Analyzing John Hopkins University COVID-19 Dataset

Using Elasticsearch Kibana and Machine Learning

Authors: William Gomez | Robin Huynh | Clever I Lemus | Luis Rodriguez | Jung Tae Cha

[rhuynh13@calstatela.edu](mailto:rhuynh13@calstatela.edu), [clemus28@calstatela.edu](mailto:clemus28@calstatela.edu), [wgomez13@calstatela.edu](mailto:wgomez13@calstatela.edu), [lrodri71@calstatela.edu](mailto:lrodri71@calstatela.edu) jcha22@calstatela.edu

Department of Information Systems | California State University, Los Angeles

CIS 3200-02 Data Processing and Analytics

**Abstract:** The following research explains and illustrates the intended structure and process used for COVID-19 cases through data manipulation and further analysis using Elasticsearch technologies and to obtain the necessary skill set of high-level data visualization analysis on our data. The mission of this research study is to provide a thorough understanding of handling and uploading data files and applying multiple data visualization processes using Elasticsearch Kibana and Git Bash. In addition, constituting eye-opening visuals such as geospatial maps, timelines, and charts based on the number of COVID-19 cases and deaths throughout the U.S counties from (2020-2022). Due to the massive size of the original data, this study designs a method and model of analysis for a smaller version of the dataset, allowing more space and flexibility for optimization.

**1. Introduction**

This project delivers the emphasis of Elasticsearch to maintain and visualize the COVID-19 dataset provided by John Hopkins University (JHU) (2020-2022). In addition, the dataset’s information about the number count of COVID-19 cases, deceased cases, locations, and more.

Throughout this research paper, we ourselves chose this dataset to utilize its trending relevance and how it reflects the ongoing challenges posed to the medical field and those affected by this disease. Since the pandemic became a typical safety concern in our daily lives, many data scientists and medical professionals grew interested in obtaining more information on COVID-19 to get the upper edge in combating against the ongoing struggles that it has brought upon us in rapid succession. Therefore, John Hopkins University, a highly reputed school for its data analysis, provided an extensive bundle of data to the open-public for the sake of expanding the need of people to work and discover many new insights about COVID-19 [1]. In addition, to join the fight against this ongoing phenomenon and to find a resolve.

2. Related Work

Our project consists of gathering U.S. Covid-19 data and learning to ingest the .csv files for Kibana for visualization. Many works or research projects regarding Covid-19 dataset have taken different approaches of gathering certain data. For example, an article has developed a different way of analyzing Covid-19 data through the social media Twitter. Lamsal conducted a study on Covid-19 related keywords [2]. This type of data is considered “socially generated data” which the article demonstrated machine learning to analyze keywords and “number of tweets” throughout each month.

Another article did a data analysis on Covid-19; however, the parameters of the study were subjected to occurrences of government policy response to the pandemic [3]. This article utilized machine-learning through the company Jataware, to collect articles based on policy implementations during covid. They used natural processing language to capture keywords of the policy event types, which are then displayed on graphical charts.

The difference between their methods of data analysis and ours is that they conducted the study based on natural language processing. The type of machine-learning for our data analysis was through regression, which deals with numerical type data. Both of our data analysis was the same regarding Covid-19 related data, however methods were executed differently based on different machine-learning methods.

**3. Specifications**

The dataset comprises of the following number of recorded COVID-19 cases and deaths in each U.S county. Since the entire JHU dataset is publicly distributed and available to the public, we were able to utilize a portion of the data provided by them through Kaggle [4]. The overall data is 309MB in size and covers two years’ worth of COVID-19 cases, and deaths in the file. Although the size not as large, we assume that similar data processes can be applied to a much more complex and bigger dataset (as large as 10GB, maybe more). Table 1 demonstrates files and size of the file from the dataset.

*Table 1 Data Specification*

|  |  |
| --- | --- |
| Data Set | Size (Total 313.45) |
| covid\_us\_county.csv | 244.89 MB |
| us\_county.csv | 356.79 kB |
| us\_county.dbf | 1.04 MB |
| us\_county.prj | 146 B |
| us\_county.shp | 67.05 MB |
| us\_county.shx | 103.14 kB |

The us\_county\_shx was not use, it is all in us\_county.csv.

The below table shows the specification for Elasticsearch Kibana we are using and Machine Learning Instance for our project.

*Table 2 H/W Specification*

|  |  |
| --- | --- |
| Elasticsearch Hot Storage | 360GB |
| Elasticsearch Memory | 8 GB |
| Master Node | 1 |
| Number of Nodes | 3 |
| Kibana Memory | 1 GB |
| Machine Learning Memory | 2 GB |
| Total Storage | 360 GB |
| Total Memory | 12 GB |

**4. Analysis and Visualization**

Chart, bar chart, histogram

Description automatically generated

**Figure 1: Top 3 Average Number of Deaths by State**

Figure 1 is a vertical bar stacked chart that shows the Top 3 states with the highest average of deaths from earliest record cases on January 21, 2020, to present time. Based off the data and period, the state with the most average deaths is Arizona with 1,865,588, followed by California with 1,624,486, and New Jersey with 1,525,764 deaths. The chart shows how both data from the states are compared in sizes and variations of each month. The horizontal axis represents the @timestamp for every 30 days. The vertical axis represents the average number of deaths. It is important to visualize the states with the most average deaths to determine the severity of Covid-19 (Figure 1).

Map

Description automatically generated

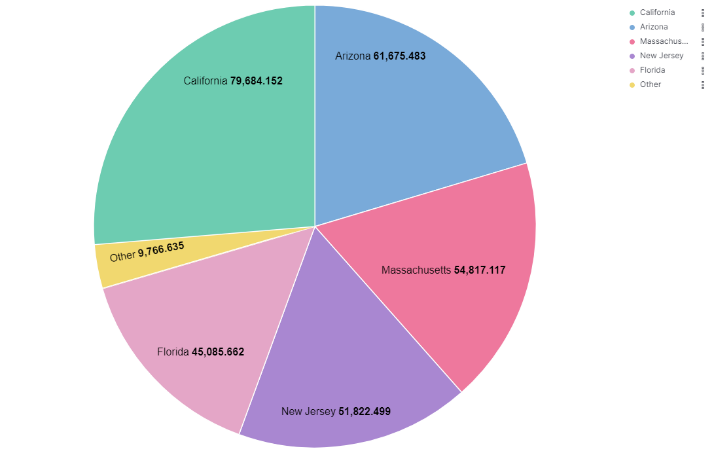
**Figure 2: Geo-Mapping of Covid-19 Deaths by County**

**Map

Description automatically generated**

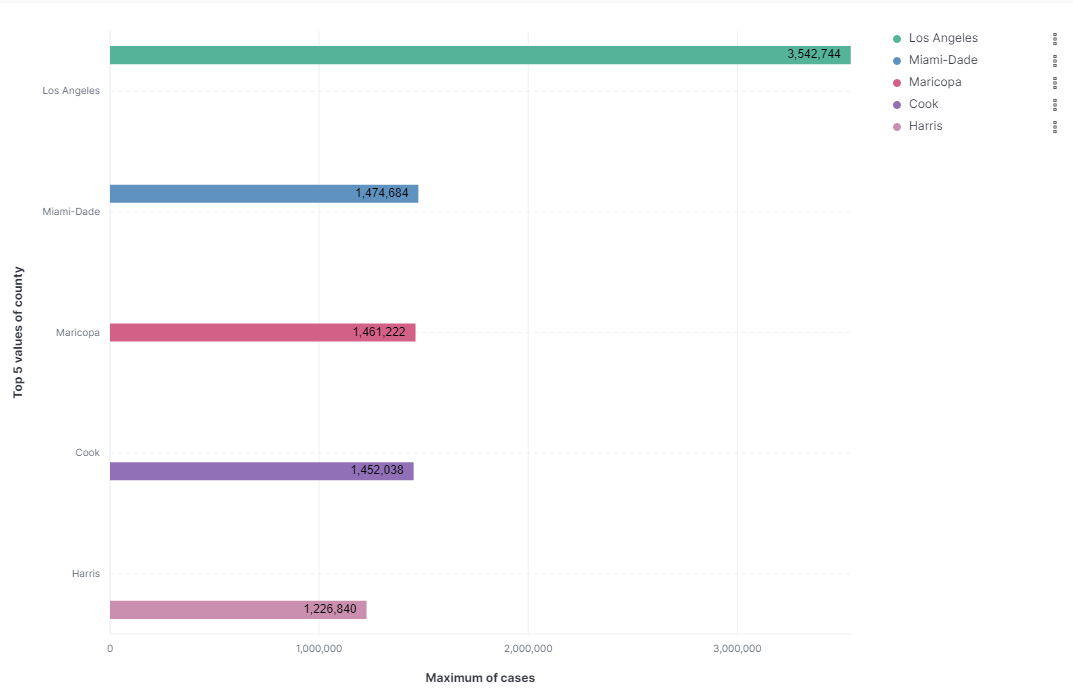
**Figure 2.1: Geo-Mapping of Covid-19 Deaths by County**

Figures 2 and 2.1 show the relevant number of Covid-19 county-level deaths. The geographical visualization shows cluster data analysis filtered through county code. As you can see the values and intensity of the color visually represents the range of cases compared to another counties. Noticeably, California shares the largest number of deaths, in addition to several other counties on the east-coast. Furthermore, all California counties combine for a total of around 45 million deaths since the pandemic initially began. (Figure 2.1).



**Figure 3: Top 5 Average Number of Cases by State**

Figure 3 shows the average number of cases each state is getting, sorted by the top five states with the most average number of cases. The pie chart represents the period from Jan. 22, 2020 – Dec. 1, 2022. From the data gathered, we can see that the state with the most average number of cases is California (79,684), followed by Arizona (61,675), Massachusetts (54,817), New Jersey (51,822), and Florida (45,085). These numbers can be seen as a good reflection of the state's overall population and how that can lead to a higher average number of cases overall. It is important to visualize the average number of cases by state to determine the states that need the most attention and assistance to combat the Covid-19 virus (Figure 3).



**Figure 4: Top 5 Maximum Number of Cases by County**

Figure 4 shows the maximum number of cases in each county, sorted by the top five counties that have the most cases. The horizontal bar chart represents the period from Jan. 22, 2020 – Dec. 1, 2022. From the data gathered, we can see that the county with the most maximum number of cases is Los Angeles (3,542,744), followed by Miami-Dade (1,474,684) and Maricopa (1,461,222). These numbers can be a good indication of which counties are highly impacted by Covid-19. The counties with the highest population are also getting the most covid cases. This bar chart will help government leaders get a better understanding of what needs to be done to stop the number of cases from increasing (Figure 4).

Graphical user interface

Description automatically generated with low confidence

**Figure 5: Forecast Average Number of Cases / County**

Graphical user interface, application

Description automatically generated

**Figure 6: Forecast Average Number of Death / County**

Figure 5 and 6 shows the forecast of the average number of cases and death in a county. The figure 5shows the prediction of the average number of cases in a county and figure 6 shows the prediction of the average number of deaths in a county.

Our latest record indicates that on November 13, 2022, the average number of cases in a county is 29,294.08. According to the forecast, by November 15, 2023, the average will increase up to 56.24% which is 45,768.134 cases. The minimum average cases per county will be 35,699.971 and the maximum will be 64,805.96.

For the average number of deaths per county, our latest record shows that on November 23, 2022, the average is 322.353 per county. The Forecast shows that, by November 15, 2023, the average will increase up to 19.18% which is 384.166. The minimum average death per county will be 297.583 and the maximum will be 470.748 (Figure 6).

Table

Description automatically generated

**Figure 7.1: Machine Learning Results**

Figure 7 shows our results from the Machine Learning Regression on Elastic. We can see that across the board, the machine from Elastic predicted a greater number of cases per county than what the probable cases are. Unfortunately, due to our data size and the length it takes to complete a machine learning task, we were unable to rerun the test to compare other possible outcomes.

Graphical user interface, application

Description automatically generated

**Figure 7.2 Machine Learning Results**

In figure 7.2, we can see our Mean Square Error and our margin of error. For some unknown reason, we got a margin of error of 1. What this would mean is that our accuracy is 100 percent, which is highly unlikely to be the case. Like mentioned before, due to the size of the data and the length of time it takes to rerun the machine learning process, we were unable to compared different results.

**5. Conclusion**

With all this said, lets sum up all the above work in the following points:

1. Increase in the number of COVID-19 total cases and deaths since 2020;
2. Los Angeles is the most impacted county with the greatest number of cases in the U.S;
3. Most number of cases in the states of California, Arizona, Massachusetts, New Jersey, and Florida.

From the available data, we discovered an interactive way for analysis and data manipulation thanks to Elasticsearch Kibana and Machine Learning. Furthermore, moving forward we hope the use of our work can be done with a larger and greater dataset to find more insights about COVID-19 and to innovate data driven conclusions on U.S county’s COVID-19 cases and deaths by using our framework.

For more information, dashboards and code visit our projects work at our GitHub link2

### References

[1] COVID-19 US County JHU Data & Demographics. (December 17, 2022) Kaggle. <https://www.kaggle.com/datasets/headsortails/covid19-us-county-jhu-data-demographics>

[2] Lamsal, R. Design and analysis of a large-scale COVID-19 tweets dataset. Appl Intell 51, 2790–2804 (2021). <https://doi.org/10.1007/s10489-020-02029-z\>

[3] Cheng, C., Barceló, J., Hartnett, A.S. et al. COVID-19 Government Response Event Dataset (CoronaNet v.1.0). Nat Hum Behav 4, 756–768 (2020). <https://doi.org/10.1038/s41562-020-0909-7>

[4] Kaggle: Your Machine Learning and Data Science Community. (n.d.). <https://www.kaggle.com/>

2 GitHub Link: <https://github.com/robee-z/cis3200/blob/master/group%202.pptx>