Covid-19 Analysis of the United States

Authors: Tony Fong, Victor Verduzco, Pablo Ilabaca, Michael Medina, John Martinez
Department of Information Systems, California State University Los Angeles
CIS-4650, Introduction to Big Data
pilabac@calstatela.edu vverfuz@calstatela.edu mmedin126@calstatela.edu tfong9@calstatela.edu jmarti168@calstatela.edu

Abstract: This paper explains the method and processes we used to determine the average age range, people who had underlying conditions and how they reacted, the mortality rate in the United States from those that had contracted COVID-19 in each state, and to see if there a large difference between both genders in death rates. By looking at characteristics and demographics like a person's age, health, and sex we can understand the severity of COVID-19 throughout the United States. Our research paper focuses on how we handle this big data set using Hadoop and beeline. Along with that, we have also analyzed the data visually using the tools Excel and Power BI. From these tools we will have created map visualizations and different types of charts..

1. Introduction

This project uses Hadoop and Hive to keep and process a CDC Covid-19 dataset, this data set was collecting data in real time and also had over nineteen different elements such as age group, sex, race, ethnicity, exposure, hospitalization, death, and underlying conditions. We also used Powerbi and Microsoft Excel to create our visualization and map. As a group, we were unable to use the whole 11GB of data but were able to condense it down into 2GB to fit in the oracle server

We chose this dataset because it was the most updated dataset, and it contained a lot of elements for us to investigate. The CDC also had some of best collection methods and worked in tangent with various hospitals across the United States to ensure the data was also as accurate as possible. Using this data, we were able to cut down the data size, clean it, and then began our analysis. We focused on how many people who died had underlying conditions, the death toll of every state per capita, and also the death rates by age group

2. Related Works

2.1 Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development

In a study done by Radanliev, P et al, data mining was used to analyze scientific research records to identify relationships between certain keywords and concepts related to three specific topics: (1) Covid-19 mortality, (2)
Covid-19 immunity, (3) Covid-19vaccine development. The data mining was conducted using scientific literature form the Web of Science Core Collection.

From the analysis of this study, they identified that Chinese Universities dominated the research topics revolving around Covid-19 during the early stages of the pandemic. It was also found that there was strong collaborative research done between China and the United States for Covid-19. Conclusively, multiple relationships were identified between keywords, synonyms and concepts, related to Covid-19 mortality, immunity, and vaccine development.[1]

2.2 Understanding epidemic data and statistics: A case study of COVID-19

Another similar study using COVID-19 data was done by Hoseinpour Dehkordi, Amirhoshang, et al, to investigate and determine early warning systems developed in previous epidemic responses to contain viruses from spreading. They began by examining records of past initial outbreaks of SARS, MERS, and HCoV (SARS-CoV-2). By reviewing past epidemic records, they could see how the contingency policies of various counties affected the transfer rate of diseases in relation to COVID-19.

By statistically analyzing the data via regression analysis they looked for linear relationships between policies & behaviors with regards to Covid-19 confirmed cases, deaths, and recovered cases. From their analysis they found that lockdown policies were effective in reducing the number of confirmed cases. Additional policies such as isolation, social distancing, immediate detection of infection, and quarantine of infected individuals were also shown to effectively reduce the number of confirmed cases and case fatality. From these results they also predicted that the mass movement of people during holidays such as Lunar New Year in China would increase the spreading of disease.[2]

2.3 The characteristics of COVID-19 transmission from case to high-risk contact, statistical analysis from contact tracing data

This research study by Chayanon Phucharoen, NichapatSangkaewab, and KristinaStosica wanted to examine the transmission of COVID-19 transmission in Phuket, Thailand. What made them want to focus on this area is that infections were kept relatively low despite the other country's infection rates being higher. The model they created contained information about the patient's sex, age, how many are in the household, and the friends they met. In this model, researchers focused on how effective quarantine measures were in stopping the spread of COVID-19.

---

[1] Radanliev, P., De Roure, D., & Walton, R. (2020). Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development-In the first wave of the Covid-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1121-1132.

[2] Hoseinpour Dehkordi, A., Alizadeh, M., Derakhshan, P., Babazadeh, P., & Jahandideh, A. (2020). Understanding epidemic data and statistics: A case study of COVID-19. Journal of medical virology, 92(7), 868–882. https://doi.org/10.1002/jmv.25885

The results of this study confirmed that quarantine policies helped lower the chance of infections. Further reinforced that households with more than one person should quarantine themselves from the infected person. Their findings confirmed that sharing the same space can increase the chances of infection. The people in Phuket followed quarantine policies and were able to lower infections in the area.[3]

2.4 Difference between our work and theirs

The difference in our work is that we tried to focus on understanding how COVID-19 affected certain age groups and people with underlying health conditions. Most of the related work tried to figure out how COVID-19 policy's helped reduce infection among people, research on COVID-19's mortality and immunity rate early in the pandemic, and how effective quarantining is when it comes to helping slow down the spread of COVID-19. Our analysis of this dataset shows a stark difference because of our focus on mortality and people with diseases that put them at higher risk of dying from COVID-19.

3. Specifications

The dataset we chose contained live data from Covid-19 cases dating back to when the outbreak had first started. It comprises the clinical date of the Covid-19 patient case, the state and county they resided in, and demographics such as their age group, sex, race, and ethnicity. It also contained specific data related to the patient's contraction of Covid-19, such as their symptom status, the number of weeks they had these symptoms, if they had to be hospitalized and/or taken to the Intensive Care Unit, if they had passed away and if they had any underlying medical conditions/risk behaviors before they had contracted Covid-19. When we had chosen and downloaded this dataset, the dataset size was 12GB. However, since the dataset is currently live, it is constantly being updated by the CDC, so the dataset size could be larger in size than before.

The table below (Table 1), shows the specifications for the Oracle Big Data server cluster, as well as the specifications of Hadoop:

Table 1 H/W Specification

Hadoop Version	Hadoop 3.1.2
# of CPUs	8 CPUs
CPU Speed	1995.309 MHz
Total # of Nodes	Total Nodes: 3
Total amount of memory	Memory Size: 58GB

4. Workflow Chart Implementation

The raw dataset containing live Covid-19 patient cases was downloaded directly from the CDC website. The dataset was split to reduce its file size, and we then uploaded the reduced dataset to Google Drive. After uploading the dataset from Google Drive to the Hadoop File System, through Beeline we used HiveQL code to create the tables we needed for visual analysis. We then downloaded the data results onto our local folders, and once the file was downloaded, we used Microsoft Excel and PowerBI to create our visual graphics for analysis. Below is the workflow chart that shows the process we took for our data analysis (Figure 1).

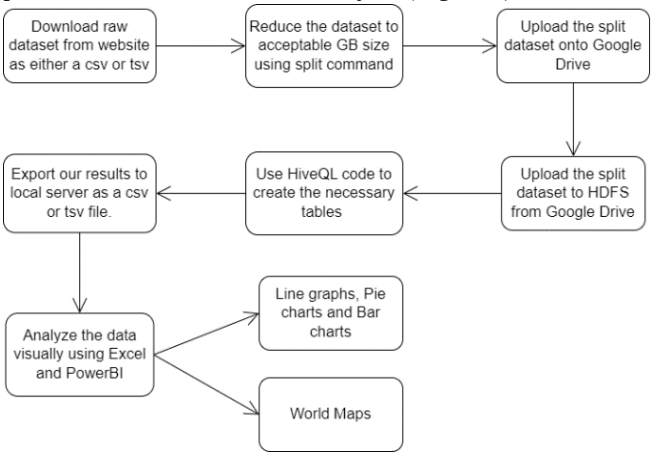

Figure 1- Workflow Chart Diagram

5. Splitting and Uploading the Dataset

Since the dataset file size was 11GB, it was way over the specific minimum limit given to us (>=2GB). Due to this, we had some complications uploading it to the Hadoop server by itself. We first tried uploading the dataset to GitHub, but GitHub has a short file size limit of 2GB and even with its highest-paid version, it still wasn't enough, only giving us a max total of 5GB. We had tried uploading to OneDrive to then upload to the Hadoop File System, but we were restricted access when trying to upload onto the file system. We did find a way to upload the entire dataset file to the Hadoop File System through Google Drive, as it more than

We managed to find a command that would enable us to split the dataset into a file that contained the number of records we wanted. The command we used to split the dataset can be found below. We split the dataset and limited the number of records to 15,000,000, which reduced the dataset file size from 12GB to 2.1GB. Below is the command we used to split the dataset:

split -l [number of records per file] [filename]
split -l 15,000,000 COVID- 19_Case_Surveillance….**csv**

6. Analysis and Visualization

After we had split the dataset, uploaded it to the Hadoop File System, and created the tables we wanted, we extracted the files and imported them into Excel and Power BI for visualization.

6.1 3D Map in Excel

In this map we visualized the death rates by age group per state. We noticed that older age groups such as 50 to 64 years

[3] Phucharoen, C., Sangkaew, N., & Stosic, K. (2020). The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. EClinicalMedicine, 27, 100543.

old and 65+ years old were the groups that had significantly higher death rates than the younger groups.

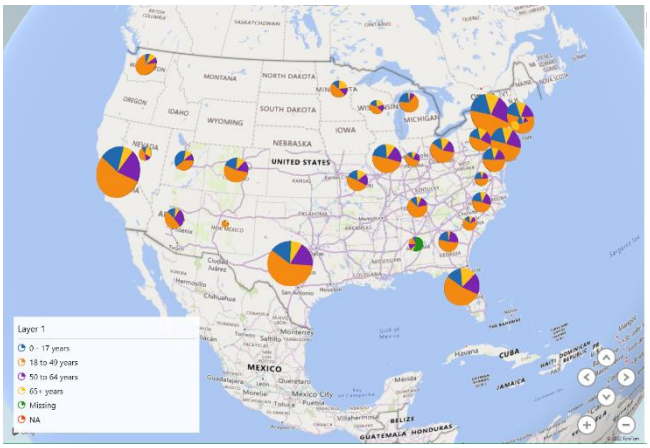

Figure 2- covid-19 cases per age group

In this map we saw people who had underlying conditions before they contracted Covid-19, we saw that it was extremely common for many people who were already sick to contract the virus in the first place

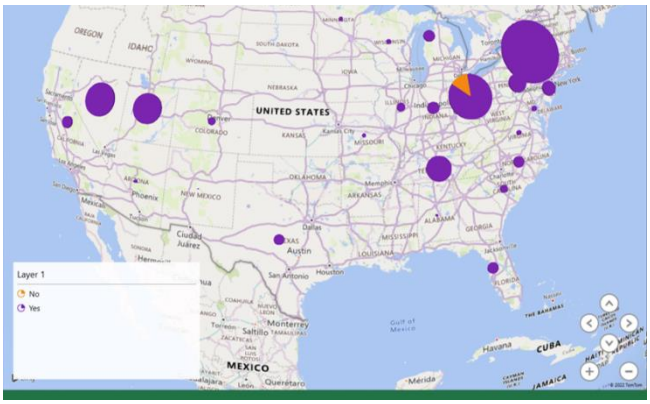

Figure 3- covid-19 cases reported on people who had underlying conditions

6.2 Power BI

In the next visual (Figure 3) we used a line chart to graph out the number of people who contracted covid over a span of 2 years. From this chart, we can see that January 2022, with 2.65 million confirmed cases, is the month with the most cases out of the 2 years. We can assume that people stopped caring around June 2021 when covid cases started to die down, resulting in a large spike in cases.

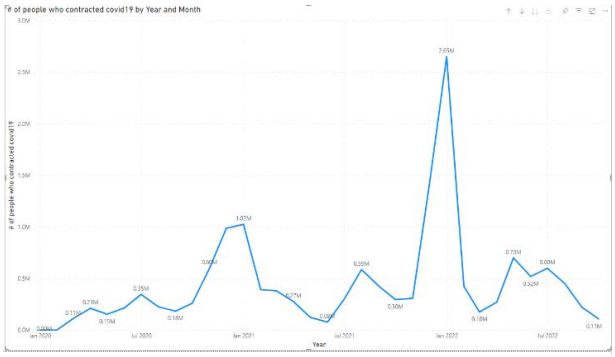

Figure 4- Number of people who contracted Covid19 by year and month

Looking at the bar chart (figure 2) we can see the amount of people who died after contracting covid in each state. The chart shows that the state with the most death is Nevada with 29.48% followed by Ohio with 24.64% and New York with 10.75%.

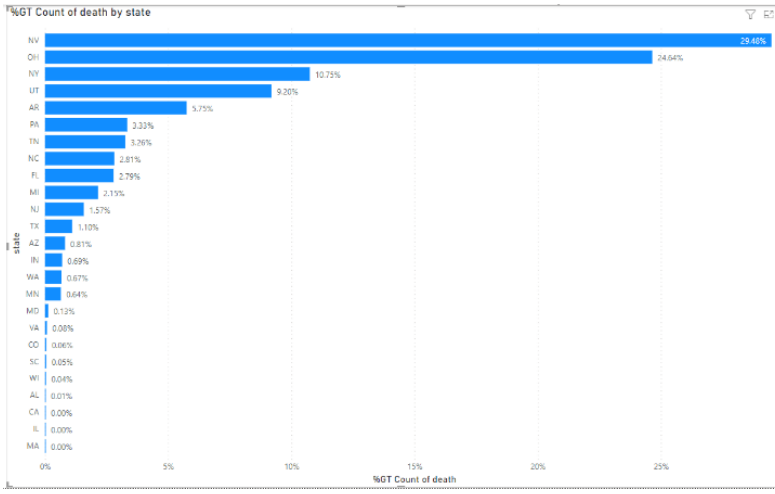

Figure 5- percent of death by state

In the pie chart (figure 3) we can see that females are slightly more prone to death after exposure to Covid-19 when compared to males. Females were reported for 54.03% of all death, while males were reported for 45.97% of all death. We can see that there isn't much of a correlation between gender and death.

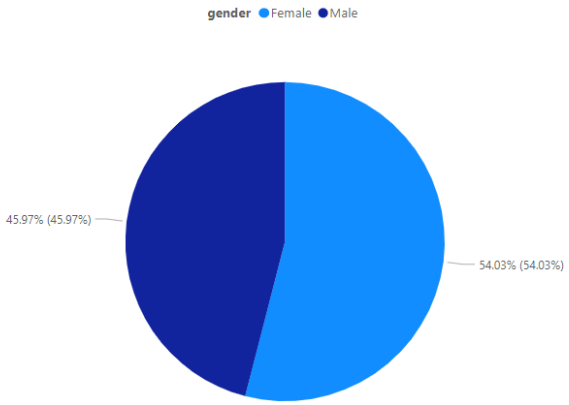

Figure 52- the percentage of death per gender

7. Conclusion

After reviewing our data we look back at the four major questions we all looked at: The death rates per state, if someone with underlying conditions would be more prone to a higher death rate, was there a difference between both genders for death rates, and also what is the most common age range that suffered the most.

While looking at the death rates per state we realized that it wasn't just the total rates that matter but more of a per capita rating seeing as states vary so much between population. Despite New York having the highest death rates it was Nevada and Ohio who had higher rates per capita than New York did. At first this did surprise us since New York has much more population density, but it was also a lot stricter on its Covid protocols than both Nevada and Ohio.[4]

When we looked at the data for death rates with underlying conditions, we noticed that nearly all people who died from covid-19 had some sort of underlying problem beforehand. This made a lot of sense to us seeing as someone who was already experiencing medical problems would also suffer greatly if they contracted a virus that affected their lungs and other organs.

When we looked at the data for genders, we did not a significant difference between for one gender having higher death rates than the other. With women at 45.87% and men at 54.03%

Finally, when we looked at the common age ranges that had the highest casualties, we saw that people that were over 50 years of age suffered the highest number of casualties. This was understandable as Covid-19 was a very aggressive virus and an older person can tend to not have as much resistance against viruses and are more prone to suffering more from all types of diseases.

For further analysis and usage of our dataset, you may visit the project's GitHub link[5]

References

[1] Radanliev, P., De Roure, D., & Walton, R. (2020). Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development-In the first wave of the Covid-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1121-1132.

[2] Hoseinpour Dehkordi, A., Alizadeh, M., Derakhshan, P., Babazadeh, P., & Jahandideh, A. (2020). Understanding epidemic data and statistics: A case study of COVID-19. Journal of medical virology, 92(7), 868–882. https://doi.org/10.1002/jmv.25885

[3] Phucharoen, C., Sangkaew, N., & Stosic, K. (2020). The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. EClinicalMedicine, 27, 100543.

[4] "Covid-19 Executive Orders." COVID-19 Executive Orders | Office of the Professions, https://www.op.nysed.gov/about/covid-19/executive-orders.

[5] https://github.com/mike0nthemic/G5_Big_Data_4560

[4] "Covid-19 Executive Orders." COVID-19 Executive Orders | Office of the Professions, https://www.op.nysed.gov/about/covid-19/executive-orders.

[5] https://github.com/mike0nthemic/G5_Big_Data_4560