EAS509 SDM Project
Project Title : COVID-19 DATA ANALYSIS

VAMSHI ADI, University at Buffalo
HAO WANG, University at Buffalo
AJINKYA ATHLYE, University at Buffalo
JASHAUL DIWAKAR, University at Buffalo
KULJOT SINGH CHADHA, University at Buffalo

ACM Reference Format:
Vamshi Adi, Hao Wang, Ajinkya Athlye, jashaul Diwakar, and Kuljot Singh Chadha. 2022. EAS509 SDM Project Project Title : COVID-19 DATA ANALYSIS. 1, 1 (May 2022), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 KEYWORDS

COVID-19,Time Series Prediction/Analysis,Survival Analysis,Visualization

2 ABSTRACT

The continuing COVID-19 epidemic has wreaked havoc on the global economy, causing governments to take drastic steps to stem its spread. Being able to examine the patterns underpinning the pandemic allows people to take precautions accordingly. This study looked at the time-series analysis and survival analysis of the data. We took into account the pandemic data collected by Organizations such as WHO, CDC that keeps track of the COVID cases globally overtime as well as hospital data that records various information of their patients for survival analysis. The sample of time series data was gathered till March 2022, and the prediction and analysis were completed until August 2022. The outputs of our analysis generally conform to the real scenario of the pandemic. We studied and analysed the various patterns associated with the collected data including prediction on the number of deaths,recovered,active and confirm cases. Applied time Series Analysis to understand the underlying causes of trends or systemic patterns over time. Survival Analysis is done to investigate the impact of COVID-19 and the risk factors associated with COVID-19 deaths. We also applied general statistical analysis on the data to extract main statistics about the data.

3 INTRODUCTION

3.1 Motivation Significance

Coronavirus disease 2019 (COVID-19) is a contagious illness caused by the coronavirus 2 (SARS-CoV-2), which causes severe acute respiratory syndrome. The first instance was discovered in December 2019 in Wuhan, China. The disease

Authors' addresses: Vamshi Adi, University at Buffalo; Hao Wang, University at Buffalo; Ajinkya Athlye, University at Buffalo; jashaul Diwakar, University at Buffalo; Kuljot Singh Chadha, University at Buffalo.

Manuscript submitted to ACM

quickly spread over the world, resulting in the COVID-19 pandemic. We are all aware of and affected by COVID-19. More than 419 million cases have been documented throughout 188 nations and territories over the globe, with over 5.8 million deaths; more than 400 million patients have recovered. In response to this ongoing public health emergency, we will study and analyze the various patterns associated with COVID data, including projecting the rate of recovery and other variables pertaining to the deaths, active, and overall recorded cases of COVID. To be specific we will apply Time Series Analysis to understand the underlying causes of trends or systemic patterns over time and the cases in the delta shift of cases. Additionally, we could identify the factors to cause in the delta change of the number of cases in relations to decisions/rules enforced across various regions from country specific to state/county specific. The goal of our project is to apply statistical learning techniques on the COVID-19 data and gain some meaningful insights.

4 DATA SOURCE

The data set we are working with is an aggregated data put together by Center for Systems Science and Engineering at John Hopkins.

- World Health Organization[1]
- European Centre for Disease Prevention and Control[2]
- US Center For Disease Control and Prevention[3]
- News[4]

  WorldoMeters[4.1]

  1Point3Arces[4.2]

  Covid Tracking Project[4.3]

  Los Angeles Times[4.4]

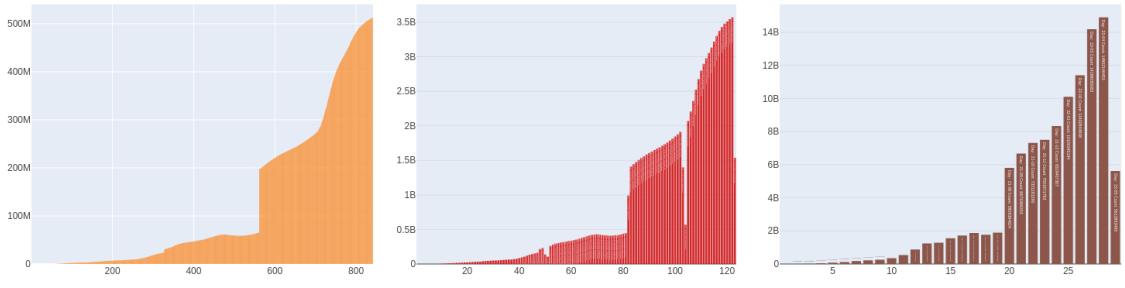
  The Mercury News[4.5]

5 METHODS

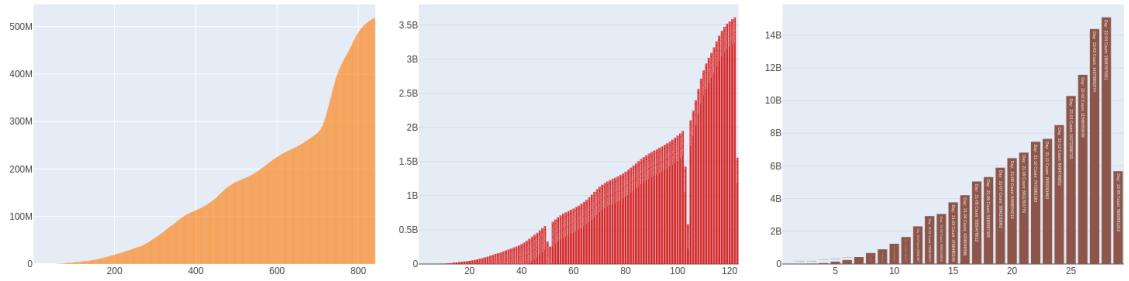
5.1 Data Trend Overtime

Visualizations are done on the COVID time-series data to generate treemaps of each country's confirmed/active/deaths cases, geographical scatter plot of global stats on a world map, line plots encompassing different COVID data over the past few years, etc. The history, trend, and the latest status of the pandemic can be examined through visualization graphs. The following Graphs shows the Daily, weekly and monthly analysis of the Cases for Active, Confirmed, Death Recovered records.

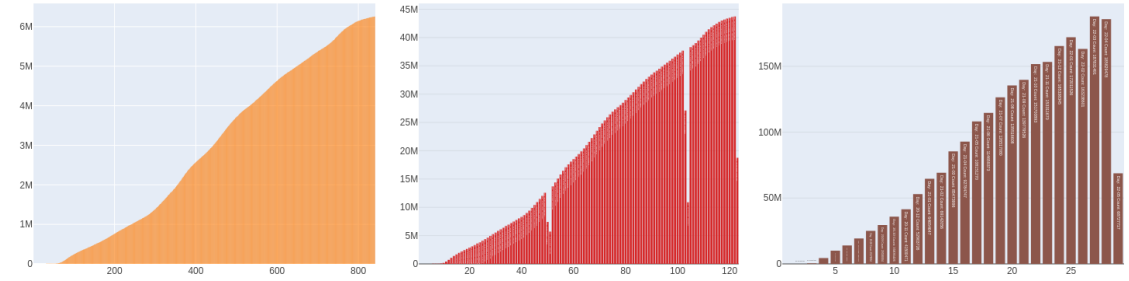
Active Cases:



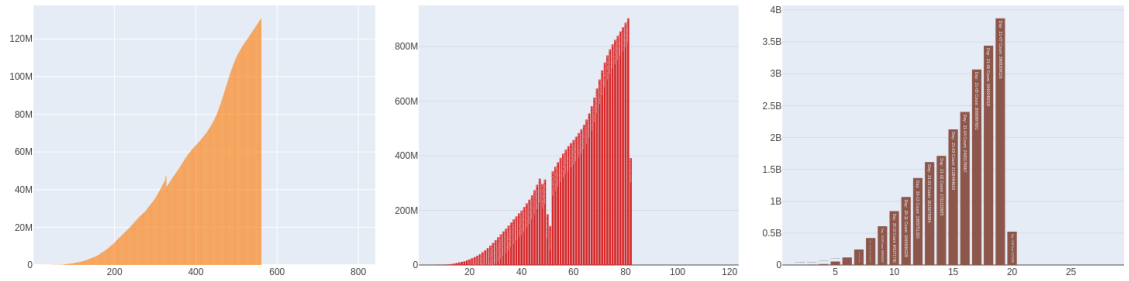
Confirmed Cases:



Death cases:



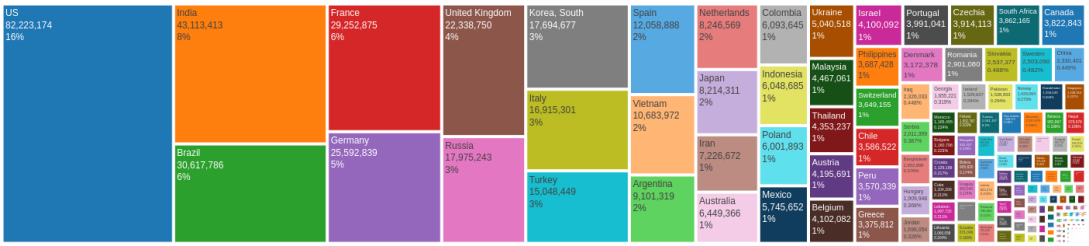
Recovered Cases:



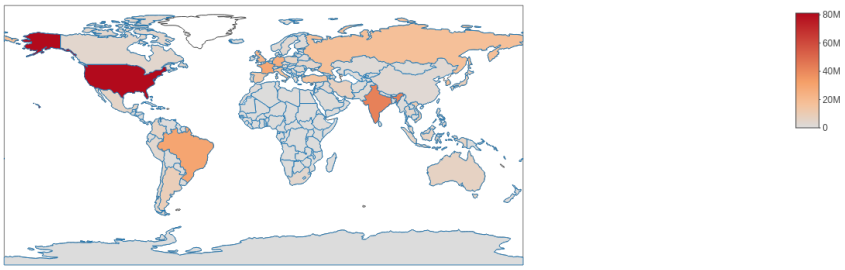
The above visualizations show the trend over various time periods and the main goal is to observe if the graph has a smooth edge i.e if the progress is gradual or if there is a spike. For instance seeing the daily statistics we cannot see any discrepancies but from monthly or even weekly statistics we can clearly see spikes be it increase or decrease, which actually correlate to some event which happened during the real time as well like the 2nd wave and 3rd wave.

5.2 Covid 19 Cases over The globe Visualization

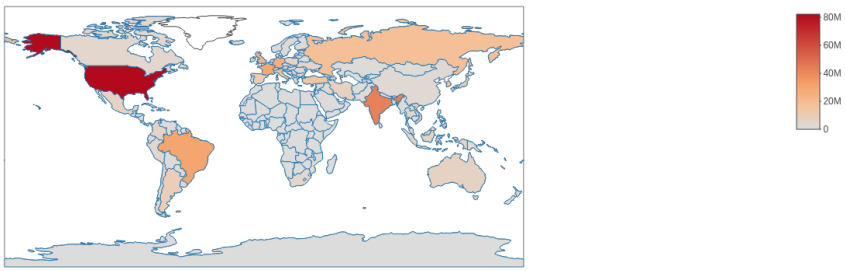
Treemap:



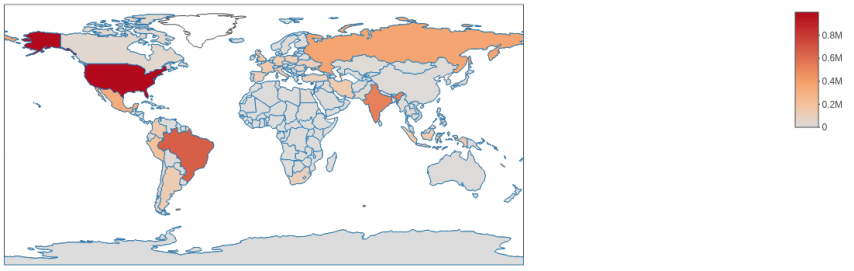
Active Cases:



Confirmed Cases:



Death cases:

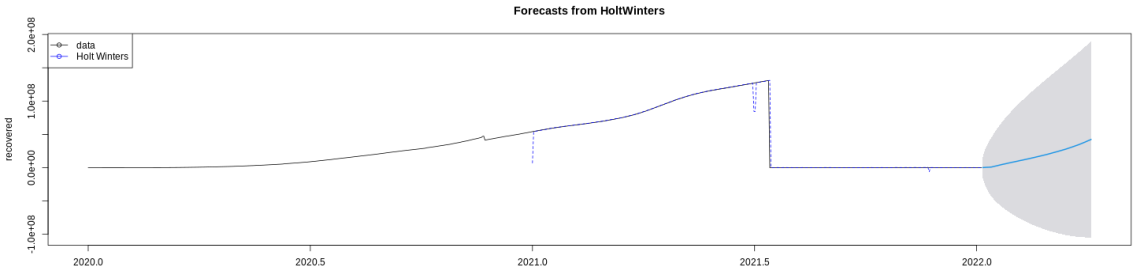


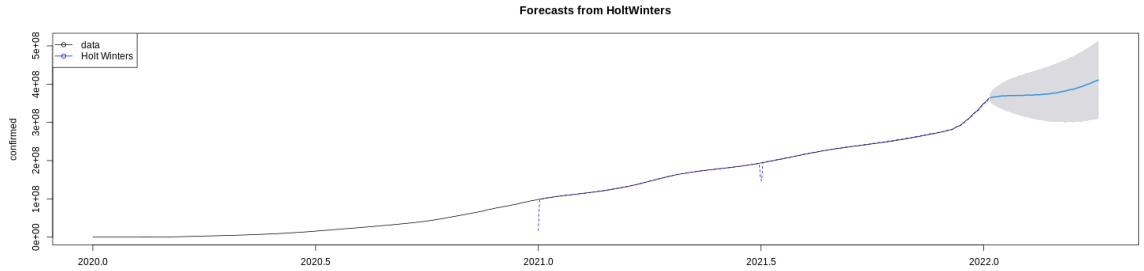
Recovered Cases:

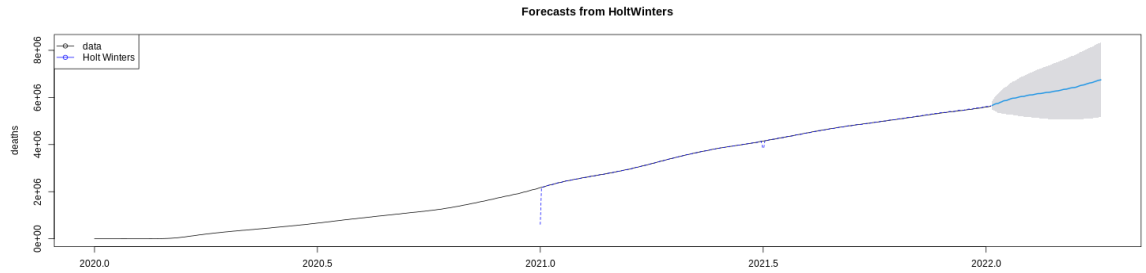
For the global pandemic visualization two types of graphs are used: Treemap and Map visualization.

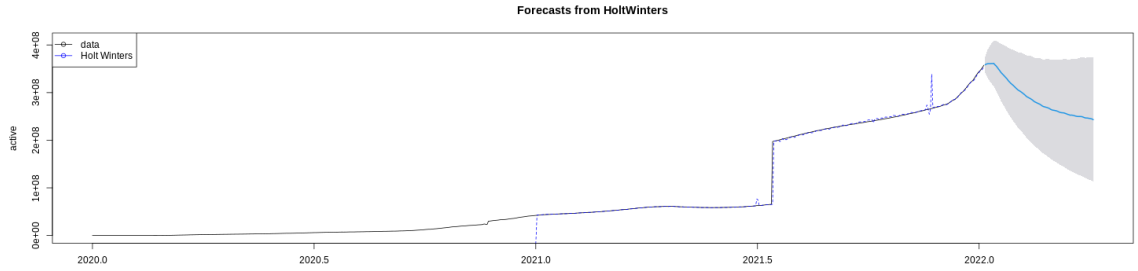
- Treemap: This is a visualization whose main purpose is to show the share of covid cases in comparison to other countries around the world.
- Map Visualization: This is a visualization which complements the previous treemap and displays the chunks in the treemap with a darker shade on the map. For instance US, Brazil and India have the majority of share in the deaths according to the map, the same can be noticed shown in the map visualization where these three countries are shaded in a darker color.

5.3 Holt Winter Analysis



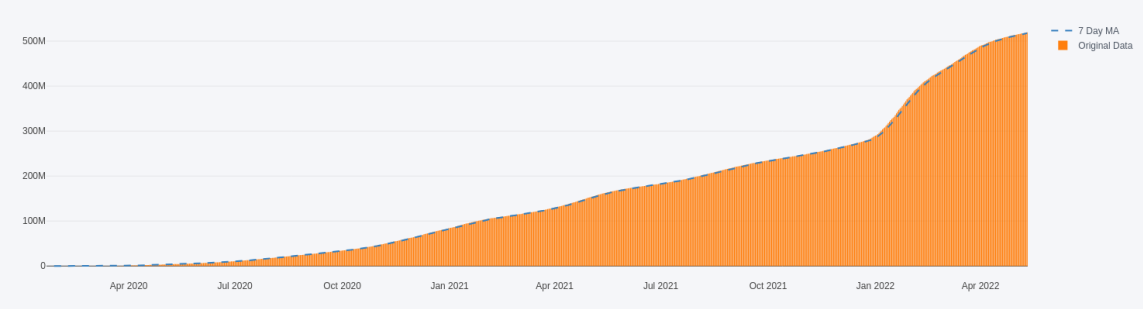






For prediction analysis, Hold-Winters algorithm is used to predict the time-series data. Since there are 4 types of cases covered, there are 4 visualizations and HW models for the Active, Confirmed, Death, and Recovered cases respectively. Additionally, there is a fifth model that was run on the difference between two consecutive days of the confirmed cases. This signifies the daily increase in cases. This was inspired from the fact that the AutoCorrelation graph for confirmed cases showed a heavy correlation between itself and its 1-time-lag, hence depicting some seasonality per unit time. From the visualization, it is evident that Holt-Winters did a good job of predicting a general trend but as we can see from the predictions made in the future, it fails quite significantly. This is to be expected due to the erratic nature of how the pandemic has been so far – which implies that sudden spikes (the second wave, new variants such as Omicron or Delta, etc.) and sudden troughs (due to vaccinations) have different features associated with them if we were to predict them. Univariate time-series forecasting would not be able to capture these occurrences.

5.4 7-Day Moving Averages



This is a visualization to overlay the moving averages on top of the original time-series data, to identify any spikes if present.

5.5 R-Shiny Application

We have implemented a COVID Insight Dashboard using R-Shiny with 2 main tab-ed layouts, Data Exploration Tab and Forecast Tab. The Data Exploration Tab shows insightly graphs of the covid data with respect to Type, Country, Date and Time. The Forecast Tab displays future trends which are predicted using the historically labeled data. We have implemented Holt Winters for the prediction of daily increase, active trend, confirmed trend, recovered trend and deaths trend. The Data Exploration Tab accepts dynamic input from users and updates the graphs displayed such as the treemap, Map Visualization, Single Moving Average Visualization, Time Based Visualization (Daily, Weekly and Monthly trend) and also Holt Winters forecast of future trends. These graphs are dynamically generated based on user input and accept values “active”, “confirmed”, “recovered”, and “deaths”.

6 RESULTS

- Visualization and exploratory analysis of COVID data. Through visualizations we were able to get a big picture of the pandemic over the globe as well as the pandemic trends in specific countries. Our findings through

the visualization graphs overall conform to the real world scenario of the pandemic such as the US, Brazil, Russia are the most affected countries, the starting point of the outbreak of the pandemic is in March 2020, the pandemic shows a steadily increasing trend, etc.

- Time-Series Decomposition and Holt-Winters prediction on the dataset. It is difficult to find seasonalities and long-term trends as well as perform predictions for daily new cases due to the nature of the dataset: time span not long enough, inconsistent data collection (hard to count exact number of new cases, counts of new cases purely based on patients reporting themselves), COVID affected by many real-world factors outside our control such as sudden waves, vaccination rollouts, new variants, etc. Therefore we cannot guarantee the usability of our predictions. However, based on our prediction, the daily increase in COVID cases will continue to rise in the US, with no turning point in near future.
- Survival Analysis to examine the risk factors associated with one's chances of survival after being affected by COVID. For this part, we are using the Data Science for COVID-19 (DS4C) dataset which offers a rich set of information (mainly from Korea and Singapore) on patient demographics, locations and weather conditions, as well as data on cities and regions where these infections are observed. In our results we found that the Proportional Hazard for age flattens as patients get much older. We also found that the variable "elderly population ratio" plays an important role in the model, with patients in cities with elderly population ratios of 1 having a 1.1970 time more likely chance of dying.

7 CONCLUSIONS

- Through the data visualized in the graphs it can be seen that the pandemic trend for total confirmed, daily increase, deaths, recovered cases are increasing steadily, and are likely to continue increasing in the near future.
- Our Holt-Winters prediction model were tested to produce results that match the real situation, and we predicted the pandemic until August, 2022. The prediction shows that the confirmed cases will keep increasing unwaveringly until August with not turning point.
- Our survival analysis shows that age is the primary risk factor that leads to COVID deaths.

8 REFERENCES

1 https://www.who.int/
2 https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
3 www.cdc.gov/coronavirus/2019-ncov/index.html
4 [4.1]https://www.worldometers.info/coronavirus/
[4.2]https://coronavirus.1point3acres.com/en
[4.3]https://covidtracking.com/data
[4.4]https://www.latimes.com/projects/california-coronavirus-casestracking-outbreak/
[4.5] https://www.mercurynews.com/tag/coronavirus/