Final Report: Brazil's COVID Response

1. Introduction and Question

Countless lives have been affected globally due to the COVID-19 pandemic, leading many countries to take drastic measures to control the spread of the virus. Some major policies implemented include (1) early lockdown, which calls for lockdown at the onset of COVID-19, as practiced in India, (2) late lockdown, which calls for lockdown when the cases are peaking, as practiced in Italy, and (3) close contact tracing, which ensures that all affected individuals and their close contacts are traced and tested, as practiced in South Korea. Surprisingly, despite being one of the hardest-hit countries by the pandemic, Brazil has yet to implement any significant measures against coronavirus.

This leads us to ask the question: which strategy, or combination of strategies, between the three listed is the most effective for containing COVID-19 in general and for Brazil?

2. Files, Pre-Processing, EDA and Limitations of Datasets

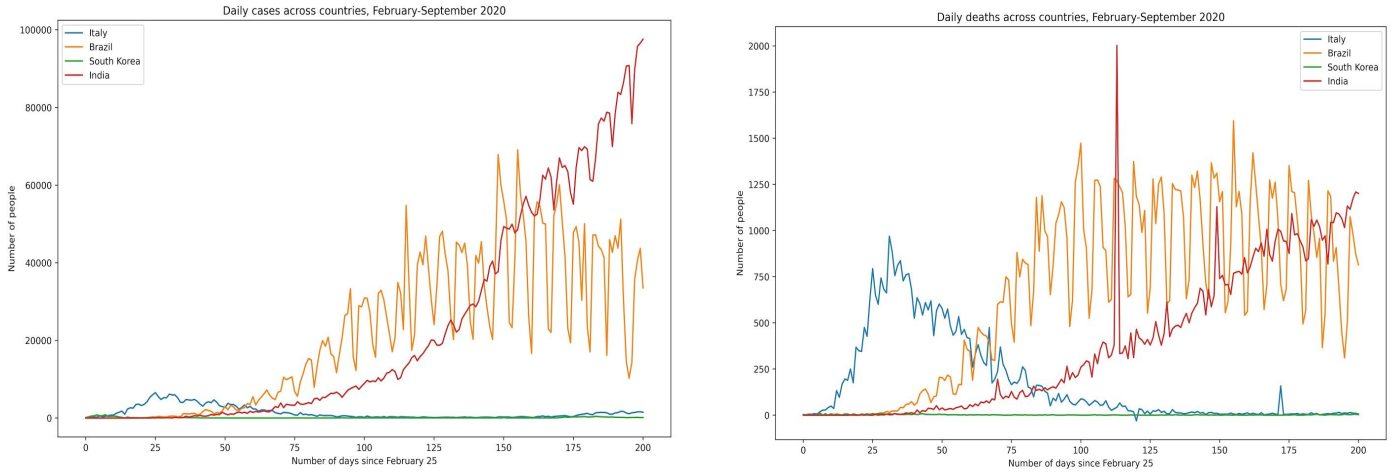

Figure 1: Combined EDA Graph: daily cases and deaths in Italy, India, South Korea, and Brazil

a. Italy

For the Italy EDA and models, we used primarily the "covi19_italy_region.csv" file, which contains daily entries for each region in Italy (20 regions in total) from February 24 to September 12. Each daily entry has a serial number and variables for date, location, hospitalizations, current positive cases, new positive cases, currently recovered, current deaths, total positive cases and tests performed. For modeling, we reduced the dataset to three main variables: "date", "daily new cases", "daily new deaths", and for testing, we used 75-25 train and validation split. To get the national daily cases variable, the dataset was grouped by "Date", summing the "new positive cases" for all the regions on a specific date. To get the number of daily new deaths, the entries were grouped by date, summing the "current deaths" for all regions, and then the difference between the cumulative values was found. For EDA, we made a plot of national daily cases and daily deaths. Daily cases (figure 1a) reached the highest peak on March 21 with 6557 daily cases, almost 2 weeks after a national lockdown was mandated on March 10. In March we can also see the highest peak for daily deaths (figure 1b), reaching almost a 1000 per day. During the summer months, daily cases and deaths were low (200-500 daily cases), a

direct result of the strict lockdown measures implemented in the spring. After restrictions were lifted in the summer, cases and deaths started to rise gradually in September.

b. India

We used the covid_19_india.csv file. For preprocessing, since there were a lot of entries that had multiple variations of the states or union territories, these entries were mapped to the correct regions. To get the daily number of deaths, cases and cured people for each day, the difference between the cumulative counts in the data was found. For further modelling and analysis, the data was reduced to three relevant variables of dates, daily new cases and daily new deaths, and a 75-25 train and validation split was used every time. For EDA, a plot with the number of new daily cases and deaths plotted was made (Figure 1). Interestingly, in the EDA, we noted that suddenly, beyond late May 2020, the cases shot up massively. This makes sense because the Indian government began lifting its lockdown around the same time. Finally, another interesting finding was the outlier peak of deaths around the end of June, 2020. As for limitations, this data lacks other helpful indicators such as number of hospitalizations.

c. South Korea

For South Korea EDA and models, we found the "covid_19_data.csv" dataset particularly useful. By taking the differences of cumulative counts recorded by the variables of "Confirmed", "Deaths" and "Recovered", we computed daily counts of cases/deaths/recovered which we used as response variables. For our analysis, we used 75-25 train and validation split.

The one limitation with the datasets was that it did not provide any information about cases in region or state level in South Korea. It could have been insightful to map if there were any regions in South Korea compared to the other regions. From the above figures, we noticed that South Korea had the lowest cases and deaths compared to the other countries, but it had the biggest spike in March before major containment policies were enacted and a small spike around September.

d. Brazil

In developing our final model, we worked primarily with the provided "brazil_covid19.csv" dataset. As done for the other countries' datasets, we computed the daily national changes in cases/deaths (our response variables) from this dataset's cumulative case and death counts, aggregated by date. Because Brazil has not yet implemented a lockdown, we did not have a specified date on which to split training and testing sets, so we elected to use the first ¾ of the dataset for model training and the remaining set for validation.

As before, we generated a plot comparing the daily changes in cases, deaths, and recoveries, although the counts of recoveries are potentially inaccurate (due to the limitation of many "NaN" and missing dataset values). We observed, as seen in the above figure, huge recurring spikes and falls in the counts of daily new cases and deaths, which makes sense given the lack of containment policy thus far in Brazil.

3. Baseline Model: Polynomial Regression

For our baseline model, we selected polynomial regression as we did not observe a linear relationship between our response variable (daily new cases) and our predictor variable (number of days since the start of lockdown measures) for India, Italy and South Korea. We used a validation set to find the most promising polynomial relationship. Firstly, we compared degrees

up to 30 and selected the best degree with the lowest mean-squared error (MSE) on the validation set. Secondly, we integrated k-fold cross-validation to the previous model to reduce the possibility of overfitting to the validation set and found the best degree similarly. Finally, we used k-fold cross validation with bootstrapping to estimate the performance of the model. For this model, we selected a maximum degree of 50 and bootstraps of 100. For each bootstrap and for each degree, we performed k-fold validation with k =10, and computed MSE on training and validation data, before selecting the best degree with the lowest validation MSE. After using these three methods of model selection, we found that the best polynomial fit degree for the India data is 16, for the Italy data is 7, and for South Korea is also 7. The graph below displays an example scatter plot of India's daily new cases since the start of lockdown measures, with the model predictions in red.

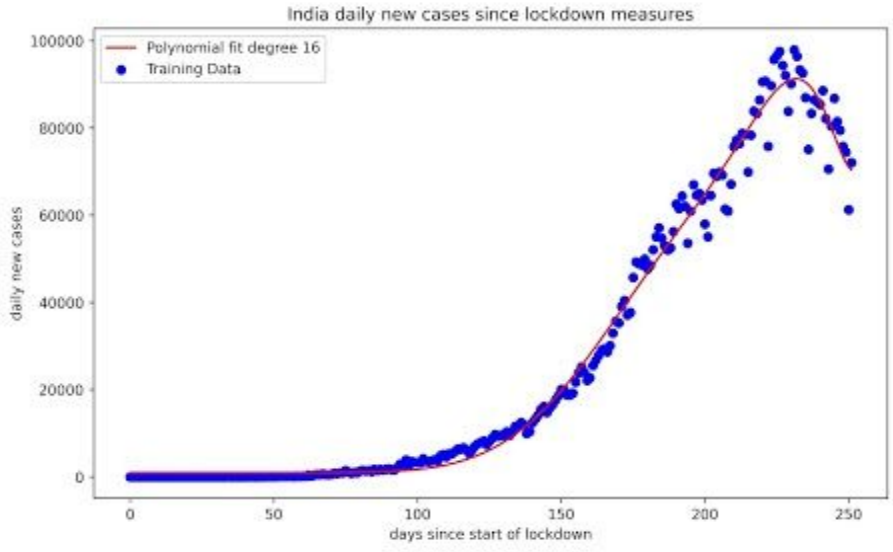

Figure 2: Polynomial Regression Model on Indian daily new cases

4. Choosing a Time-Series Model

Though we opted for a polynomial-regression baseline model, we felt that a time-series model would be more appropriate, as time-series models are able to account for more complex features, such as seasonality and trends, and are more reliable in extrapolating findings for predictions than linear/polynomial-regression models are[1]. Given the time-ordered nature of the data and the observed repeating trends in daily new cases/deaths, we elected to use a time-series analysis in building our final prediction model.

Step 1: Seasonal Decompositions

In order to prepare our datasets for time-series modeling, we first performed seasonal decompositions of each dataset to investigate the presence of any trends in the daily spread of COVID. As seen below in the "seasonal" component of Brazil's decomposed case data, we noticed that each dataset exhibited similar patterns of seasonality, confirming our choice of a time series as an appropriate model.

[1]https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-one/?fbclid=IwAR3_4epMKFE0fMxjOtZ4DYbNGJrOJbLRfKjkX8qd2cq-QYedYaFFdIZ3Ve0

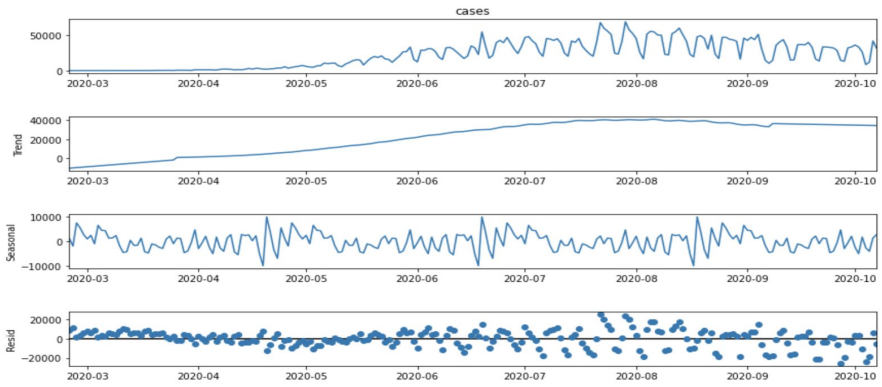

Figure 3: Example of a seasonal decomposition on Brazil's daily case data, performed using the 'seasonal_decompose' function of the 'statsmodels' library.

Step 2: Stationarity Check

We finally proceeded to check for stationarity in our data, which is an essential assumption for predicting with most time-series models. After detrending and differencing our datasets, we successfully validated the stationarity condition through ADF tests[2], allowing us to proceed with fitting models.

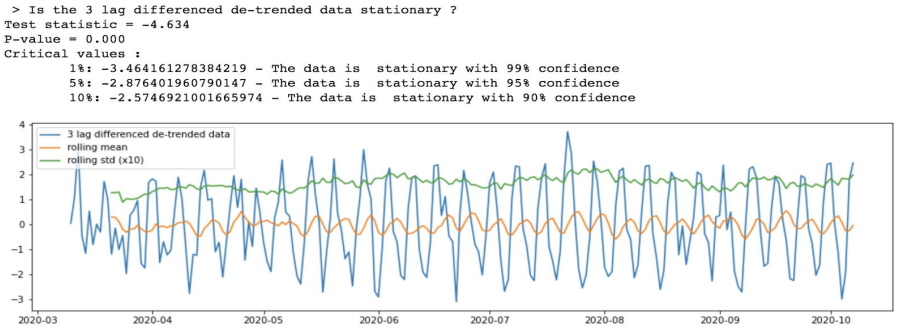

Figure 4: Example of an Augmented Dickey-Fuller (ADF) stationarity test on Brazil's differenced, detrended daily case data, performed using the 'adfuller' function of the 'statsmodels' library.

Step 3: Time-Series Model Selection

To determine the best time series model for our data, we chose to look at 5 different prediction models and calculate the root-mean-square error (RMSE) of each model by using 75/25 train/validation split. Not all of these models were suitable for each country, so we determined which models to use based on trend and seasonality of the country data.

- Simple Exponential Smoothing (SES)[3]: this model is suitable for time series data without trend or seasonal components. The model calculates the forecasting data using weighted averages. This model uses a smoothing parameter: α between 0 and 1. When α=0, the forecasts are equal to the average of the historical data. When α = 1, the forecasts will be equal to the value of the last observation. We implemented two SES models, one with α = 0.8 and one using the Python 'statsmodels' module to automatically find an optimized value for the dataset, which gives a lower error. From the visualization of the results we can see that the forecasted data points form a straight horizontal line because the model

[2]https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788
[3]https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-two/

does not predict any fluctuation given by the trend or seasonality of the data. Since all of the countries showed some kind of trend and seasonality in the data, we chose to use this model only as a baseline for comparison.

- Holt's Linear Trend Method[4]: this model is suitable for time series data with a trend component but without a seasonal component. In addition to the level smoothing parameter α, the Holt method adds the trend smoothing parameter β (also between 0 and 1). We chose to use the parameter values $\alpha = 0.6$, $\beta = 0.2$ and we implemented two Holt models, the default Holt's additive model and the exponential model. The exponential model is appropriate for situations where the increase or decrease starts slowly but then accelerates rapidly. The visualization of the results shows a trend that is more dramatic than the actual observations because it does not include seasonality and focuses on the overall trend of the data.
- Holt-Winters' Seasonal Method[4]: this model is suitable for time series data with trend and/or seasonal components. The method includes this seasonality smoothing parameter (γ) We implemented two Holt-Winters models based on the type of seasonality (additive or multiplicative) We used the additive model when seasonal changes in the data stayed roughly the same over time and didn't fluctuate in relation to the overall data. We used the multiplicative model when the seasonal variation changed in relation to the overall changes in the data. For instance, if the data was trending upward, the seasonal differences would grow proportionally as well. The model also requires a frequency of seasonality parameter, which we set to 14, since covid symptoms usually appear within two weeks (14 days).
- ARIMA/SARIMA[4]: this model is suitable for time series data with trend and/or seasonal components. Auto-Regressive Integrated Moving Average (ARIMA) models look at autocorrelations or serial correlations in the data. In other words, differences between values in the time series. SARIMA builds upon the concept of ARIMA but extends it to model the seasonal elements in your data. The model has 7 parameters: 3 trend elements, 3 seasonal elements and the seasonal period. We chose these parameters by using "grid search" to iteratively explore different combinations of parameters and select the one with the lowest AIC (Akaike Information Criterion) value.

After comparing the RMSE values of the models, we found that SARIMA displayed the lowest RMSE value among the suitable models for each country, so we chose to use SARIMA to make forecasts of future COVID cases and deaths.

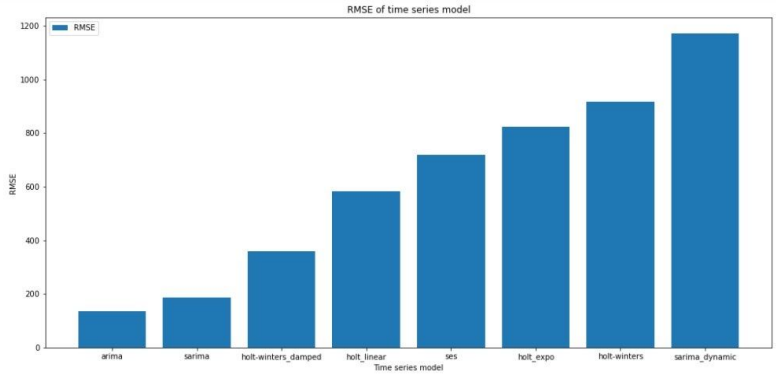

Figure 5: RMSE of each time series model trained and tested on the Italian daily cases.

[4]https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-two/

5. Time-Series Analysis

5.1 Daily Case Model

After identifying SARIMA as the best model, we trained it to predict new daily cases in all the countries. We conducted two main types of analyses. First, the naive analysis trains on all the new daily cases in the dataset and predicts till 2021. This naive analysis helps us isolate how the overall case count will grow for each country. Second, the advanced analysis trains on all the new daily cases in the dataset till the policy is implemented, and then forecasts into the future. Comparing this forecast with the actual case count helps us determine the efficacy of the implemented policy. For the specific insights from the two analyses on each country, see below:

Italy: The naive analysis tells us that the daily new cases will continue to increase linearly well into 2021, reaching almost 5000 cases per day by April 2021. The advanced analysis, which forecasts case counts after the late lockdown in Italy was mandated on March 10 2020, shows a rising trend. If lockdown measures hadn't been implemented, the daily new cases would have been much higher than the actual observed cases, reaching almost 60,000 daily cases by September 2020 instead of the 1,500 in the actual data This reveals that late lockdowns (when cases are rising considerably) are very effective at flattening the curve of COVID-19 spread.

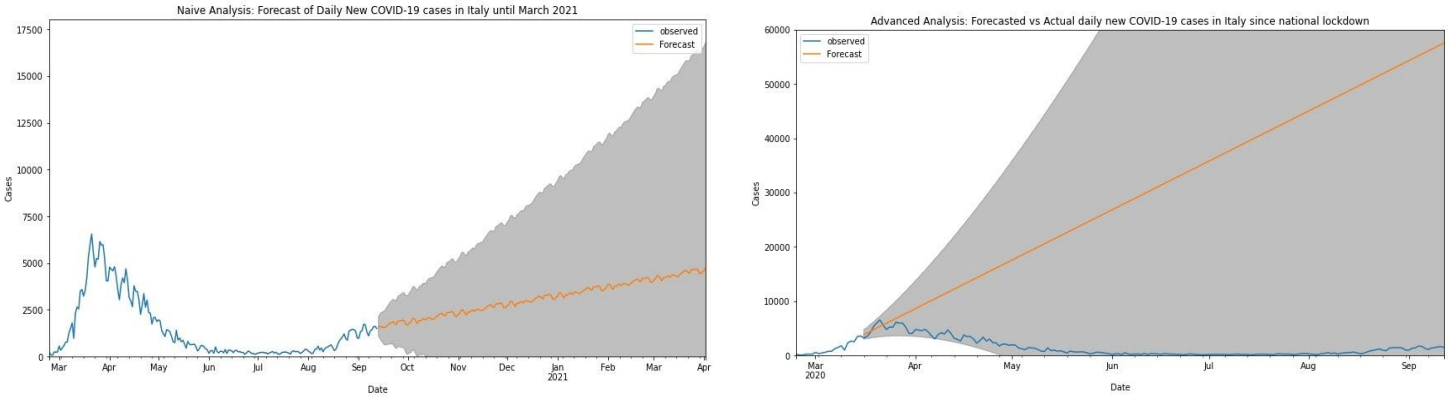

Figure 6: Italy Case Count - Naive (left) and Advanced (right) Analysis

India: In the case of India, as seen in the graphs below (Figure 7), the naive analysis tells us that the case count will continue to rise well into 2021. The advanced analysis, which forecasts case counts after the early lockdown in India (post 30 May, 2020) shows a rising trend. However, the actual cases post-lockdown were a lot higher than the forecasts. This reveals that early lockdowns are not that effective, and that they simply delay the rise of cases once the lockdown is ended.

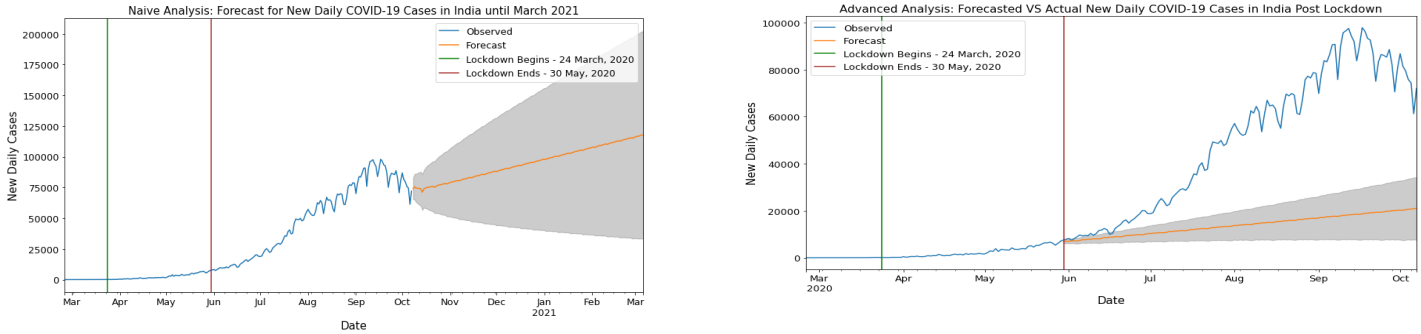

Figure 7: India Case Count - Naive (left) and Advanced (right) Analysis

South Korea: The naive analysis of South Korea data shows us that the cases will continue to rise and fall with monthly seasonality. For advanced analysis, we trained our data only up to 5 April, 2020. As we could not find the exact date when South Korea implemented contact tracing, we guesstimated this date based on the Infectious Disease Control and Prevention Act which took effect on 5 April, 2020. Our advanced analysis reveals that close contact tracing was very effective in containing coronavirus because the forecasted cases were higher than the observed cases.

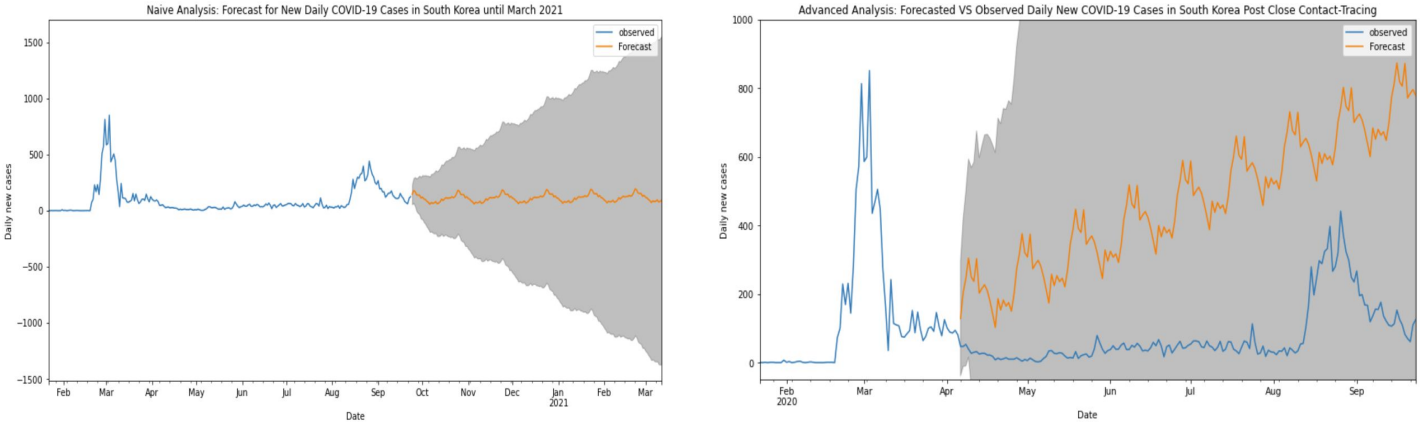

Figure 8: South Korea Case Count - Naive (left) and Advanced (right) Analysis

5.2 Daily Death Model

Similar to the daily case analyses, we used the SARIMA model to predict new daily deaths in all the countries. Again, we conducted two main types of analyses. First, the naive analysis trains on all the new daily deaths in the dataset, and predicts till 2021. This naive analysis helps us isolate how the overall death count will grow for each country. Second, the advanced analysis trains on all the new daily deaths in the dataset till the policy is implemented, and then forecasts into the future. Comparing this forecast with the actual death count helps us determine the efficacy of the implemented policy. For the specific insights from the two analyses on each country, see below:

Italy: The naive analysis shows that the overall daily death count will decrease quickly and hover around zero from October 2020 to March 2021. The advanced analysis shows that the late lockdown mandated by the Italian government was very effective at saving lives. If lockdown measures hadn't been implemented, we would have seen daily deaths increase linearly and reaching 400 daily deaths by September 2020 as shown by the forecasted yellow line. Thanks to the lockdown, daily deaths are actually much lower (between 0 and 10 in September 2020).

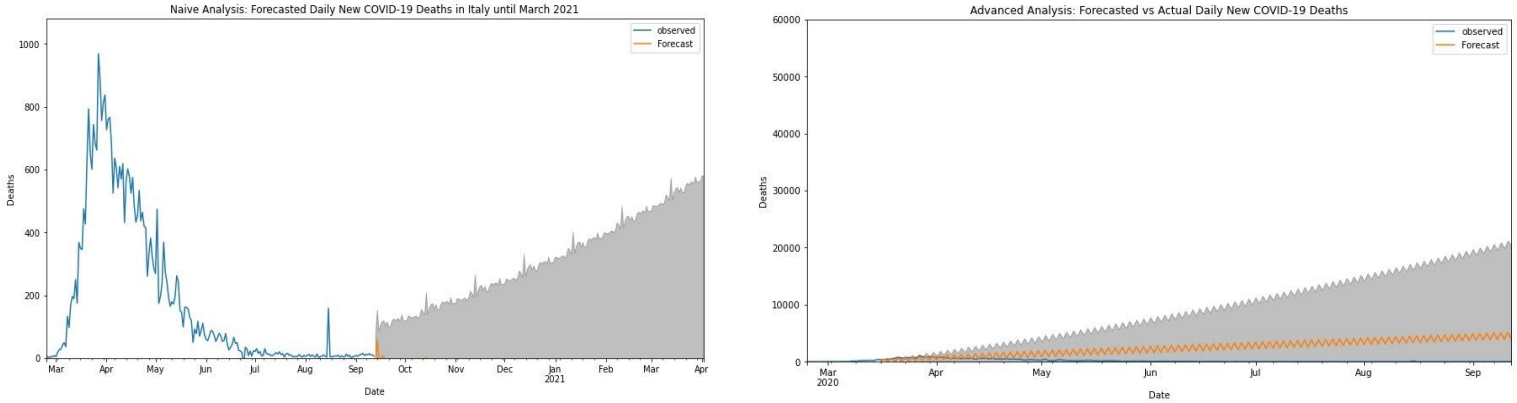

Figure 9:Italy Death Count - Naive (left) and Advanced (right) Analysis

India: As can be seen in the graphs below (Figure 10), the naive analysis shows that the overall daily death count will continue to rise well into 2021. Furthermore, the advanced analysis reveals that there have been a lot more deaths than forecasted post the lockdown. Together, these analyses suggest that the Indian early lockdown has not been as effective at controlling the daily death counts.

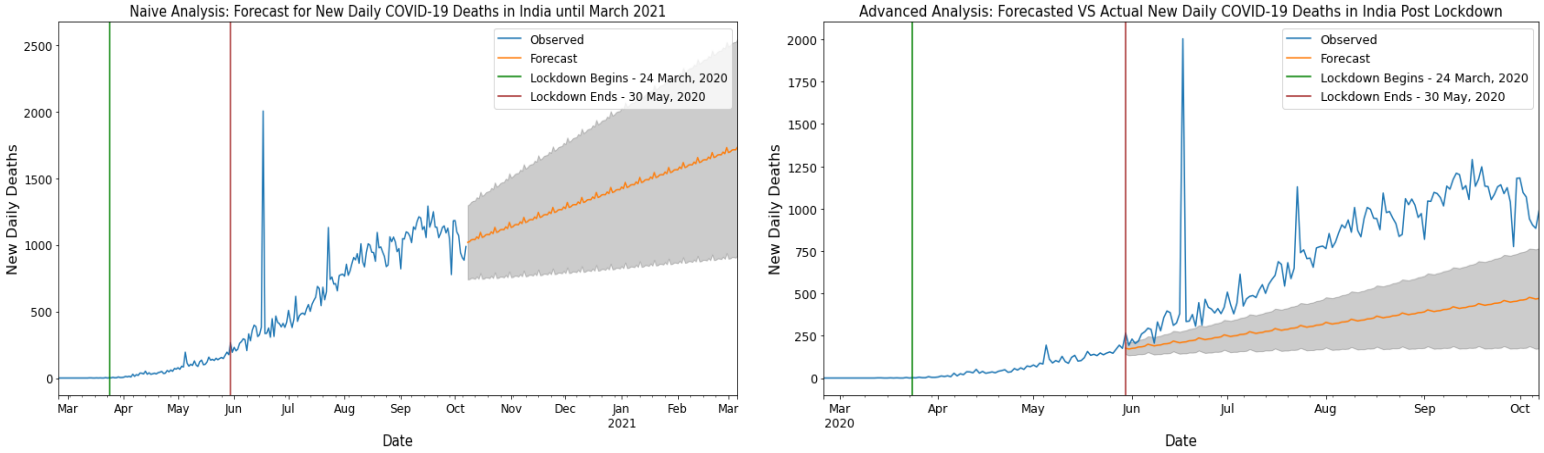

Figure 10: India Death Count - Naive (left) and Advanced (right) Analysis

South Korea: The naive analysis of South Korea reveals that the daily death count will continue to rise into March 2021 although the death count is very small. Our advanced analysis shows that early contact tracing adopted by South Korea was very successful at saving lives because the forecasted daily death count was higher than the observed daily death count.

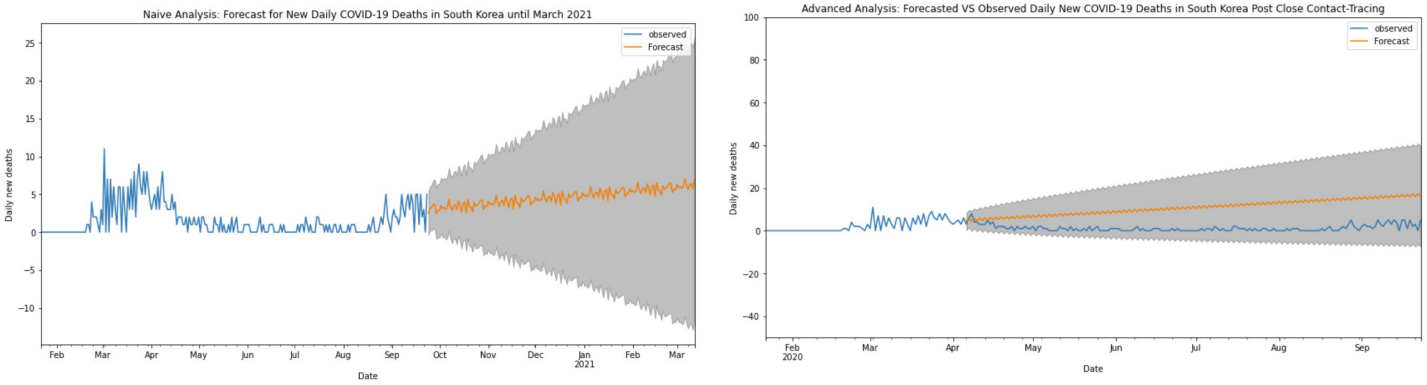

Figure 11: South Korea Death Count - Naive (left) and Advanced (right) Analysis

6. Conclusion

6.1 General Conclusions

Overall, our SARIMA model analyses suggest that adopting contact tracing and mandating lockdowns when cases are peaking are effective containment measures. We infer this by the fact that South Korea and Italy both observed a fewer actual number of cases than predicted by a SARIMA model trained before policies were enacted.

In contrast, our models suggest that an early lockdown is not an effective tactic. In India, the opposite phenomenon occurred where the observed number of cases was much higher than the SARIMA model, trained on data before lockdown, had predicted. This inefficacy was initially surprising but is logically coherent. We reasoned that an early lockdown may temporarily halt the initial spread of the virus but, once lifted, allows for uncontrolled spread once again.

6.2 Recommendations for Brazil

As we have seen in our prior modeling, lockdowns at the peak of cases and close contact tracing seem to be the two most effective strategies of the three. However, at Brazil's current point of COVID-19 spread, there has not yet been a lockdown, but both cases and deaths have steadily decreased since August 2020. This suggests that Brazil is likely already past the pandemic's peak. Our SARIMA model forecast supports this suspicion, as the model predicts a steadily declining trend into 2021.

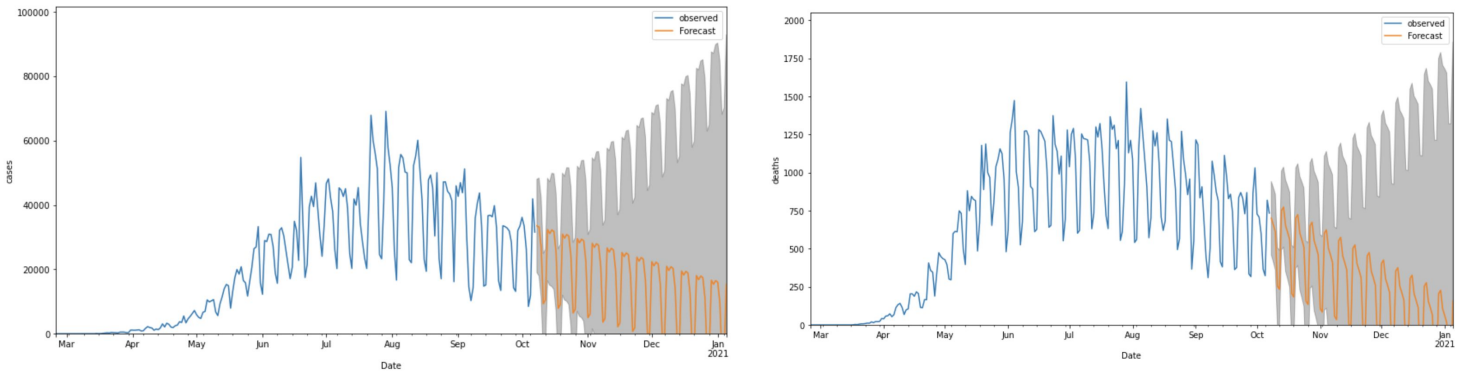

Figure 12: SARIMA Forecast of Cases (left) and Deaths (right) in Brazil

As Brazil's COVID spread is likely past its peak, where a late lockdown would be effective, we recommend administration to focus primarily on implementing close contact tracing to ensure cases and deaths continue to decline as forecasted.

7. Impact Statement

The immediate impact of our findings is clear: with Brazil in a major health crisis, an effective containment strategy could mitigate the current situation and save many lives. More broadly, though, these results can be generalized to the global understanding of effective pandemic response. In particular, these models indicate that the early lockdown strategy that was adopted by many regions, including some of the most populous U.S. states[5], may not be as effective as was once believed.

The human impacts of our data include potential unreported cases due to a lack of testing and medical resources. Some studies estimate that COVID-19 mortality counts in Brazil may be underreported by more than 40%[6], reflecting the broader socioeconomic impacts of the pandemic. As case and death metrics are likely underrepresented, the pandemic is probably even more severe in Brazil than our models indicate, making an effective pandemic response all the more crucial for the nation's future.

8. Limitations and Future Steps

Beyond the aforementioned underrepresentation of case/death counts, our analysis is limited by our omission of several important predictors, including social distancing, mask policies, and stringency of policy enforcement. In addition, we worked only to predict daily changes in cases and deaths while not taking into account other response variables that may be more pertinent to public health concerns, such as ICU capacity and population-adjusted metrics.

Should we revisit this project in the future, we would love to more closely examine the relationships between these additional predictors and response variables to recommend more potential pandemic response strategies. We would also like to take into consideration unreported cases/deaths by appropriately adjusting model biases to more accurately model the current pandemic situation. However, given the raw data we do have, our models are valid, and we stand firmly by our recommendations for Brazil.

9. References

Since time series analysis is beyond of the scope of this course, we relied heavily on the following resources to implement the 5 different time series models:

- https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-one/
- https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-two/
- https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[5]https://www.sfchronicle.com/bayarea/article/Bay-Area-coronavirus-decision-Behind-the-scenes-15148425.php
[6]https://www.jmir.org/2020/8/e21413/