# The Impact on Death Rate and Mental Health by Covid-19 Over Time in California

H. E. Heum, I. S. Slorer, S. Rynning-Toennesen

**Abstract**

This study examines how the Covid-19 predictions has affected the population of California in terms of death rate and mental health issues. To answer this question we have performed both data inference and created two data models. These plots and models are based on several datasets containing information about vaccination rates, restrictions, anxiety and depression, and death and health conditions. Then the data has gone through an extensive data cleaning and exploratory data analysis. What we found out was that there is a clear correlation between Covid-19 restrictions, especially the maximum gathering ban, and mental health. However, we are not able to say exactly which restrictions that contribute to what degree. In addition we found out that by training a linear model on the restrictions, total vaccinations and daily infections it is possible to give a pretty accurate estimate on the number of deaths.

## 1. Introduction

As we are heading into the middle of 2022 the Covid-19 pandemic has controlled most of our lives for the last 2,5 years. The community has seen restrictions vary in strength depending on the current infection rate and the deadliness of the virus. And the overall goal of the restrictions and political policies has been to minimize the overall consequences that the pandemic has on society. It is therefore interesting to look into how the policies have affected society and try to get an understanding of the consequences of inferring the different restrictions.

Our goal and research question in this study is to first get an understanding of how the pandemic has evolved over time. Then to use this knowledge to see if there is any correlation between the restrictions implemented in different states and the amount of Covid cases. Also, we will look into how the different restrictions have affected people's mental health and if there are any differences between the younger and older population. Our goal is to give more insight into the restrictions' effect on mental health. Then the politicians will have better knowledge about the consequences and can make better choices the next time a pandemic strikes.

The overall research on the Covid-19 pandemic is extensive and governments have spent a lot of research on trying to understand how the virus behaves. In the US a lot of the research can be accessed through CDC (Centers for Disease Control and Preventions) website [1]. However, most of this research is rooted in the goal of stopping the virus and embracing studies about the variants, how infectious they are, and how the virus spreads. In this study, we will also focus on people's mental issues and how the restrictions affect people's mental health. There is not done as much research into mental health and Covid but there are also some previous studies done on this. One example is "COVID-19 restrictions and age-specific mental health—U.S. probability-based panel evidence" [2] by Sojie. E, Tham. W.W, Bryant. R McAleer. M. We will complement this study with additional data and plots showing the effects in California.

## 2. Data

### *2.1. Covid data*

For analyzing and predicting covid deaths we used the dataset 'csse_covid_19_daily_reports_us.csv'. For analyzing death counts by sex/age and death counts by conditions we are using the data sets "cdc death counts by sex age state" and "cdc_death_counts_by_conditons.csv". The datasets are collected from data.cdc.gov, Centers for Disease Control and Prevention, which is an American federal agency under the Department of Health and Care Services. The agency is responsible for the protection of public health, especially infectious diseases, environmental medicine, occupational medicine, prevention, and education.

The dataset "csse_covid_19_daily_reports_us.csv" contains State and Region, number of confirmed daily cases, daily deaths, people recovered daily, confirmed cases that have not yet been resolved, and Federal Information Processing Standards code that uniquely identifies counties within the US, incident rate (cases per 100,000 persons), number of people hospitalized, testing rate, hospitalization rate, etc. The dataset "cdc_death_counts_by_sex_age_state.csv" contains start and end date, year, month in which death occurred, state, sex, and age group, Covid-19 deaths, total deaths, along with deaths from Pneumonia and Influenza. The "cdc_death_counts_by_conditons.csv" dataset contains much of the same but also additional data about conditions contributing to deaths involving Covid-19.

### *2.2. Restriction data*

To analyze how the number of Covid cases and death rate and mental health are affected by Covid restrictions we have collected two additional datasets about Covid restrictions and vaccination rate; "state_social_distancing_actions.zip" and "COVID-19_Vaccinations_in_the_United_States_ Jurisdiction.csv.zip". The dataset “state_social_distancing_ actions.zip” is collected from kff.org, KFF (Kaiser Family Foundation) is an American non-profit organization that focuses on major health care issues facing the nation and the U.S role in global health policy. This dataset contains state actions to mitigate the spread of Covid-19 and contains data about statewide face mask requirements, emergency declaration, the status of reopening, stay at home orders, mandatory quarantine for travelers, non-essential business closures, large gatherings ban, restaurant limits, bar closures from June 2020 to November 2021. A challenge we faced with using this dataset was that it lacked data for many of the states for some of the categories. The dataset "COVID-19_Vaccinations_in_the_United_States_Jurisdiction .csv.zip" was collected from data.cdc.gov and contains data about types of vaccines, doses delivered, the number of administered vaccines, the number of people fully vaccinated, the number of people that have received a booster vaccine, and more.

### *2.3. Mental health data*

For analyzing how mental health has evolved along with Covid and restrictions we are using the dataset “nchs_covid_ indicators_of_anxiety_depression.csv” from [data.cdc.gov] (http://data.cdc.gov/). This dataset is collected from a survey conducted by the National Center for Health Statistics (NCHS), which is a 20-minute online survey designed to rapidly monitor changes in mental health with the impact of the coronavirus and to complement the ability of the federal statistical system to rapidly respond and provide relevant information about the impact of the coronavirus pandemic in the U.S. The data is collected in different phases; Phase 1 occurred between April 23, 2020, and July 21, 2020, Phase 2 occurred between July 21, 2021, and October 11, 2021, Phase 3.3 occurred between December 1, 2021, and February 23, 2022, and will continue through May 2, 2022. The questions in the survey are made to obtain information on the frequency of anxiety and depression and are collecting information on symptoms over the last 7 days, and are divided into Patient Health Questionnaire (PHQ-2) and Generalized Anxiety Disorder (GAD-2). An example of a PHQ-2 question is: “Over the last 7 days, how often have you been bothered by ... having little interest or pleasure in doing things? Would you say not at all, several days, more than half the days, or nearly every day?” The alternatives for answers are assigned a numeric value where 0 corresponds to “not at all” and 3 corresponds to “nearly every day”. A person is categorized as having depression with a sum equal to three or greater on the PHQ-2 and categorized as having anxiety with a sum equal to three or greater. Answers to both questions were required to calculate the scores, and people with missing responses were not included in the percentage calculation.

The table contains the percentage of a certain group that is categorized as having anxiety only, depression only, and both. The different groups are state, sex, age, race, education, sexual orientation, and disability status with subgroups within each group. The table also contains which phase the data was collected in, with the start date and end date.

## 3. Description of Methods

### *3.1. Data Cleaning*

The first step of making prediction from the data is making sure the data is possible to work with. This involves cleaning up the data and for our datasets the necessary steps was mainly to fill inn NULL values (Missing values) and missing rows. In addition we had to see if any of the current data did not make sense. The process of finding null values started with plotting a heatmap displaying the missing values as yellow. Here is an example from the restriction data:

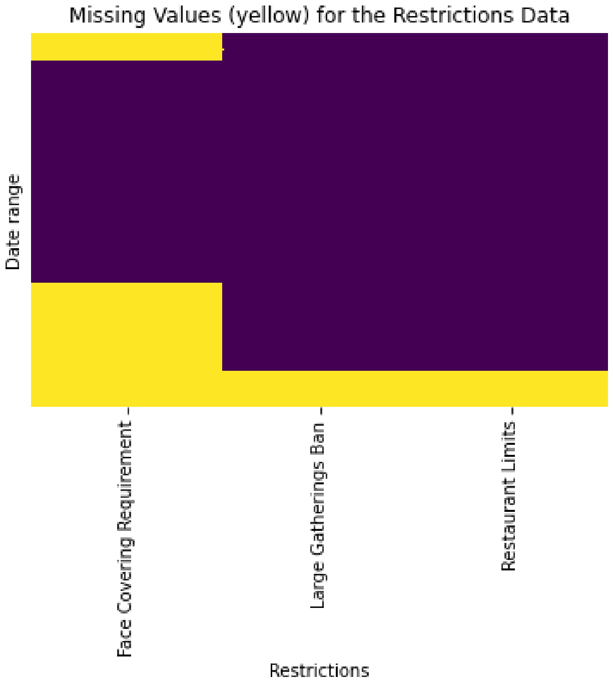


**Figure 1:** Illustration of how we found the missing data for some of the features of the restriction data

Here we can see that for the Face Covering Requirements the missing values are all either at the start or the end. This also made us wonder why the same data was missing for Large Gatherings Ban and Restaurant Limits. We found out that this was because these rules were not in affect anymore so we could fill in the missing data with no restrictions. For the face coverings we found out that the data was not very accurate. Therefore we used Covid-19 US State Policies' [3] to manually update the mask mandate with the correct data.

When it came to missing data we had to use two different approaches. For some data for instance also in the restrictions data we did not have all the dates. What we had to do here was to fill inn all the missing date with the data on the previous date with valid data. We could do this because the data was only updated when the restrictions were changed. The other approach was to add inn data inn the beginning for the dataset that did not have data from the start of the pandemic. This was the case for the dataset vaccination dataset. What we did here was simply setting the number of administered vaccination to 0 because the missing data were from the period before the first vaccines were administered.

When looking for data that did not make sense we specifically looked for negative data. We found at that around 10 dates had a negative number of daily infections or deaths. Because having negative values for these number does not make sense and that there were so few values we decided to replace the values with 0 since it want make a huge impact on the predictions or analysis.

After the different datasets were cleaned and fitted to the right format we had chose to merge sine of them together. The important step here was to make sure that they covered the same time range. This was done by taking the time interval which had no missing data for any of the datasets.

### 3.2. Exploratory Data Analysis (EDA)

To get a better understanding of the restrictions dataset, we decided to make plots describing the datasets features correlation. As we can see on the correlation plot, and also in the list of Daily Death's most correlated values, we can see that the columns adressing openings of restaurants and Face Covering Requirements ha the higher relation with Daily Deaths. This makes them important features when it comes to predicting Daily Deaths.

Looking at the map of correlations we can see that Face Covering Requirements and Large Gathering Bans are heavily correlated. This makes sense as the government often implements several restrictions at the same time. Another noticable trait in the plot is the heavy negative correlation between Large Gathering Bans and number of total cases. This can be interpreted as when the large gathering bans value raises (which determines how strict the gathering bans are) the total cases of Covid gets reduced.

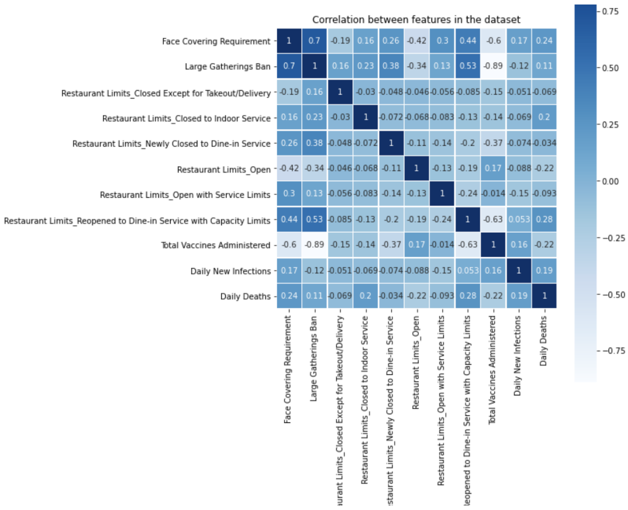


**Figure 2:** Correlation plot of the fetaures in the Restrictions dataset

To get a better understanding of the shape, variability and the center of our statistical data, we made several boxplots.

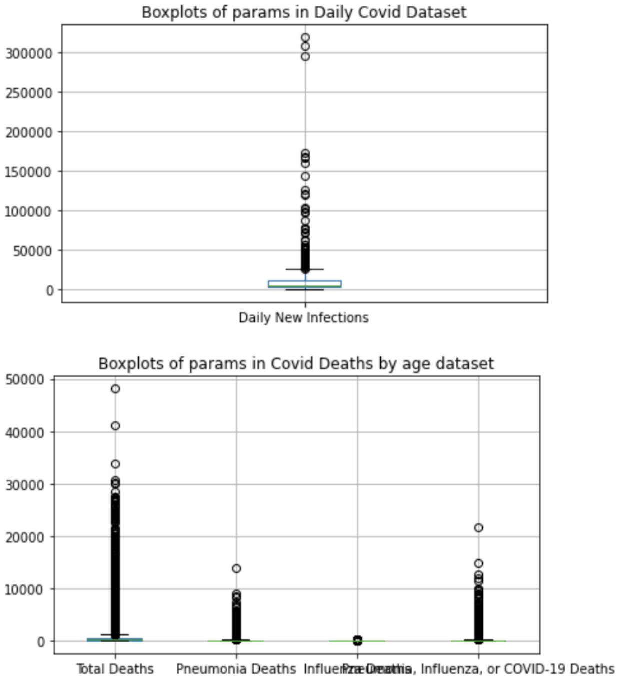


**Figure 3:** Boxplots of outliers in the datasets.

Looking at the Daily New Infections from the first dataset it is noticable many datapoints outside the maximum whisker (Q3+1.5*IQR). These datapoints would normally be interpreted as outliers and hence removed. Due to some domain knowledge and research regarding Covid-19 we understood that the number of Daily infections is a variable with high variance. For instance, the parameter would have a sudden increase during the development of a new contagious covid variant. Therefore, even though it most likely will affect our model and predictions, we decided not to interpret them as outliers and keep the datapoints.

Looking at the boxplot for the Mental Health Score that determines levels of mental health, there are no outliers. This is good for decreasing the variance in our model and increasing the statistical power. One reason for the absence of outliers is that we averaged our mental health data values for each group of anxiety-category. This will cause outliers to be less outstanding.

### *3.3. Model selection*

Our first model was a simple linear regression model with no hyperparameter tuning. To improve our prediction even further we tried several different data science approaches. We first tried to predict with other linear models from the sklearn library, such as the stochastic gradient descent regressor, Lasso and Ridge. Without any hyperparameter tuning we measured each models' performance by using root mean squared error on the prediction and the true values. The performance of the Lasso and Ridge model was slightly worse compared to the normal linear regression model. Surprisingly the stochastic gradient descent regressor performed very bad, with an extreme high RMSE score. This regressor has a lot of hyperparameters we could tune, which could explain its bad performance. Nevertheless, we decided to explore the Ridge and Lasso model further.

### *3.4. Hyperparameter tuning*

The performance of Ridge and Lasso was quite similar to the normal linear regressor, but both models have a hyperparameter we can tune. The models introduces regularization by penalizing large weights on features. How much the weights are penalized is determined by the alpha value. We hyperparameter tuned both models by using sklearn's GridsearchCV. This library uses cross validation to calculate the optimal alpha value. As a scoring function for determining the alpha after the cross validations we used a l1-loss function. This approach is more resistant to outliers, which our dataset contains. After predicting again with optimal alphas, our RSME score got much better and almost the same as the linear regressor. To make our model even more robust to new data we decided to take the average of all the models' predictions. This did not improve the RMSE score significantly, but it will make the model more robust.

### *3.5. Overfitting*

By hyperparameter tuning on a cross-validation set, we reduce the probability of overfitting. After the model is fitted, we test its accuracy on a separate test set. This gives us feedback on its performance on an unknown dataset.

### *3.6. More advanced models*

There exists plenty of advanced open source libraries for models with a special focus on time series. We also tried to make a prediction of the future by using Facebook's forecasting procedure Prophet.

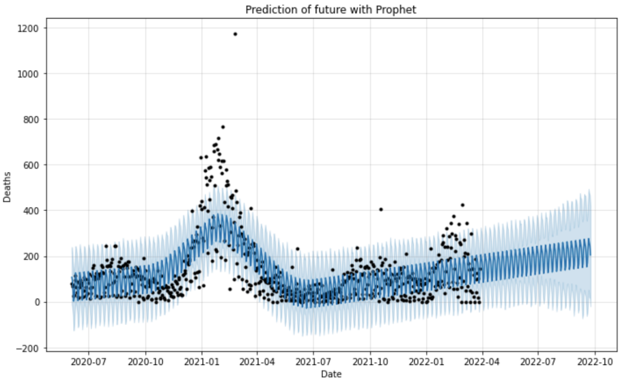


**Figure 4:** Plot of future daily deaths

### *3.7. Feature engineering*

As a final part of the data manipulation we had to modify the features to be able to make as good predictions as possible. Because we are using a linear regression model we want our data to be at specific formats to be able to use it and to find the optimal estimator. The first feature modification we did was to log transform the cumulative vaccination data. The reason we want to log transform this data is to make it linear and we want it linear because a non-linear data indicates a lack of association for the data.

The second feature modification was one hot encoding. We use one hot encoding to be able to use categorical data as input to the linear regression model. Our categorical data that we chose to transform was the restaurant limits for the restriction data. When we one hot encode we add each of the categorical variable as a column and set the value to either 1 or 0 depending if it is the original value for the current row or not.

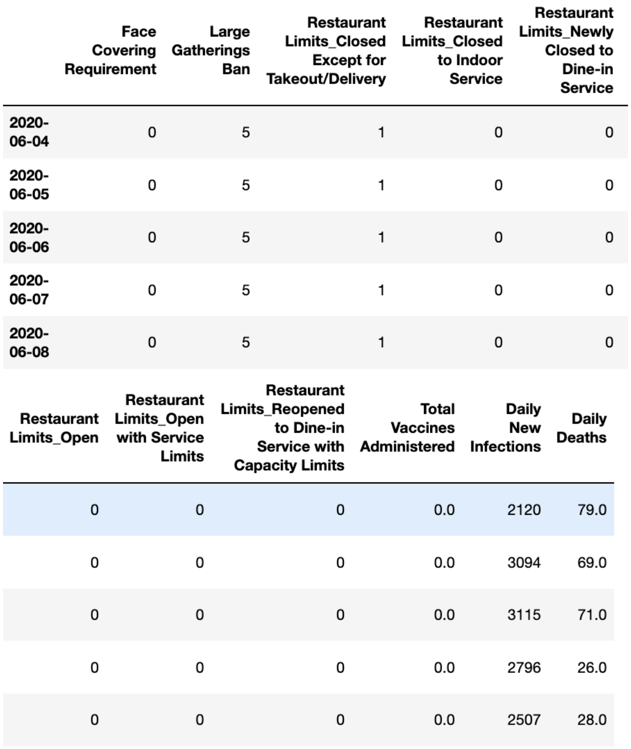
The third manipulation we did was to convert the cumulative infection and death data to rather display the number of new cases. This makes sure the model does not care about how many people that have been infected or died before but only make predictions based on the new daily values.

The fourth manipulation we did was to convert text based data about the maximum number of people that can gather to a discrete scale from 1 to 5. Here was the lowest level of restriction saying that there were no restrictions at all and 5 said that all gatherings were prohibited.

Also as part of the process of modifying the features we have to figure out which features that should be part of the model. This feature selection and modification process was done by trying out different combinations and optimizing the RMSE. RMSE stands for Root Mean Square Errors and is the standard deviation of the residuals (prediction errors). Or more understandable it is a measure of how well your model fits the to the testing data points.

As a result of doing many iterations with modifying the features we did some discoveries. For instance the prediction got worse when using the total number of infections or the daily number of vaccinations. For our feature selection we tried out many different features we did not end up using. On example was data containing the percentage of the infected that was infected with the different variants of the virus. We know from other research that the Delta variant is more contagious than the Alpha variant that was present in the start of the pandemic. However, the model was not able to understand or use this information in the predictions. So we chose not to use it.

For the two final predictions of the number of deaths and mental health score we ended up using the following features:



| | Face Covering Requirement | Large Gatherings Ban | Restaurant Limits_Closed Except for Takeout/Delivery | Restaurant Limits_Closed to Indoor Service | Restaurant Limits_Newly Closed to Dine-in Service | Restaurant Limits_Open | Restaurant Limits_Open with Service Limits | Restaurant Limits_Reopened to Dine-in Service with Capacity Limits | Total Vaccines Administered | Daily New Infections | Daily Deaths |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2020-06-04 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0.0 | 2120 | 79.0 |
| 2020-06-05 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0.0 | 3094 | 69.0 |
| 2020-06-06 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0.0 | 3115 | 71.0 |
| 2020-06-07 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0.0 | 2796 | 26.0 |
| 2020-06-08 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0.0 | 2507 | 28.0 |

**Figure 5:** The variables used for the predictions. The rows are taken from the estimation of the number of death prediction

Note that the rows here are taken from the prediction of the deaths. The rows for the predictions of the mental health score will be a weekly sum of these rows.

## 4. Summary of Results

### *4.1. Interesting findings*

From plotting the Covid data with the death rate for the different age groups we can see that the death rate increases with age. This is not surprising considering that older age groups in general have a worse physical health condition.

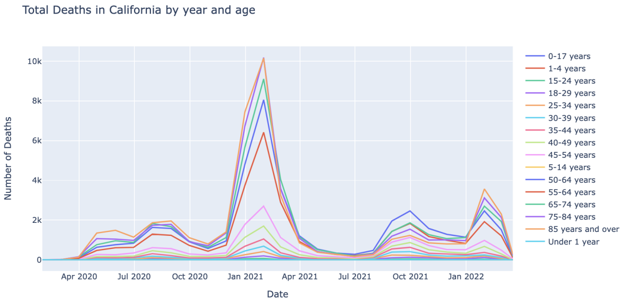


**Figure 6:** Covid deaths in California

From plotting the mental health data we can see that the percentage of people having anxiety/depression is the highest among young people and is lower for higher age groups. We were expecting it to be a difference between young and elder people, considering that younger people tend to have larger circles and in general a bigger need for socializing with larger groups, but were surprised by how significant the difference was.

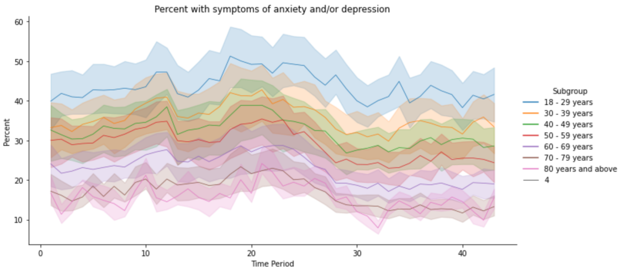


**Figure 7:** Mental health data by age

### *4.2. Analysis*

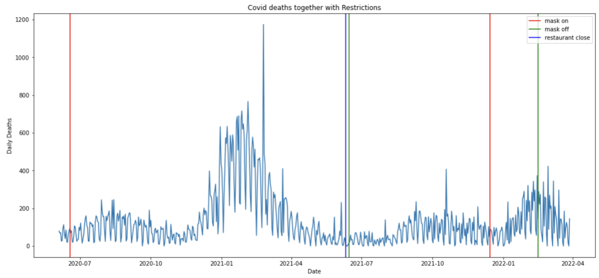


**Figure 8:** Covid deaths with restrictions

The first part of our research question was to get an understanding of how the pandemic has evolved regarding death rates and Covid cases and how this is affected by restrctions. Figure 2 is visualizing the number of Covid deaths in California from July 2020 through April 2022 along with mask restrictions and closing restaurants which is represented by the lines. We can not see any clear correlation between number of Covid deaths and restrictions from this plot, hence the data plotted does not provide us any useful insight on whether the restrictions are helping decrease Covid deaths or not.

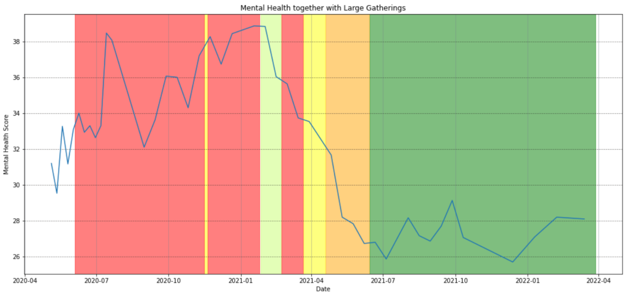


**Figure 9:** Mental health and gatherings

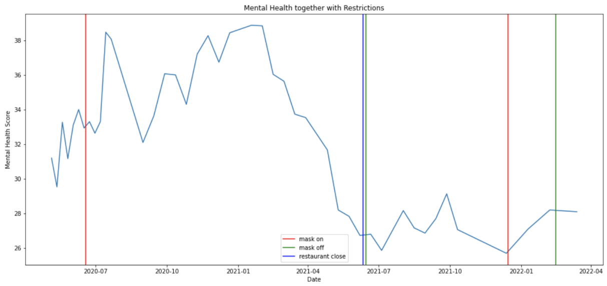


**Figure 10:** Mental health and restrictions

The second part of our research question was to get an understanding of how the different restrictions have affected people'a mental health. Figure 3 is showing the average score of the populations mental health in California, where the mental health score represents the percentage of people in the U.S in all ages being categorized as having depression, anxiety or both. Figure 3 is plotted along with the same mask and restaurant restriction data as in Figure 2, and Figure 4 is showing mental health data plotted along with data of the level of strictness for large gatherings where the red color represents 5 in strictness, the highest level of strictness, and green represents 1, the lowest level of strictness.

From Figure 4 we can see a clear correlation between gathering restrictions and the percentage of people having anxiety and/or depression. Between July 2020 and February 2021, when the restrictions for large gatherings were at the highest level of strictness, the percentage of people with depression and/or anxiety were at its highest with around 10%- higher depression rate than between July 2021 and April 2022, around a year later, when the restrictions for large gatherings were at the lowest level of strictness. In the period between February 2021 and July 2021, when the restrictions were slowly being removed the percentage of people with anxiety/depression was drastically decreasing.

Figure 3 is slightly harder to analyze. We do not think it makes sense that the mask requirements have any direct effect to mental health, but we decided to interpret the lines as points in time when stricter restrictions in general were put into place. The graph shows that mask is required in the same periods as the gathering restrictions are at its strictest, and the mask requirement is removed along with the gathering restrictions in June 2021. The plot shows that the percentage of people with depression/anxiety is already sinking by around 10%- before the mask requirement is removed, as the gathering restrictions are beginning to loosen up. As mask required again around December 2021 we can see a small increase in percentage of people with anxiety/depression again. The mask requirement were put into place again along with other restrictions so it is hard to tell exactly what restriction is causing this since we only have data on the mask requirements, but we can conclude that the restrictions are in general having a negative effect on people's mental health.

### *4.3. Predictions*

As described in the methods section the predictions are made with the features from Figure 5. The first prediction we did was trying to predict the total number of daily deaths in California. The resulting prediction is:

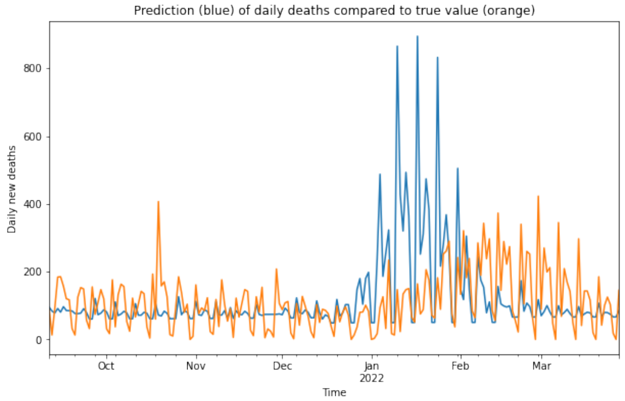


**Figure 11:** Our prediction of daily deaths made with a linear regression model

The first noticeable finding is that our prediction is fluctuating a bit from day to day. However, this fluctuation is either much smaller or larger than the actual values, which seem to fluctuate by almost the same amount for the whole time period. In addition we can see that the prediction is also predicting an increase in the number of death. But the prediction of this spike is more aggressive and suspects that there will be more deaths than it actually is and that they will occur over a smaller period of time. A summary would be to say that the prediction is not completely accurate but it is still able to catch some of the properties of the real data.

With the second prediction we are trying to predict the average mental health score over time for people of all ages. The prediction is:

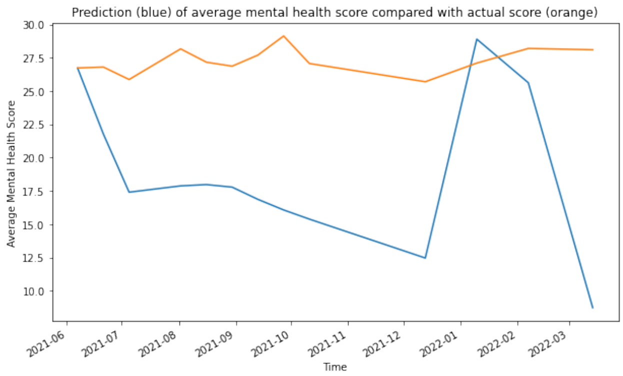


**Figure 12:** Our prediction of average mental health score made with a linear regression model

Here we can see that the prediction (in blue) is really far way from the actual data (in orange). This could be an indicator that Covid restrictions are not good features for estimating mental health. But it could also just be an indicator that we have not added enough variables and it could still be possible to create a good model by working on our features and tuning them.

## 5. Discussion

In our research question we stated that we wanted to look into if there are any differences in mental health between young people and old people and compare that with restrictions. When we started working with the mental health data we realized that it was only possible to look at mental health data in either age or state, and not both simultaneously. Our current restriction data only includes restrictions in California. This made it challenging to draw conclusions regarding how restrictions affected mental health among different age groups. To make further analysis, and to be able to confirm/deny our hypothesis we have to gather more data regarding mental health and construct more restriction data across several states.

In our current prediction model we used three linear models that are all quite similar[4], but different types of regularization loss functions. To make even more robust models with improved performance, we could use other types of predictions than linear regressors. For instance, there exists advanced regressors such as LightGBM and XGBoost[5] that most likely, with optimized parameters, will make better predictions. By utilizing these models together with advanced machine learning techniques, such as bagging, ensemble and stacking [6], our model would perform better.

When we predicted Covid-19 deaths we used restrictions, vaccine rates and new Covid cases based on a **daily basis**. The prediction of new Covid deaths is an advanced and complex analysis to understand. There are more factors that would affect Covid deaths than just daily changes. For instance, the longer Covid has occurred - the predictions should be improved. We should take advantage that there exists Covid data since patient one, and us all currently existing data to predict new cases. As we slightly mentioned above, there exists open source libraries with advanced prediction models of Time Series. These should be used for making more complex models and hence get better predictions.

## References

[1] *Science and Research.* https://www.cdc.gov/coronavirus/2019-ncov/science/science-and-research.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fmore%2Fscience-and-research.html, (Retrieved 04.29.2022).

[2] E. Sojli, W. W. Tham, R. Bryant M. McAleer: *COVID-19 restrictions and age-specific mental health—U.S. probability-based panel evidence.* https://www.nature.com/articles/s41398-021-01537-x, (Retrieved 04.29.2022).

[3] *Face Mask Mandates.* https://statepolicies.com/data/graphs/face-masks/, (Retrieved 04.29.2022).

[4] *Regression Models.* https://favtutor.com/blogs/ridge-and-lasso-regression, (Retrieved 04.29.2022).

[5] *ML models.* https://neptune.ai/blog/xgboost-vs-lightgbm, (Retrieved 04.29.2022).

[6] *Machine Learning Techniques.* https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205, (Retrieved 04.29.2022).