COVID-19-Data Analysis Using Elastic Search

Authors: NATHAN KWOK,AN NGUYEN,TYLER TRINH,HANS ERON VALDERAMA
Department of Information Systems, California State University, Los Angeles
CIS 4560-01 Introduction to Big Data
e-mail: nkwok2@calstatela.edu, anguye219@calstatela.edu, hvalder@calstatela.edu, and ttrinh5@calstatela.edu

Abstract: In this project, we have chosen a COVID-19 data set and the data that is collected in the USA. The data that was collected by CDC for the cases of COVID-19 which is sorted by States and Date. While the coronavirus lasted from 2019-2023 many people are still concerned with this disease as it has caused lots of trouble for the citizens as many were forced to stay home and not have physical interactions outside isolation at home. In this project we would like to figure out which states had the most cases of COVID-19in each year. We also want to figure out which cities had the least cases each year. With this dataset, we would want to see which State and Age as well as the rate to see what is the most common place and age range of getting COVID-19.

1. Introduction

In this project we utilize the open source software, elastic search to process the dataset which contains valuable information within the years of covid-19.While the database is constantly being updated conclusively the dates that will be set forth will be March 7th 2020 to October 12th of 2024. The COVID 19 dataset by CDC includes data that points to State,Searson,Weekend Date,age category,sex label,race,type,weekly rate and cumulative rate.

We have chosen this topic and dataset because of the impact Covid-19 had a great impact on our nation. While many from young to old have all experienced challenges with a crude rate of cases.

2. Background

The COVID-NET Surveillance System provides weekly data on lab-confirmed COVID-19 hospitalizations across the U.S. This system gathers information from over 250 hospitals in 14 states to capture trends in hospitalization rates, stratified by age, sex, race, and underlying conditions. It aids in monitoring the pandemic's impact and identifying vulnerable populations, offering critical data to guide public health decisions and policy-making efforts throughout the pandemic response.

There are other website we looked at and study the number of Covid-19 cases such as:

- Worldometer is a real-time statistics website that provides data on a wide range of topics, including population, economy, health and the environment. For coronavirus (COVID-19), Worldometer offers detailed and frequently updated information.
- The CDC's COVID Data Tracker is a comprehensive resource providing regularly updated data on COVID-19 trends in the United States. It includes information such as test positivity rates, emergency department visits, hospitalizations, deaths, and vaccination statistics. The platform offers detailed dashboards and interactive tools to explore data by demographics, location, and other factors.
- The Los Angeles County Department of Public Health provides weekly updates on COVID-19 trends. Current metrics include a 7-day test positivity average of 2.2% and low hospitalization rates. Data encompasses cases, emergency department visits, deaths and vaccination efforts. Wastewater surveillance is also utilized for early detection of viral trends. Updates reflect ongoing improvements in public health response and are subject to periodic revisions as additional information becomes available.
- The WHO Health Emergencies Programme plays a critical role in managing global health crises, including the COVID-19 pandemic. It coordinates preparedness, surveillance, containment, treatment and research efforts globally.
- The California Department of Corrections and Rehabilitation (CDCR) tracks COVID-19 data for its incarcerated population through a public tool that provides information on testing, positive cases, and outcomes across facilities. The system, created in partnership with California Correctional Health Care Services (CCHCS), ensures transparency in COVID-19 management within state correctional institutions.

3 . Related Work

Related work #1: A decision support system for demand management in healthcare supply chains considering the epidemic outbreaks: A case study of coronavirus disease 2019 (COVID-19)

This study started in late 2019 on cases of Covid-19 happening around the globe. They find that places that lack equipment to encounter unknown viruses are prone to have dramatic increases in the number of people infected.

Related work #2: COVID-19 open source data sets: a comprehensive survey

This article was written in late 2020, almost a year after the COVID-19 pandemic arose. The paper utilizes open-source datasets from various sources such as the statistics from the World Health Organization (WHO). In

determining and diagnosing whether a disease would be caused by COVID-19, there is the use of CT scans, X-ray images, and cough sounds that detect the virus. Analysis of epidemiological, demographic, and the mobility data is adopted to track the virus and where it is mostly spread. The survey was conducted of online and open-source datasets, which consisted of a patient's age, sex, PCR status and their indications. The article also consisted of the use of Artificial Intelligence and Machine Learning with the help of a Neural Network to aid in identifying the medical images of patients. Hence, the data is transformed and analyzed to show future results. Compared to our own research and dataset, we would also use a similar type of data where we generally count how many diseases are in an area or state. Using the data, we inserted it into ElasticSearch and used the machine learning feature of Kibana to figure out the regression of the data. Our dataset and analysis focused more on the number of cases caused by COVID-19 through counting data received from hospitals, while the article mentioned here examined the internal workings within the hospitals.

Related work #3: COVID-19 Incidence and Death Rates Among Unvaccinated and Fully Vaccinated Adults with and Without Booster Doses During Periods of Delta and Omicron Variant Emergence

In this report written by Johnson et al. (2022) the authors examine the death rates of both vaccinated and unvaccinated individuals during the COVID-19 variants in 2021. This report also shows a visualization of a table that has information on specific time periods of unvaccinated or vaccinated death rates associated with COVID-19 such as July-November of 2021 or April-May of 2021 along with the total number of deaths. Additionally, one of the figures shows the trends of cases and deaths on a weekly basis comparing patterns between vaccinated and unvaccinated individuals. Our dataset focuses on a specific time frame, however in contrast, we want to explore a wider range of factors associated with COVID-19 such as age, demographic characteristics and geographic locations compared to the research that was only on vaccinated and unvaccinated individuals in 2021.

Related work #4: Age and sex-specific risks of myocarditis and pericarditis following Covid-19 messenger RNA vaccines

Cases of myocarditis and pericarditis have been linked to COVID-19 mRNA vaccines, with risks being highest within the first week following vaccination, particularly after the second dose, as shown by an analysis of hospital and vaccine data from France between May and October 2021. This study identified 1,612 myocarditis and 1,613 pericarditis cases, noting significantly increased odds ratios for myocarditis of 8.1 for the Pfizer-BioNTech (BNT162b2) vaccine and 30 for the Moderna (mRNA-1273) vaccine, especially among males aged 18 to 24. These conditions, often presenting as mild and resolving after short hospital stays, were more common in males for myocarditis but showed wider demographic impacts for pericarditis. The study underscores the importance of understanding vaccine-associated risks across age and sex groups, given the continued expansion of vaccination campaigns and the need to balance these rare adverse effects against the benefits of vaccination in preventing severe COVID-19 outcomes.

4. Our Work

The dataset that we processed through Elastic Cloud was done successfully without problems. The graphs are done with guidance from the labs that were given in class. We've implemented creating graphs as well as using machine learning to predict the outcomes. COVID-19 rates. Which is approximately 11.1 MB in size.

Table 1. *Dataset Specifications*

Data Set	Total Size (11 MB)
(COVID-19) Hospitalization Surveillance Network (COVID-NET)	11.4 KB

Table 2. *Elastic Search Specification*

Size per zone	180 GB Storage 4Ghz
Avalibility Zone	2 Zones
Total (Size x Zone)	360 GB | 8GB |UP TO 5 vCpu

Table 3. *Machine Learning Specification*

Size per zone	1GB Ram |0.5 v CPU
Avalibility Zone	1 Zones
Total (Size x Zone)	1 GB RAM| 0.5 vCPU up to 8 vCPU

Table 4. *Kibana Instances*

Size per zone	1 GB ram | Up to 8vCPU
Avalibility Zone	1 Zones
Total (Size x Zone)	1 GB ram | Up to 8vCPU

Table 5. *Integration Server Instances*

Size per zone	1GB Ram |8 v CPU
Avalibility Zone	1 Zones
Total (Size x Zone)	1 GB RAM| up to 8 vCPU

Table 6. *Enterprise Search Instances*

Size per zone	2GB Ram |8 v CPU
Avalibility Zone	1 Zones
Total (Size x Zone)	2 GB RAM| 8vCPU up to 8 vCPU

4.1. Top 5 States with Covid cases

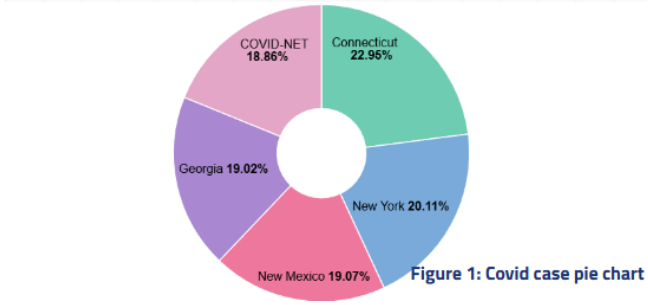

Figure 1: Covid case pie chart

Figure 1 shows the top 5 states with the most covid-19 cases reported since 2019. The graph was created by using visual tools in ElasticSearch. The COVID-NET also listed as a "state" in the database and placed in the top 5 states when we create the graph. COVID-NET means that the entry was reported through the Coronavirus Disease 2019 Hospitalization Surveillance Network (COVID-NET).

4.2. Record Count of Covid Cases By Season

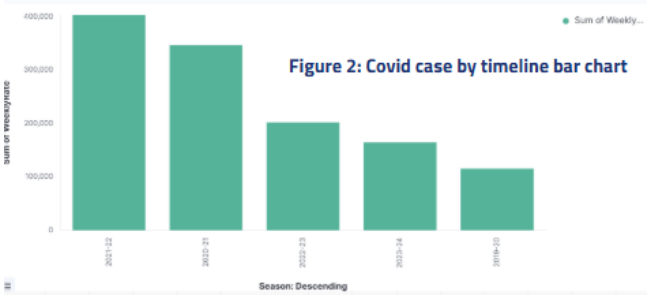

Figure 2: Covid case by timeline bar chart

Figure 2 shows the top 5 seasons that got the most Covid-19 cases report, those are all from 2020 to 2023 and all in May. Lots of people go out during the Summer time when there is no school.

4.3. Age Range

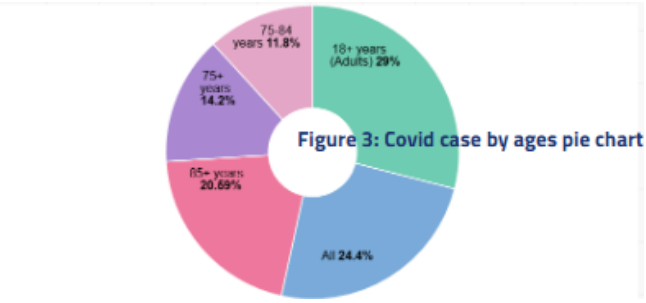

Figure 3: Covid case by ages pie chart

Figure 3 illustrates a pie chart of all the age ranges from infants all the way to adults from 18+ that was recorded that had Covid-19.

The majority of people who reported for COVID-19 were the adults group. Consider that during Summer and Winter time people would graduate or on break from school which would cause people to want to go out more and potentially get infected.

4.4. Race

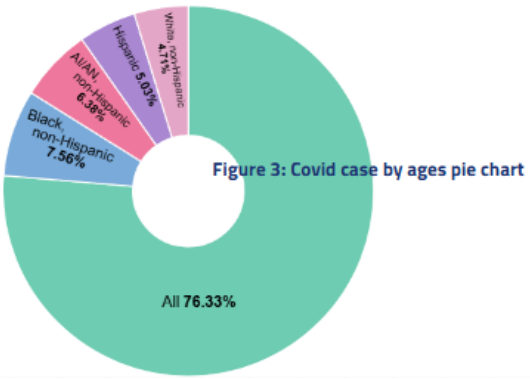

Figure 3: Covid case by ages pie chart

Figure 4 shows the races that were reported for COVID-19. As for the "all" variable it means that some people when reporting for COVID-19 put it as non-disclosure or the report itself comes from the hospital only with the state and the date without race input.

The most race beside "all" is African-American, this could be due to some factors such as high property rates, lack of

medical workers in the area, crowded living area or simply lack of access to the vaccines.

4.5. Weekly Hospitalization Count

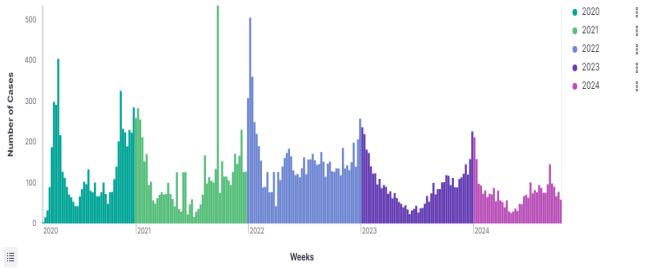

Fig. 5: Visual Chart: Weekly Hospitalizations

Figure 5 shows a vertical bar graph of the number of Covid cases on a weekly basis. This bar graph spans across the start of our data range of March 7, 2020 until October 12, 2024, all while receiving the count of hospitalizations recorded every Monday of the week.

In this bar graph, the different colors indicate the span of weeks within a year, splitting the data visually into 5 different years. Each bar represents the weeks and the number of hospitalizations in our data at a weekly rate. In ElasticSearch, the y-axis was set to aggregate the maximum number of the weekly rate. In the x-axes, the buckets were split into timestamps of *7 days* or equivalent to *1 week*. Then another split series was added, this time setting the timestamp to *per year*.

We can utilize this visualization of this bar graph to look into the trends of Covid cases. As we can see in the graphs, it would seem that the number of hospitalization cases increase during the end of the year until the start of the year. This can be hypothesized as maybe people are more prone to being infected with the disease or that people tend to gather more during the winter season, therefore spreading more infections and the number of cases leading to increase.

4.6. Weekly Prediction

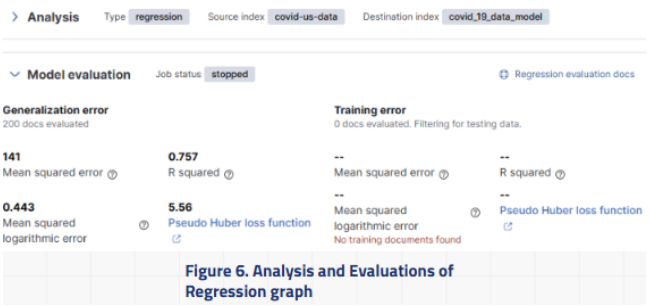

Figure 6. Analysis and Evaluations of Regression graph

Figure 6 shows R squared is 0.757. The closer to 1 R squared gets, the higher the accuracy. It is also highly unlikely that you will get an R squared of 1. If you do, you may need to rebuild and check your model for any errors.

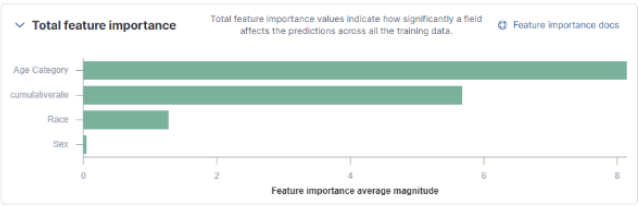

Figure 7. Weekly Prediction from Machine Learning

Figure 7 shows total feature importance. In this particular model, we see that the field, Age Category, had the highest impact on the dependent variable "Weekly Cases".

5. Conclusion

The COVID-19 pandemic, spanning from 2019 to 2024 has had a profound and multifaceted impact on the United States with varying effects across states, age groups, seasons and racial and ethnic demographics. Different states experienced surges influenced by population density, healthcare capacity, vaccination rates and public health policies. States like New York and California were early epicenters due to urban density and international travel, while others, like Florida and Texas, faced waves linked to seasonal travel and varying mitigation measures. The collective response evolved from crisis management to long-term resilience, emphasizing vaccination, antiviral treatments and public health education. While the acute phase of the pandemic has largely subsided, COVID-19 remains a significant public health concern, reshaping policies and deepening understanding of health disparities and pandemic preparedness.

References

Author links open overlay panelKannan Govindan a b, et al. "A Decision Support System for Demand Management in Healthcare Supply Chains Considering the Epidemic Outbreaks: A Case Study of Coronavirus Disease 2019 (Covid-19)." Transportation Research Part E: Logistics and Transportation Review, Pergamon, 7 May 2020, www.sciencedirect.com/science/article/pii/S1366554520306189.

"Population Covid-19 Tracking." COVID-19 Information, 29 Apr. 2023, www.cdcr.ca.gov/covid19/population-status-tracking/.

"Covid-19 Cases | WHO COVID-19 Dashboard." World Health Organization, World Health Organization, data.who.int/dashboards/covid19/cases?n=c. Accessed 30 Nov. 2024.

"CDC Covid Data Tracker." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/#datatracker-home. Accessed 30 Nov. 2024.

"Weekly Rates of Laboratory-Confirmed COVID-19 Hospitalizations from the COVID-Net Surveillance System." Centers for Disease Control and Prevention,

Centers for Disease Control and Prevention, data.cdc.gov/Public-Health-Surveillance/Weekly-Rates-of-Laboratory-Confirmed-COVID-19-Hosp/6jg4-xsqq/data_preview. Accessed 30 Nov. 2024.

LA County Daily Covid-19 Data - LA County Department of Public Health, publichealth.lacounty.gov/media/coronavirus/data/. Accessed 30 Nov. 2024.

"Coronavirus Cases:" Worldometer, www.worldometers.info/coronavirus/. Accessed 30 Nov. 2024.

Shuja, J., Alanazi, E., Alasmary, W. et al. COVID-19 open source data sets: a comprehensive survey. Appl Intell 51, 1296–1325 (2021). https://doi.org/10.1007/s10489-020-01862-6 https://rdcu.be/d0ojw

COVID-19 Incidence and Death Rates Among Unvaccinated and Fully Vaccinated Adults with and Without Booster Doses During Periods of Delta and Omicron Variant Emergence — 25 U.S. Jurisdictions, April 4–December 25, 2021. (2022, January 21). PMC Home. https://pmc.ncbi.nlm.nih.gov/articles/PMC9351531/pdf/mm7104e2.pdf

Le Vu, Stéphane, et al. "Age and Sex-Specific Risks of Myocarditis and Pericarditis Following Covid-19 Messenger RNA Vaccines." Nature News, Nature Publishing Group, 25 June 2022, www.nature.com/articles/s41467-022-31401-5.

Latoya Hill and Samantha Artiga Published: Aug 22, 2022. "Covid-19 Cases and Deaths by Race/Ethnicity: Current Data and Changes over Time." KFF, 22 Aug. 2022, www.kff.org/racial-equity-and-health-policy/issue-brief/covid-19-cases-and-deaths-by-race-ethnicity-current-data-and-changes-over-time/#:~:text=Total%20cumulative%20data%20show%20that,age%20by%20race%20and%20ethnicity.