**Open Ended Modeling Report for the COVID Dataset**

Group members:

Yucheng Jin, 3037442109, yuchengjin@berkeley.edu

Tinna Liu, 3035729670, tinnaliu@berkeley.edu

Jiayin Guo, 3035418572, guojiayin@berkeley.edu

**Part 1: Open-Ended EDA**

In this part, we introduce the open-ended EDA performed on the COVID dataset. By making a line plot of the cases versus people fully/partially vaccinated in California, we found that after people started mass vaccination, the growth rate of COVID infection slowed down significantly. Then we investigated the relationship between the mean increase of new COVID-19 cases and mask-use situation. From the plot, there's no clear relationship between the two variables: there are counties where people wear masks more frequently but observe a high mean increase value, while there are also counties where people do not wear masks very often but observe a low mean increase value.

**1a. Relationship between Cases and Vaccination in California**

We first investigated the relationship between the number of new COVID cases and vaccination in California.

图表, 折线图

描述已自动生成

Figure 1. Number Per Capita of COVID-19 Cases/People Fully Vaccinated/People Partially Vaccinated in California from 2020-12-10 to 2021-9-26

From the above plot, before the COVID vaccines were put into use, the number of new COVID cases in California increased rapidly every day. However, after the start of mass vaccination, the growth rate slowed down significantly, and the number of new infections remained almost unchanged for nearly one year. Although there were many other factors that influenced the situation, from the above plot we can conclude that vaccination has positive effect on preventing the spread of COVID-19.

**1b. Relationship between Mean Increase and Mask Use in California**

Then we investigated the relationship between the mean increase of COVID cases and mask usage situation in California.

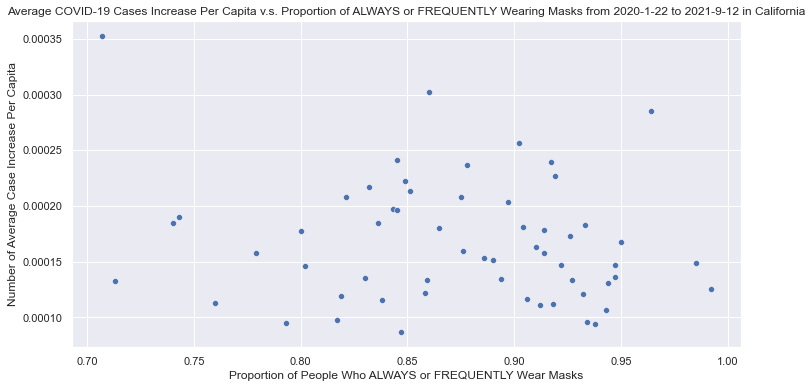


Figure 2. Average COVID-19 Cases Increase Per Capita versus Proportion of ALWAYS or FREQUENTLY Wearing Masks from 2020-1-22 to 2021-9-12 in California

The plot above shows the relationship between the proportion of people who reported to "ALWAYS" or "FREQUENTLY" wear a mask and the number of mean COVID-19 cases increase per capita from 2020-1-22 to 2021-9-12 in California, where each point in the plot represents one county.

There's no clear relationship between the two variables in the plot: there are counties where people do not always wear masks but have low mean increase, and there are counties where over 90% of people always wearing a mask but have relatively high mean increase.

It seems that wearing a mask is not the only factor that influences the increase of COVID cases. In addition, we notice that, among those who reported to always/frequently wear a mask, they may not wear the mask properly and they may also lie about the fact. Therefore, we brought up some open-ended questions that we would like to further explore:

* What are some possible underlying factors that complicated the mask-cases relationship?
* If we were to include these additional features as part of our feature selection, would we be able to uncover a clear relationship between mask usage and cases increase? If yes, what type of relationship would that be?
* How does the current number of cases per capita influence the pace at which the cases increase?

**Part 2: Problem**

**2a. Hypothesis**

We propose our hypothesis as: Mask usage and population density are both decisive factors that influences the number of cases per capita. In particular, the frequency of people wearing masks is negatively related to cases per capita, and population density is positively related to cases per capita.

Our hypothesis is intuitive. From the open-ended EDA, there is no obvious correlation between wearing masks and the increase of cases per capita. To our reasoning, wearing masks is an effective measure against the spread of COVID, so we deduce that there might be other underlying factors that affect the relationship. While wearing masks could reduce the increase of COVID cases, it is also sensible to assume that individuals who live in more crowded regions are more likely to wear masks. If it is true that regions with denser populations are more likely to observe a rapid increase in COVID cases per capita, it might look like that wearing masks is positively correlated with cases increase.

As a result, in Part 2 of the project, we directly investigate the effect of population density on the number of COVID cases. Population density is not a feature that can be obtained from the existing datasets in Part 1, so we read in an additional database that contains population density from the U.S. Census Bureau.

**2b. Accepting/Rejecting Hypothesis**

We will conduct a multiple linear regression on the mask usage, population density, and cases per capita data, and examine the sign and magnitude of the coefficient for each of the features. We expect features that are positively correlated with cases per capita have a positive coefficient in front of them, and features that are negatively correlated with cases per capita have a negative coefficient in front of them. The magnitude of the coefficient as well as RMSE determines the strength of the relationship. Therefore, we will accept our hypothesis if we observe a large positive coefficient for population density, a relatively more positive coefficient for less frequent mask usage compared to that of more frequent mask usage, and a small RMSE. Otherwise, we will reject our hypothesis.

**Part 3: Modeling**

In this part, we introduce the modeling process. This includes the modeling techniques we used, feature selection, the input and output of the models, and the reasons why we chose these models. We start with our baseline model.

**3a. Baseline Model**

The baseline model f\_base uses a multiple linear regression. Input to the model is the population density for each county and the mask usage percentage data for each county, as separated into always, frequently, sometimes, rarely, and never. The output to our model, which is our prediction, is the number of COVID cases per capita for each county on Aug. 2nd, 2021. We chose this date specifically so that there are enough dates before it whose cases data could be used to improve the model in the next steps, and there are also enough dates after it so that we could analyze the long-term prediction accuracy of our model. We chose our inputs as to closely follow the variables we are interested in in our hypothesis, and we picked our model as multiple linear regression because we deducted that there would be a somewhat linear/positive relationship between our input features and our output variable.

**3b. Improved Model**

The improved model f\_improved also uses multiple linear regression method. From our baseline model, we computed the RMSE for our training and testing datasets, and we found the RMSE’s to be rather large (around 0.03) compared to the scale of cases per capita. Therefore, we decided to include more features in our model prediction. In particular, we included the cases per capita for each county on the first day of each month from April 2020 until August 2021, and we included the one hot encoding of the state that each county is located in. We selected the past cases per capita data because we wanted to explore how the past trend could affect future case growth, and we selected the state information based on an EDA from part I of the project, where we observed that the increase of cases per capita differs among different states due to different political, economic, and climate environments. The output of the model is still the number of COVID cases per capita for each county in Aug. 2nd, 2021. We chose a multiple linear regression model because we believed that the time series data would fit perfectly with a linear model, and the one hot encoding also fits a linear model.

**Part 4: Model Evaluation and Analysis**

In this part, we evaluate and analyze the performance and results of our models. We used time-series method in our experiment and obtained some meaningful visualizations.

**4a. Model Evaluation**

To evaluate the result of our model, we split our data into training set and test set, and we computed the RMSE on both the training set and the test set. For our baseline model, the training RMSE is 0.02996, and testing RMSE is 0.03521. Since the scale of our variable, cases per capita, is rather small, the RMSE’s for our baseline model are relatively large compared to the scale of our predictions. This indicates that our baseline model does not provide us an accurate prediction. Also, since the testing RMSE is higher than the training RMSE, we believe that there might be a slight overfitting to our training set.

For our improved model, the training RMSE is 0.00031, and the testing RMSE is 0.00034. Both RMSE’s are significantly smaller than those of the baseline model (in fact, almost 100x smaller), which makes us believe that our improved model is much better than our naïve baseline model.

**4b. Model Visualization and Analysis**

In this part, we will show plots that illustrates the coefficient of each feature that we use in our improved multiple linear regression model and analyze how each feature contributes to our model prediction. Since we are dealing with time series data, we will also use a visualization to illustrate the accuracy of our model predicting cases per capita in the near future and the relatively distant future.

**4bi. Parameters**

Cases per Capita in the Past

图表, 散点图

描述已自动生成

Figure 3. Parameters of Cases per Capita Over Time in the Improved Model

In Figure 3, we visualized the coefficients of the cases per capita in the past 17 months, counted back from 8/1/2021. We see that the coefficients corresponding to closer dates have a really high impact on our model prediction. In particular, the parameter corresponding to 8/1/2021 is about 1, which means that our model predicts that the cases per capita for each county on 8/2/2021 will be very close to that of 8/1/2021. Parameters for earlier dates stay very close to 0 and scatter above or below, which means that the data from further back are less valuable to predict the present, and there is randomness involved when using a distant past to predict the present.

Population Density

The coefficient for population density is 0.0, which shows that under this set of feature selection, population density seems to be uncorrelated with cases per capita. One possible cause is that we use the number of cases per capita instead of the direct number to represent our cases information, that is to say, the "population density" information has already been included in the predicted feature. Another possible cause is that the magnitude of cases per capita is too small comparing to the magnitude of population density, such that a small coefficient for population density could contribute a lot on our model prediction.

Mask Usage

图表, 散点图

描述已自动生成

Figure 4. Parameters of Mask Usage in the Improved Model

The parameters of the mask usage features are all negative, which could be caused by the fact that some other features that we used to predict affects the absolute scale of the mask usage parameters, so that we cannot observe positive parameters for NEVER and RARELY, for example. This does not necessarily mean that mask usage has no effect on the spread of COVID, because it might be that the effect of mask is overridden by some other features (possibly those that are stronger) that we have included in our modeling.

OHE States

图表, 散点图

描述已自动生成

Figure 5. Parameters of States in the Improved Model

The parameters of the states are listed in the above figure. Florida has the lowest parameter and Louisiana has the highest. We can intuitively interpret a positive parameter as the corresponding state does not do very well in preventing the COVID-19 spread, or it has a better detection ability as the data collected here is clearly not the actual number of cases. While experimenting with different models on the data, we found that the RMSE with one-hot encoded state information is much lower than the RMSE without, which indicates that state helps make our model prediction more accurate.

In sum, the visualization of coefficients illustrates the strength of each feature. From our result, the cases per capita from the nearest date has the strongest predictive power on our prediction, and population density seems to have no obvious correlation with cases per capita. From the hindsight, the reason that population density and mask usage data do not have a strong relationship to cases per capita is probably that their effect on cases per capita is too week compared to that of the cases per capita on the nearest date, which makes their effect not so obvious.

**4bii. Time Series Comparison**

Chart, line chart

Description automatically generated

Figure 6. RMSE for Two Models & One-day Growth Rate for Two Models as a Function of Time

Since we work with time series features, we were curious how our baseline model and improved model perform when predicting cases per capita in the near future and the more distant future. Therefore, we graphed the RMSE for each of the prediction date from 8/2/2021 to 9/12/2021. Over the 40+ days span, we observed that both models have increasing RMSE as time goes by, while our improved model consistently gives a much smaller RMSE than the baseline model does. This means that both models do worse when trying to predict a farther future, but the improved model predicts better than the baseline model.

This plot shows there is room for improvement with regards to predicting future cases. There might be other useful feature, such as vaccination data, hospitalization data, and medication data that we could possibly include into our model, or it might be possible to change our multiple linear regression model into another type of model to increase prediction accuracy.

**Part 5: Answer**

In this section, we will answer our hypothesis and provide a short justification of why we believed so.

From our baseline model and improved model, we fail to observe a strong correlation between both mask usage and population density and cases per capita. In our baseline model where the only features that we selected are mask usage and population density, the RMSE of our model is rather large. In our improved model, we failed to show the relationship because the small or random coefficients in linear regression model for both features, as shown in part 4bi.

**Part 6: Model Improvement**

In this section, we will discuss some steps that we have taken along our way to improve our model.

**Improvement 1: Adding Cases Info from Past Year**

In our original design for the improved model, we only included cases per capita data from one selected date, and the date was picked rather arbitrarily. As we were exploring the relationship between RMSE and time past the selected date, we realized that the RMSE increases rapidly as time increases. We deduced that adding more data points for cases per capita from the past might help with the situation. Therefore, we picked the cases information of the first day of each month since there has been a significant amount of COVID cases in the U.S., which happened around April 2020, as our additional features. We again plotted RMSE vs. time since a chosen date, and the result is shown in Figure 6 above. Although RMSE still increases with time, the rate of increase is much slower when all the additional features are present in our model input, which means that adding cases per capita information from history helps us predict the output in the future.

**Improvement 2: Using Cases per Capita Instead of Cases**

Originally, the variable that we were predicting was the number of COVID cases for each county. With this, when predicting the number of COVID cases on Aug. 2nd, 2021, we achieved a training RMSE of 63.24, while the testing RMSE was a lot bigger, at 205.27. The large difference between training and testing RMSE indicates that there might be overfitting to our data. The overfitting could occur because the absolute number of cases for each county depends on the county’s population, and the situation for one county cannot be generalized to the others. Therefore, our solution to this overfitting problem was to use cases per capita instead of cases, because when analyzing cases per capita, all counties are treated on the same scale. Doing so, we achieved much closer training and testing RMSE, with 0.00031 for training and 0.00034 for testing.

However, one thing to notice is that with the use of cases, we were able to retrieve a non-zero coefficient for population density, at 0.000826. This number is still fairly small, and it is possible that the coefficient happened to be positive due to some randomness in data selection, but this could possibly hint at the fact that population density is positively correlated with the number of cases. To confirm this hypothesis, we will need further tests which isolate the effect of population density on the number of cases, and we will need more careful analysis on the result.

**Part 7: Future Work**

A direction of future work could be to transform feature before using them as inputs to the multiple linear regression model. For example, we could try to log the cases per capita data and feed the logged data into the model because the number of COVID cases followed an exponential growth during some phases of the pandemic. As our current model does not perform as good when predicting cases per capita for a relatively distant future, using log-transformed features might help our model gain accuracy when predicting future data and trends.

Another step that we could possibly take is to study the political and economic stance of each state, and manually cluster the states with similar conditions into a group. We could use this group number as a replacement to the current one-hot-encoding of the states. This might help reduce overfitting to specific states. This could also shed light on which political party has the best strategy to prevent against COVID and might provide interesting observations.