernels.

| | accuracy | precision | \ |
|--|----------|-----------|---|
| Logistic Regression | 0.282 | 0.282 | |
| Logistic Regression (without U.S. as a State) | 0.398 | 0.397 | |
| Random Forest Classifier | 0.437 | 0.437 | |
| Random Forest Classifier (without U.S. as a State) | 0.432 | 0.432 | |
| KNN Classifier | 0.437 | 0.439 | |
| KNN Classifier (without U.S. as a State) | 0.466 | 0.468 | |
| SVM | 0.310 | 0.383 | |
| SVM (without U.S. as a State) | 0.336 | 0.369 | |
| | recall | f1score | |
| Logistic Regression | 0.282 | 0.281 | |
| Logistic Regression (without U.S. as a State) | 0.398 | 0.394 | |
| Random Forest Classifier | 0.437 | 0.437 | |
| Random Forest Classifier (without U.S. as a State) | 0.432 | 0.432 | |
| KNN Classifier | 0.437 | 0.436 | |
| KNN Classifier (without U.S. as a State) | 0.466 | 0.464 | |
| SVM | 0.310 | 0.287 | |
| SVM (without U.S. as a State) | 0.336 | 0.319 | |

Figure 16. Classifier Model Metrics Comparison

Figure 16 shows the accuracy, precision, recall, and F1 score metrics for each of the classification models. Since I removed all rows with "U.S." as a value for "State" while creating plots due to the large difference in scale, I assumed that it might make a large difference in model training too. Thus, I ran 2 versions of each model - one that included "U.S" as a value for "State" and one that did not. When all data was included KNN performed the best in all metrics and achieved about 46.6% accuracy. Thai was followed by Random Forest, SVM polynomial kernel, and Logistic Regression. When rows with "U.S" as the state were dropped, Random Forest and KNN performed equally well, achieving a peak of about 43.7% accuracy. Following

this was Logistic Regression and SVM. Removing “U.S.” as a state improved overall performance in all models except for Random Forest, which suffered a slight dip in performance. Logistic Regression in particular benefited greatly from this change, which makes sense due to less variability in the feature types.

```
-- Logistic Regression
|   0   1   2   3
0  156  83  138 143
1  126 140  131 123
2  137 111  157 116
3  140 142  105 133

-- Logistic Regression (without U.S. as a State)
|   0   1   2   3
0  173  42  65  57
1   52 118 105  61
2   90  67 143  37
3  104  89  42 102
```

Figure 17. Confusion Matrices of Logistic Regression Models

```
-- Random Forest Classifier
|   0   1   2   3
0  229  67  118 106
1   73 225  90 132
2   94 131 232  64
3  127 100  69 224

-- Random Forest Classifier (without U.S. as a State)
|   0   1   2   3
0  147  30  78  82
1   41 157  63  75
2   80  81 130  46
3   68  68  53 148
```

Figure 18. Confusion Matrices of Random Forest Models

```
-- KNN Classifier
|   0   1   2   3
0  254  81  108  77
1   92 238  83 107
2  115 151 200  55
3  129 113  61 217

-- KNN Classifier (without U.S. as a State)
|   0   1   2   3
0  174  40  70  53
1   53 183  43  57
2   72 106 131  28
3   82  67  48 140
```

Figure 19. Confusion Matrices of KNN Models

```
-- SVM
|   0   1   2   3
0  109 315  29  67
1   45 348  18 109
2   36 330  83  72
3   87 312  15 106

-- SVM (without U.S. as a State)
|   0   1   2   3
0   46  50 185  56
1   11  99 153  73
2   18  66 189  64
3   24  80 115 118
```

Figure 20. Confusion Matrices of SVM Models

Figures 17-20 display the confusion matrices between the different classification models. At a glance, the numbers in each confusion matrix look almost evenly distributed through, with a slight increase in the size of the numbers along the diagonal line starting at the top left. This means that while a sizable chunk of the predictions were true positives, the majority of the predictions were incorrect, leading to the below 50% accuracy on all models. A significant amount of miscalculation like this implies that features heavily overlap across categories and the models struggle to separate them. It also means each class is hard to identify, further leading to inseparability.

Despite the improvements present between models, the highest accuracy achieved between all models was only about 46.6%. Therefore, it is safe to say that this hypothesis is not possible given the following data. This is likely due to a lack of relevant features available in the original datasets found through the CDC. As such, the only available features to predict off of

were death totals and state, which may have been more useful when coupled with information like sex, race etc. That being said, even if these potential features did improve model accuracy, one could argue that it defeats the purpose since the hypothesis implied that mental health indicators could be predicted off of COVID-19 death rates alone. Each metric used a weighted average to reflect class imbalances.

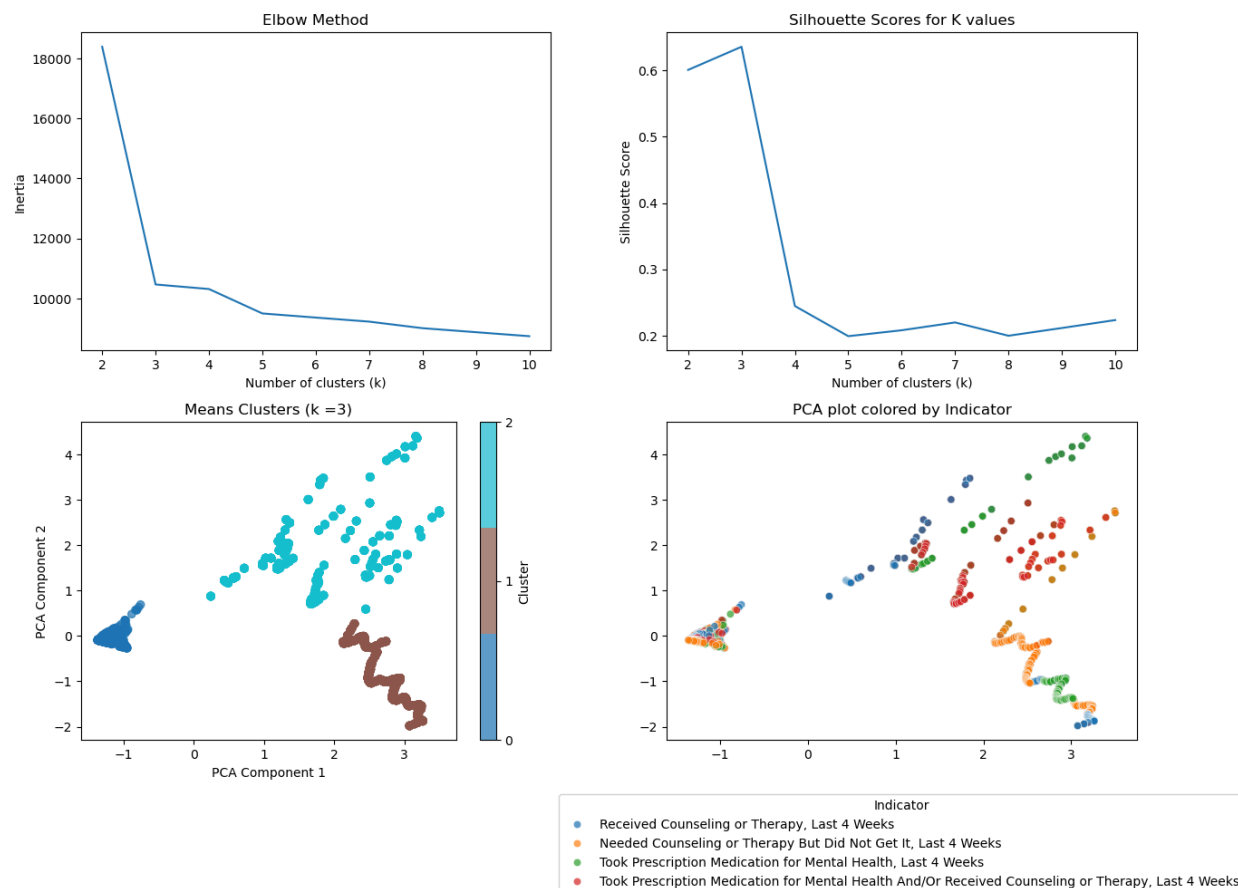


Figure 21. Elbow Method plot for number of K values plot (top left), Silhouette Scores plot for K values plot (top right), Cluster plot via PCA (bottom left), PCA plot colored by Indicator (bottom right) - All data included

The output from the K-Means model was able to shed some light on why the classification models did not perform well. First, I found the inertia values and silhouette scores for K values 2 through 10. Once plotted, it was easy to see that the optimal K value for the model when “U.S” was included as a state was K = 3 due to the elbow in the elbow method plot in Figure 21 and the peak of the silhouette scores plot in Figure 21. Once this was determined, the PCA visualization of the clustering was found in the bottom left plot of Figure 21. The points in all clusters are globally separated, which shows that the features contain different groups and that the projection maintained enough variance to show a meaningful structure. However, the large distance between some points in clusters signifies that the data is very spread out and features have high variability. This is to be expected due to the inclusion of “U.S” as a state, as it effectively creates outliers across the board. The plot on the bottom right of Figure 21 shows the

mental health indicator for each point in the clusters. This plot displays clear groupings but fails to match those in the original PCA visualization plot. It also shows lots of overlap which implies that there was not enough information to differentiate the indicators.

| Cluster | COVID-19 Deaths | Total Deaths | Cumulative COVID-19 Deaths |
|---------|-----------------|--------------|----------------------------|
| 0 | 79.687813 | 1194.495502 | 1.571984e+04 |
| 1 | 2212.433174 | 60017.413079 | 1.068219e+06 |
| 2 | 7160.340143 | 65026.941027 | 3.612618e+05 |

| Cluster | Cumulative Total Deaths |
|---------|-------------------------|
| 0 | 1.851472e+05 |
| 1 | 1.286901e+07 |
| 2 | 3.502721e+06 |

Figure 22. Average by cluster for each numeric feature - with “U.S”

```

---- Distribution of Indicators by cluster
Cluster Indicator
0 Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks 0.250000
  Received Counseling or Therapy, Last 4 Weeks 0.250000
  Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks 0.250000
  Took Prescription Medication for Mental Health, Last 4 Weeks 0.250000
1 Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks 0.334242
  Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks 0.271571
  Took Prescription Medication for Mental Health, Last 4 Weeks 0.256585
  Received Counseling or Therapy, Last 4 Weeks 0.137602
2 Received Counseling or Therapy, Last 4 Weeks 0.418367
  Took Prescription Medication for Mental Health, Last 4 Weeks 0.240136
  Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks 0.217687
  Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks 0.123810
Name: proportion, dtype: float64

```

Figure 23. Distribution of Indicators per cluster - with “U.S”

```

---- Counts for each unique Indicator per cluster
Indicator Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks \
Cluster
0 1683
1 736
2 182

Indicator Received Counseling or Therapy, Last 4 Weeks \
Cluster
0 1683
1 303
2 615

Indicator Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks \
Cluster
0 1683
1 598
2 320

Indicator Took Prescription Medication for Mental Health, Last 4 Weeks
Cluster
0 1683
1 565
2 353

```

Figure 24. Counts for each Indicator per cluster - with “U.S”

The averages for each numeric feature by cluster (Figure 22) show that the death totals have a large difference in average value. This explains why the clusters were globally separated. The distribution of indicators by cluster (Figure 23) and indicator counts per cluster (Figure 24) show that each cluster has a significant amount of points from each indicator, typically having a similar proportional distribution between clusters. These imply that there is no dominant indicator per cluster, which is not ideal. This further confirms that there is not enough variance to differentiate indicators within clusters.

Interestingly, the K-means model that was run without rows showing “U.S” as a state (Figure 25) performed worse than the original. I assumed that because the other models performed better with this change so would the K-means model. Upon further reflection, however, it is to be expected that the second K-means model would perform worse due to the increased variation. Without the “U.S” state, a major factor that determined cluster dominance shifted, and since all other states are equally distributed, it was unlikely that any meaningful clusters would form. The elbow method failed to show any meaningful bend in the plot, and the

silhouette scores showed that the highest score was $K = 2$. The PCA visualization plot in the bottom left shows a far worse distribution than the previous iteration of the model, as one cluster is very compact and one is very spread out. Like before, this shows that there is just enough variance to provide a structure, but not enough to create meaningful clusters based on the features given. The plot in the bottom right confirms that the clusters were not meaningful, as there is little to no separation between indicators.

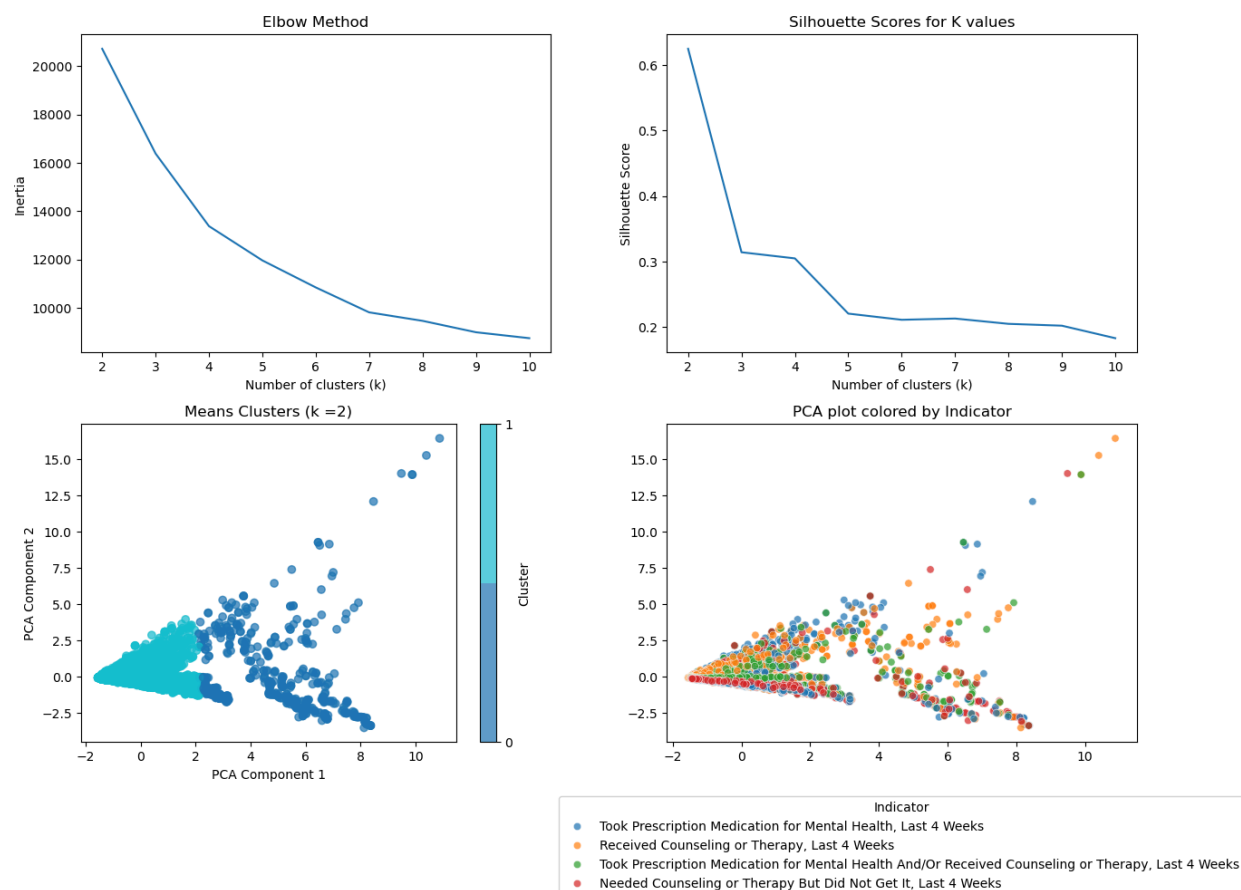


Figure 25. Elbow Method plot for number of K values plot (top left), Silhouette Scores plot for K values plot (top right), Cluster plot via PCA (bottom left), PCA plot colored by Indicator (bottom right) - without "U.S" as a State

The averages for each numeric feature by cluster (Figure 26) once again show that the death totals have a large difference in average value, explaining the global separation of clusters. The distribution of indicators by cluster (Figure 27) and indicator counts per cluster (Figure 28) show that each cluster has a significant amount of points from each indicator, typically having a similar proportional distribution between clusters. Since there is no dominant indicator in any of the clusters like before, there must still not be enough variance to differentiate indicators within clusters. Based on the distribution seen in these figures, I feel that the data became more skewed without rows with "U.S" as a state, since it supplied about 1/3 of all points in the data and basically defined one or two clusters due to the outliers it created. I am not confident in the results due to the poor performance of the models. I feel as though the data fed to the models was lacking in features and prevented class differentiation.

```
---- Averages by cluster for each numeric feature:

      COVID-19 Deaths  Total Deaths  Cumulative COVID-19 Deaths  \
Cluster
0      309.586622      4441.085873      66717.362342
1       60.403112       922.160357      11441.987369

      Cumulative Total Deaths
Cluster
0      754323.046065
1     137402.726830
```

Figure 26. Averages by cluster for each numeric feature - no "U.S"

```
---- Distribution of Indicators by cluster

Cluster  Indicator
0      Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks  0.284069
      Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks  0.266795
      Received Counseling or Therapy, Last 4 Weeks  0.232246
      Took Prescription Medication for Mental Health, Last 4 Weeks  0.216891
1      Took Prescription Medication for Mental Health, Last 4 Weeks  0.252777
      Received Counseling or Therapy, Last 4 Weeks  0.251489
      Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks  0.248591
      Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks  0.247142
Name: proportion, dtype: float64
```

Figure 27. Distribution of Indicators per cluster - no "U.S"

```
---- Counts for each unique Indicator per cluster

Indicator  Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks  \
Cluster
0      148
1     1535

Indicator  Received Counseling or Therapy, Last 4 Weeks  \
Cluster
0      121
1     1562

Indicator  Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks  \
Cluster
0      139
1     1544

Indicator  Took Prescription Medication for Mental Health, Last 4 Weeks
Cluster
0      113
1     1570
```

Figure 28. Counts for each Indicator per cluster - no "U.S"

Conclusions and Discussion (3%):

My hypothesis was unable to be proven correct due to the lack of usable features to train the data on. Should I find another dataset that includes similar information along with other relevant fields, I feel as though the hypothesis could be proven correct. That being said, for the purposes of this project, the results are clear.

None of the classification models or K-means clustering were able to predict mental health indicators from COVID-19 mortality rate. I do not feel that the models were poorly chosen. It is unlikely that I would have been able to find a better model suited for this data than KNN and Random Forest. Given more time and resources, the models could have been tuned to provide greater output. I experimented

with fine tuning the SVM model for many hours and failed to achieve a version of the model which via fine tuning would finish running in 10 minutes and did not bring my laptop's CPU temperature to over 165°F. After that, my main method of experimentation was through finding different models that could use a similar pipeline of data to ensure that it could run smoothly and efficiently on my system. After I began my preliminary analysis of the CDC datasets, I became much less confident in my ability to find a result that would prove my hypothesis. If I were to perform a subsequent exploration of this topic, I would look for better defined datasets with more features. Alternatively, I could continue to use the datasets I have and once again look for datasets that can be linked via the state and time frame columns. Regardless, the goal would be to find a larger number of useful features to feed the models for predictions.

Citations:

"Mental Health Care in the Last 4 Weeks." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, data.cdc.gov/National-Center-for-Health-Statistics/Mental-Health-Care-in-the-Last-4-Weeks/yni7-er2q/about_data. Accessed 12 Dec. 2025.

"Provisional Covid-19 Death Counts by Week Ending Date and State." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, data.cdc.gov/National-Center-for-Health-Statistics/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab/about_data. Accessed 12 Dec. 2025.

Bureau, US Census. "Household Pulse Survey Data Tables." *Census.Gov*, 15 Apr. 2025, www.census.gov/programs-surveys/household-pulse-survey/data/tables.html.