INMR96 – Digital Health and Data Analytics

# Problem Analysis

## Context of the Problem, Challenge Opportunity or Issue

The world has been grappling with the severe acute respiratory syndrome, the novel Corona virus disease 2019 (COVID-19). The virus whose origin is suspected to be from the Huanan sea food market in December 2019 as Singhal, (2020) points out. Exponential rise in the number of cases in and beyond the epicenter of Wuhan to other parts of China, Thailand, Japan and South Korea countries led to suggestions being raised of there being possibilities of human-to-human transmission. The infection rate of the virus is high, with the number of those infected and those who succumbed given the early challenges of surveillance and testing being underestimated to a larger extent. According Kumar, Malviya and Sharma (2020) the virus can be spread through airborne zoonotic droplets when an infected individual sneezes or cough – some of the noticeable symptoms. (Bentout et al., 2020) implementation of mathematical models to provide understanding of the COVID-19 in Algeria. They assert the need for the inclusion of the models as one of the sources of information regarding the spread and how well government and institutions can plan ahead of time when provided the projected figures on confirmed cases. The incorporation of data health solutions as Kapoor et al., (2020) states can provide a framework for data collection, early COVID-19 virus detection and tracking, decision making and health sector planning and ensuring the strict adherence to safety guidelines put in place by relevant authorities.

## Objectives of the Study

Data used for the study entails records of COVID-19 confirmed cases, recoveries and death toll within daily and accumulation contexts. The data was analyzed to infer trends of the spread and extent of infection. Various statistical descriptions and models are imposed upon the data and conclusions made. Exploratory analysis on the data was conducted to check the spread of the various variables within the data. An assessment with regard to digital health solutions is provided in the context of the data.

## Limitations of the Study

The data is based solely upon counts of cases and does not encompass other variables of interest such as health sector situation in the country the data comes from, government, institutions and the general public initiatives that might influence the number of cases and cohort data that might provide information on the age cohorts within the population susceptible to the virus.

## Assumptions of the Study

It is assumed that the data is complete in its entirety and analysis conducted on it provide a true picture of the general population.

## Significance of the Study

### What are the goals of the Study

The study and analysis is conducted with an aim of understanding the spread of COVID-19, and propose models that would accurately provide prediction of the variables of interest. The emphasis is on the incorporation of statistical models within the health sector to provide sound evidence for decision making process.

### Significance of the Topic to Stakeholders

The results of the analysis of this study would be of significance to the general public, institutions and the health sector in general during the pandemic period as they try to find ways to curb the spread of the virus. Health professionals and the government will be able to incorporate the findings in the drafting of policies regarding COVID-19 and use the information in their decision making process.

### Broader Implications of the Topic

#### Industry implications.

Individual governments and their health care sector, with emphasis departments that come up with policies and implement them to counter the spread of COVID-19 undertake their operations and make decisions that are based on prior research to come up with policies that the general population can abide by.

#### Global implications.

Interest in the spread of COVID-19 within nations vary, as such localized assessment and analysis of data can also vary when other factors are put into consideration. The general assessment can be extended past the factors we consider in this study, putting into consideration location based factors among others.

## Information and Literature Review

### Brief Summary of the Literature on the Subject

Bentout et al., (2020) implement mathematical models on the daily reported new cases and the accumulated cases in Algeria. The system of ordinary differential equations - SEIR epidemic model they use encompasses the interactions between various states of the pandemic that include infection and recovery and or death. Moreover, they incorporate instances such as the contact rate, removal rate, onset rate, the average infectious and incubation period, identification rate, basic reproduction number and the population of Algeria. Their simulations provide prediction of COVID-19 cases along with model parameters through least square methods, with the best model based upon that which provides the least sum of squared residuals. They incorporate other aspects that would influence the number of cases such as the lockdown as they try and point out the extent to which such aspects would completely phase out the spread and number of cases reported. The recommendations made include the need for stricter compliance and early government interventions.

Liu et al., (2020) use the SEIR epidemic model on China data to predict the number of cumulative reported cases based on the reported cases while incorporating other aspects such as the implementation of policies by the government to curb the spread of the virus, the identification process and assessment of the impact the coronavirus has had. Their model does incorporate the time the pandemic began, number of individual susceptible to the virus, reported symptomatic and asymptomatic infectious individuals, unreported symptomatic and asymptomatic infectious individuals at various time stamps and their associated rates. Their assumptions is that cases are handles as soon as they are detected. Based on their findings, they ascertain the effects the various government policies and their restrictiveness have had on the COVID-19 cases in China.

### Systematic Review of the Literature

#### Summary and Relevance of the Literature to the problem or questions

The research outline the key features to incorporate into mathematical models when predicting the number of COVID-19 cases and how the locality of the study may also influence the suitability of the model. Herein, we are predicting the number of newly confirmed cases and death tolls per day.

#### Thematic Findings in Literature as Applied to the Topic

The number of new COVID-19 cases is of interest in policy making and assessing the effectiveness of the measures being put in place by the government as well as the general public regarding health.

# Descriptive Statistics

This section is undertaken to answer the following questions:

* What is the types and distribution of the various variables in the data?
* Are there anomalies and outliers within the data?
* What are the trends recurrent within the data?

## Type of variables in the data set.

The time series data comprises of 7 variables, the date column and records of various instances on the Covid-19. The data comprises of 146 records in 7 variables, with no missing data.

|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-Null Count | Variable Type |
| 0 | date | 146 | Date Time |
| 1 | newly\_confirmed | 146 | Numeric |
| 2 | cumulative\_confirmed | 146 | Numeric |
| 3 | existing\_confirmed | 146 | Numeric |
| 4 | cumulative\_cured | 146 | Numeric |
| 5 | new\_cured | 146 | Numeric |
| 6 | cumulative\_death | 146 | Numeric |
| 7 | new\_death\_toll | 146 | Numeric |

Table 1 Data Variables and Types

## Descriptive Statistics of Discrete and Continuous Variables:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| newly\_confirmed | 146 | 18617.13 | 20976.51 | 0 | 4158.5 | 10195.5 | 26616 | 123238 |
| cumulative\_confirmed | 146 | 3643008 | 871429.7 | 1723242 | 3129901 | 4064827 | 4317573 | 4424072 |
| existing\_confirmed | 146 | 3525280 | 843377.8 | 1658261 | 3040431 | 3935184 | 4178619 | 4274422 |
| cumulative\_cured | 146 | 9952.589 | 3373.661 | 3736 | 7199.75 | 11060.5 | 12524.5 | 14684 |
| new\_cured | 146 | 74.9863 | 69.46207 | 0 | 21 | 53.5 | 109.5 | 400 |
| cumulative\_death | 146 | 105942.8 | 23665.08 | 61245 | 82270.75 | 118564 | 126429.3 | 127517 |
| new\_death\_toll | 146 | 453.9178 | 481.2265 | 0 | 51.25 | 306.5 | 618.75 | 1915 |
| cases\_rolling\_avg | 140 | 19018.33 | 17492.95 | 2216.286 | 5348.25 | 11431.36 | 31885.68 | 64848 |
| deaths\_rolling\_avg | 140 | 466.4469 | 406.208 | 18.85714 | 68.5 | 419.9286 | 752.4286 | 1248.429 |
| Mortality Rate% | 142 | 2.481056 | 1.88676 | 0 | 1.14 | 1.93 | 3.275 | 8.6 |

Table Data Description

## Descriptive Graphs:

### Boxplots

The Boxplots of newly confirmed, newly cured and new death tolls of the daily reported cases depict higher means for the newly confirmed cases with a couple of outlier data points. The daily new death tolls have a higher mean than the new cured cases, with the new death tolls having outliers.

|  |  |
| --- | --- |
|  |  |

Table 4 Boxplots

### Line Plots

The plots Figure 1 cumulative confirmed cases vs cumulative cured cases and Figure 2 cumulative confirmed cases vs cumulative deaths depict a trend that is concurrently and generally rising for both cumulative cured cases and cumulative deaths as the confirmed cases are increasing by day.

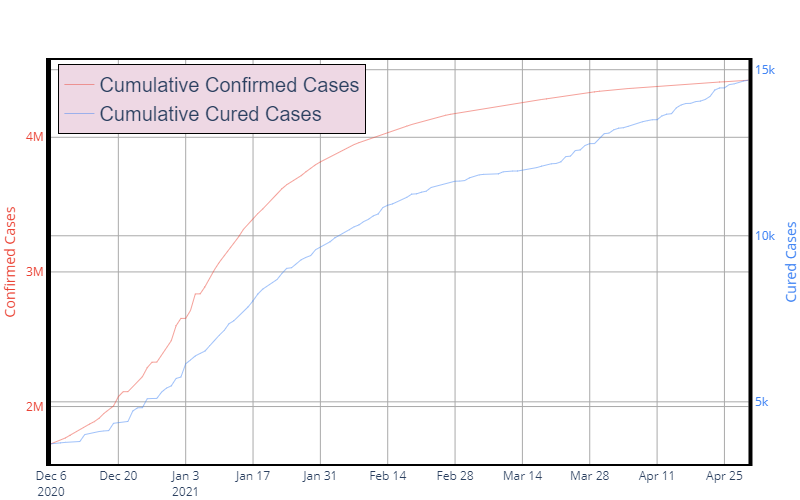


Figure cumulative confirmed cases vs cumulative cured cases

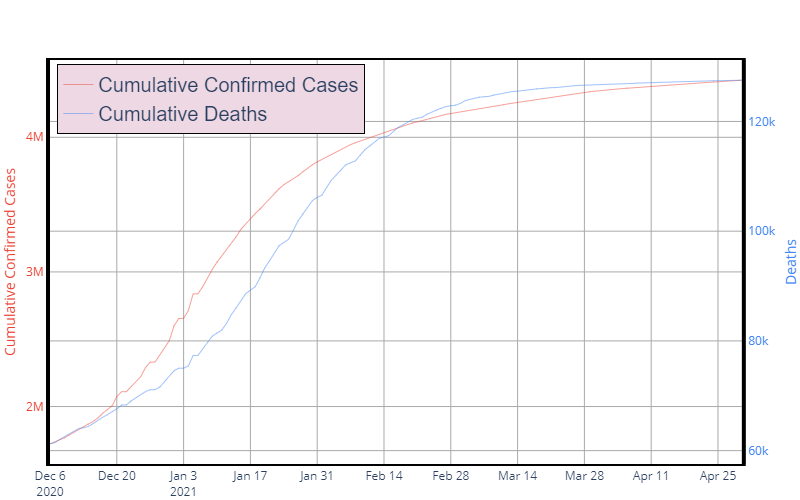


Figure cumulative confirmed cases vs cumulative deaths

A 7 day rolling mean of the data as depicted by the Figure 3 7-Day Rolling Averages show a general trend in reduction across both instances. Moreover, there was a slight variation on the mortality rate of the virus across the period of study as seen in Figure 4 Mortality Rates.

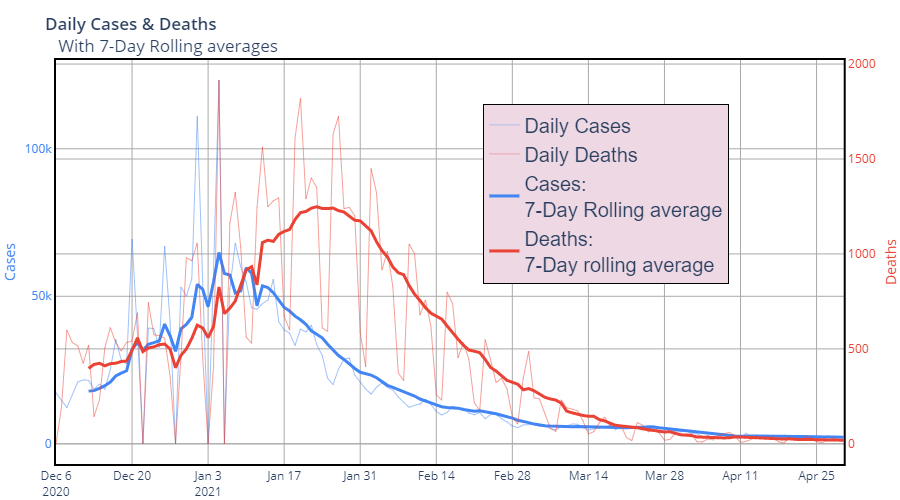


Figure 7-Day Rolling Averages

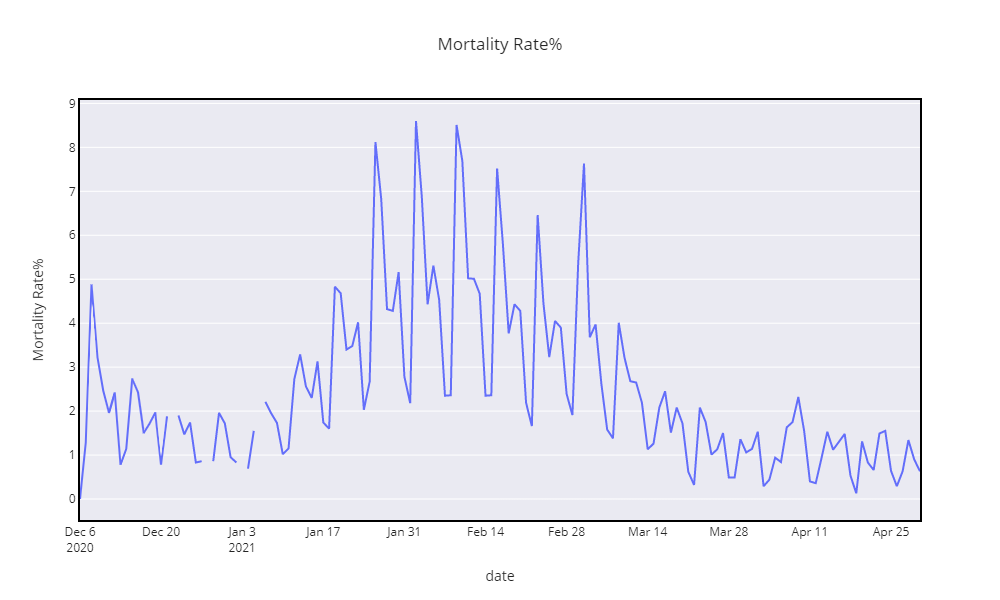


Figure Mortality Rates

## Data Preparation

The data was based upon the date aspect to carry out predictions. The prediction data possesses the format depicted in Table 4 Data format. The new variables extracted from time will be used as the predictors for the new confirmed cases and new death toll per day.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | day | month | dayofweek | dayofyear | quarter | weekofyear | newly\_confirmed | new\_death\_toll |
| 12/6/2020 | 6 | 12 | 6 | 341 | 4 | 49 | 17271 | 0 |
| 12/7/2020 | 7 | 12 | 0 | 342 | 4 | 50 | 14718 | 189 |
| 12/8/2020 | 8 | 12 | 1 | 343 | 4 | 50 | 12281 | 599 |
| 12/9/2020 | 9 | 12 | 2 | 344 | 4 | 50 | 16578 | 533 |
| 12/10/2020 | 10 | 12 | 3 | 345 | 4 | 50 | 20964 | 516 |

Table Data format

# Analytical Methods

## Analysis of the Literature Review Research Findings

The question being sought if the various features derived from the date column can be able to adequately predict the number of newly confirmed cases and number of deaths. All independent and dependent variables were numerical in nature. An XGBRegressor model was implemented on the data based on initial assessment of the same model under a TimeSeriesSplit that took into consideration the time series nature of the data. The independent variables were day, month, dayofweek, dayofyear, quarter and weekofyear that were extracted initially during preprocessing while the dependent variables modelled independently were newly\_confirmed and new\_death\_toll. The evaluation metrics are detailed in Table 5 Metrics.

|  |  |
| --- | --- |
|  | score |
| explained\_variance New Confirmed | 0.001914 |
| explained\_variance Death | -1.20841 |
| MAE New Confirmed | 369.9372 |
| MAE Death | 2383.337 |
| R\_2 New Confirmed | -0.16206 |
| R\_2 Death | -5.75698 |
| MSE New Confirmed | 206140 |
| MSE Death | 835.3484 |
| RMSE New Confirmed | 454.0264 |
| RMSE Death | 2418.96 |

Table Metrics

# Results and Evaluation

The explained variances for both models were close to 0 and the r-squared values were negative for which we can categorically state that our model did not perform well. The Figure 5 New Cases and Figure 6 Deaths provide a comparison of the predicted values against the actual values, for which we can note there were minor deviations that due to scale may have influenced the accuracy of the model.

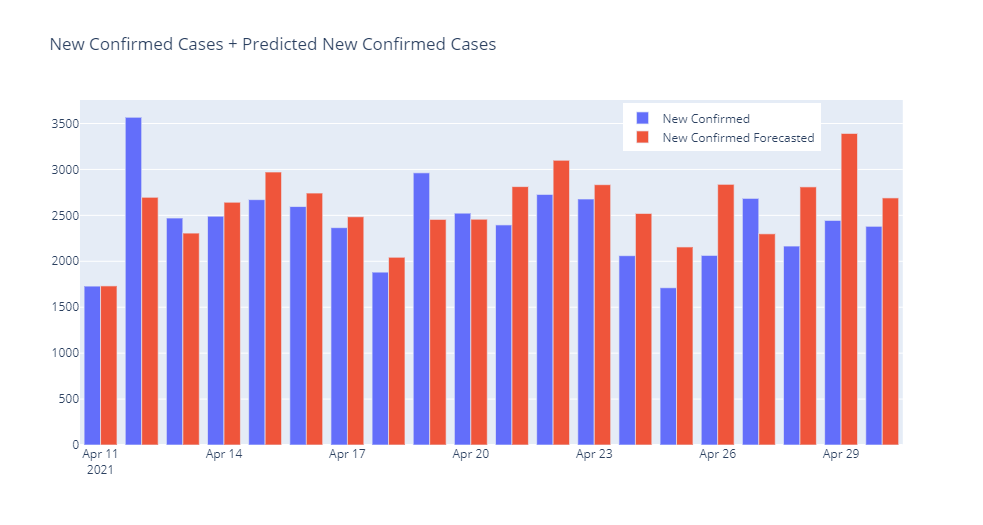


Figure New Cases

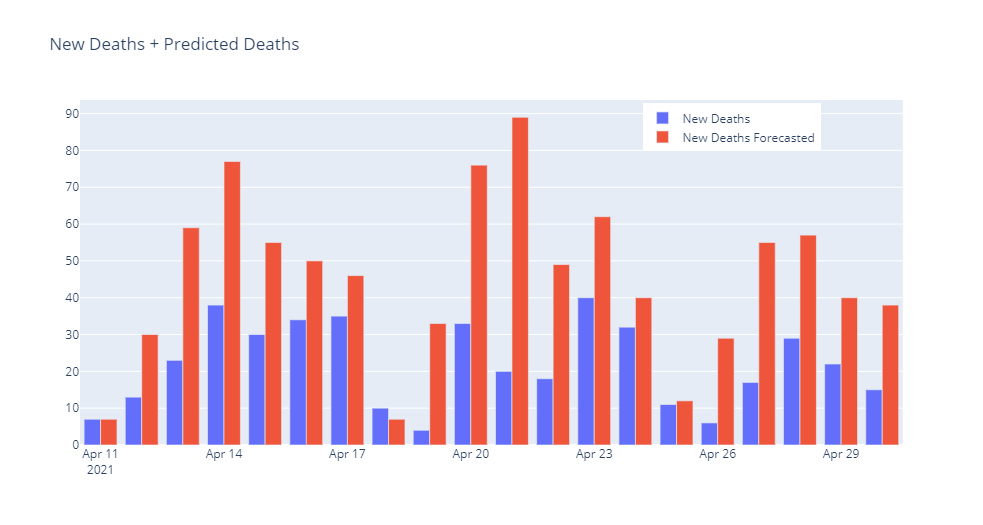


Figure Deaths

The prediction of new deaths and new cases per day as per the results show that there exists a small bit of pattern within the data used that would be able to forecast the COVID-19 cases. We can assume that the inclusion of more variables that we may think have an impact on the overall trends observed in the COVID-19 cases.

# Solution Development

The digital health solutions aspects include the provision of access for data electronically to practitioners and patients, monitoring of health metrics remotely, ease of collection of records through the use of tech such as wearables, mobile apps and the incorporation of analytics within the solutions. Sust et al. (2020) appreciates the role mathematical models play in predicting of scenarios around the COVID-19 while highlighting digital strategies that have been implemented in Catalonia to curb the spread of COVID-19. Some of the strategies they point out include citizen registration, virtual visit system expansion for primary and specialized care of patients, mobile health app for citizens’ self-assessment monitoring of patient progress and the incorporation of data analytics for policy formulation and health care system planning. Our case study’s implementation of digital solution strategies would have come in handy to provide patient data such as the age cohorts and identify the COVID-19 trends, up-to date figures on the cases and extent of spread of the epidemic as well as curb the spread through ease of access of health services virtually. The health system would be strengthened when provided predictions of the epidemic by planning ahead of time. The various aspects that would play a critical role in mathematical epidemic models as in the study by (Bentout et al., 2020) such as early detection and management of cases through self-assessment on mobile based applications would come in handy. The influence of other diseases on the mortality rate of COVID-19 would be considered given the availability of the data. However, the strategies would be dependent upon the health care system and the willingness of the general public to register in the implemented platforms.

Based upon Sust et al. (2020) findings, the adoption and success of digital health solutions may be subject to financial constraints, ethical issues around patient data and metrics to assess its progress. Moreover, they point out the reduced patient visits and overcrowding of health facilities as effective in prevention of human to human transmission of the virus. The e-health system’s implementation, adoption and full utilization as researched by Ross et al., (2016) study are influenced by factors that include adaptability of the technology within the localities, complexity with regard to use, modification, synchronizing of data, real-time access and reliability, associated cost of setup, maintenance and revenue, policies and legislation. Moreover, the engagements of stakeholders in development and implementation, training of staff on the use and skill barriers were also stated. As it is within our case study, deeper understanding of stakeholders and their engagement would influence the use of digital solutions within the locality while also providing complementary solutions through data analytics to counter the spread of COVID-19.

# Conclusions

There are various factors that influence the count of the number of COVID-19 cases and deaths within the sample population used. In this study, we have undertaken analysis to uncover trends, modelled regression to assess the predictability of COVID-19 cases. The limited data used did not provide a better model that would effectively predict future number of cases.

## Specific Recommendations and Suggestions for Future Research

The study can further be undertaken while incorporating the data on various aspects of epidemic modelling to statistically determine their influence on the cases and provide more robust predictive models. Limiting of the variables would fail to capture various statistically significant factors that would have changed the results. Furthermore, full data from the onset of the epidemic time, the number of individual susceptible to the virus, the reported symptomatic and asymptomatic infectious individuals, unreported symptomatic and asymptomatic infectious individuals at various time stamps, their associated rates and results from implementation of policies can offer more predictive power to our model. The use of mathematical models that are well suited for epidemic modelling would very well accommodate the recommended aspects (Bentout et al., 2020). Adoption and implementation of the digital health solutions within our case study’ locality would provide more data that would be analyzed to offer better insights for governments and health service providers in their fight against COVID-19.

## Ethical considerations

The collection of case data is important for policy formulation by the relevant bodies. It is imperative to educate health practitioners, the general public on the need for up to date records and the capture as much data on the area of interest to suitably model and make better policies for the development of the entire healthcare system and phasing out of the virus.

# References

Bentout, S., Chekroun, A. &amp; Kuniya, T., 2020. Parameter estimation and prediction for coronavirus disease outbreak 2019 (COVID-19) in Algeria. AIMS Public Health, 7(2), pp.306–318.

Kapoor, A., Guha, S., Das, M.K., Goswami, K.C. and Yadav, R., 2020. Digital healthcare: The only solution for better healthcare during COVID-19 pandemic?. *Indian Heart Journal*.

Kumar, D., Malviya, R. and Sharma, P.K., 2020. Corona virus: a review of COVID-19. *EJMO*, *4*(1), pp.8-25.

Liu, Z., Magal, P., Seydi, O. and Webb, G., 2020. Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data. *arXiv preprint arXiv:2002.12298*.

Ross, J., Stevenson, F., Lau, R. and Murray, E., 2016. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). *Implementation science*, *11*(1), pp.1-12.

Singhal, T., 2020. A Review of Coronavirus Disease-2019 (COVID-19). *The Indian Journal of Pediatrics*, 87(4), pp.281–286.

Sust, P.P., Solans, O., Fajardo, J.C., Peralta, M.M., Rodenas, P., Gabaldà, J., Eroles, L.G., Comella, A., Muñoz, C.V., Ribes, J.S. and Monfa, R.R., 2020. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR public health and surveillance, 6(2), p.e19106.

# Appendices

## Appendix A: Data Set

Data Set is provided within the accompanying xlsx file.

## Appendix B: Code

Code snippets in accompanying python script and notebook