

CIS 3200 - COVID-19 Analysis Term Paper

Team 3 - Hector Pedro, Israel Alegria, Julio Montiel, Samuel Mendoza, Marco Rodriguez
Department of Information Systems, California State University - LA
Los Angeles, CA

Abstract: This document goes into detail on how we used Elasticsearch, Kibana, and Azure Machine Learning Studio to process, visualize, and analyze the worldwide COVID-19 case data. Due to the restrictions in Elasticsearch, we have taken a subsample of the data repository. We are able to compare multiple metrics. Such as the ratio of confirmed cases to confirmed deaths, countries with the highest deaths, cases per day, and deaths around the world.

1. Introduction

The global COVID-19 pandemic has 3,822,951 cases and 265,084 deaths worldwide.[1] In the wake of this, we have found it beneficial to analyze the case data for the COVID-19 pandemic in the hopes that we can discover any patterns that prove useful in curbing the spread of this disease.

1.1 Elasticsearch

For this project, we will be using Elasticsearch to be able to quickly and efficiently search our dataset. Elasticsearch is an open-source search engine that allows full-text search and JSON documents [2]

1.2 Kibana

Kibana is a data visualization dashboard that runs with Elasticsearch. Kibana uses data from Elasticsearch to create bar, line, and scatter plots. Kibana is also capable of placing location data on a map [3]

1.3 Azure Machine Learning Studio

The Azure ML studio is a web-based drag and drop tool that allows users to build, test, and deploy machine learning models using datasets. This will prove very useful to us as it will allow us to find any patterns in the data using R, Python, and SQL code.[4]

1.4 Johns Hopkins COVID-19 Data Repository

We are using a sample of the COVID-19 data repository from the Center for Systems Science and Engineering(CSSE) at Johns Hopkins University. The data Size of the John Hopkins repository is 133.41mb where we sample down to 11.2mb. [5]

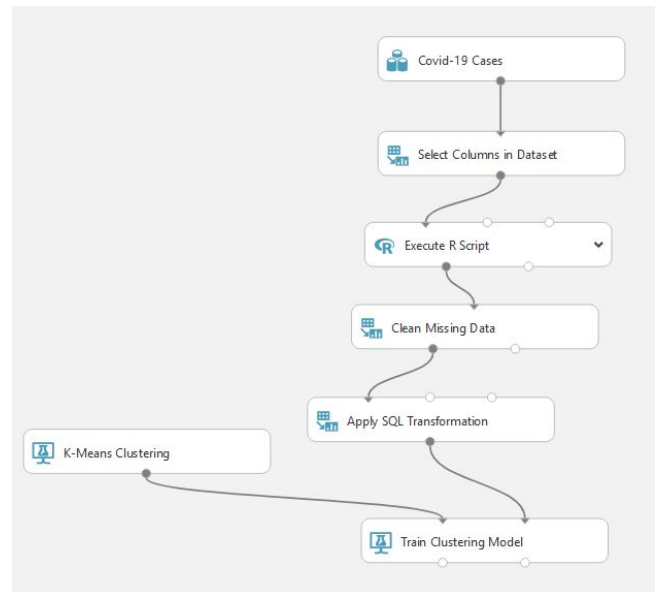


Figure 1. Azure ML studio workspace

2. Problem Definition

As the COVID-19 virus spreads and cases rise, more and more statistical data is being collected. Data is collected daily on new confirmed cases and deaths from different regions. Metrics such as Latitude/Longitude, country of the reported case/death, and difference. We want to take such data and visualize it and map it to discover any potential information such as areas of highest spread, rate of spread, case/death ratio.

3. Proposed Methodology

Our project will follow the following steps:

1. We will download and load into our Elasticsearch the datasets from Johns Hopkins CSSE
2. Import datasets into Elasticsearch
3. Create a geospatial map of death cases
4. Create a histogram of the sum of cases per day.
5. Create a pie chart of the total case percentage by country.
2. Upload dataset to Azure ML studio
3. Sanitize data by removing unnecessary columns

4. Remove rows with missing data

4. Train K-Means Clustering model to show the ratio of cases to deaths.

4. Implementation and Results

Our ELK implementation consists of 6 AWS instances located in one region and 3 availability zones. All instances are on v7.5.2 and one instance has 512MB of RAM, three instances have 1GB of RAM, and 2 have 4GB of RAM.

Our data sample will consist of two CSV files. One file regards all confirmed cases and the other all confirmed deaths



 COVID-19 Cases - COVID-19 Confirmed.csv	Add files via upload
 COVID-19 Cases - COVID-19 Deaths.csv	Add files via upload

Figure 2. CSV files as seen in our GitHub.

We will begin by importing our CSV files into elastic search. Before importing our CSV files into Elasticsearch, we must create mappings for all the fields in Kibana. We will do so by going into Kibana and Machine Learning->Data Visualizer. We will select “Import Data” and click on “Import.” We will be asked to select a name for the newly created index.

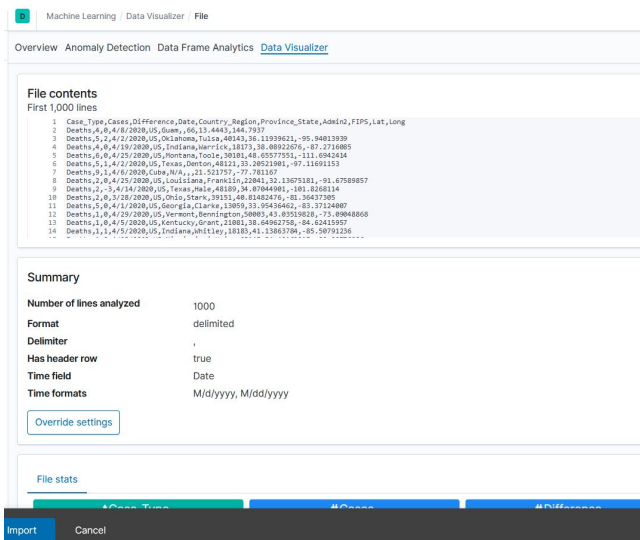


Figure 3. Data Visualizer showing data to be imported

To be able to use the data to the best of our ability we will need to create a new field in the index. We will be creating a “Coordinates” field that will comprise the latitude and longitude data of a row.



Figure 4. Creating the coordinates field

Once we have done so we click on “Import” and once the importation is complete, we will be able to select “Index Pattern Management” on the bottom of the screen. We will do so and create an index pattern based on the index we just created so we can use the data in our visualizations. Once we have done so we click on “Create index pattern” and search for the index we just created, in this case, “covid19confirmed.” After we click on next, we need to select the “Time” field as the Time filter field. Once we create the index pattern. We can use the data for our visualizations. We repeat the previous steps for the Covid-19 death data.

Uploading our dataset to Azure ML is a simpler task. In the Azure ML main page, we went to Datasets -> New and uploaded our CSV file. For Azure ML we combined both CSV files mentioned previously into one.

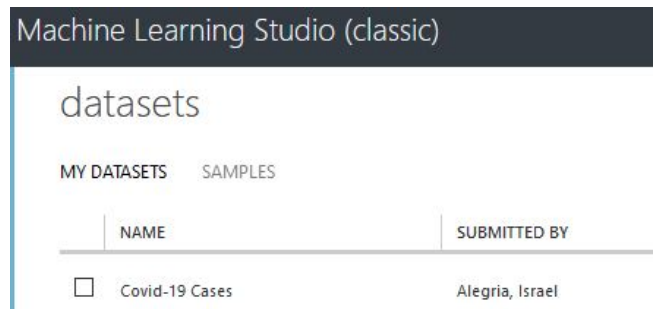


Figure 5. Partial view of the dataset as a lister in Azure ML

4.1 Mapping COVID-19 Deaths

These datasets include latitude and longitude data. By combining those two fields into the coordinates field as mentioned before, we are able to map all deaths on a map as shown below.

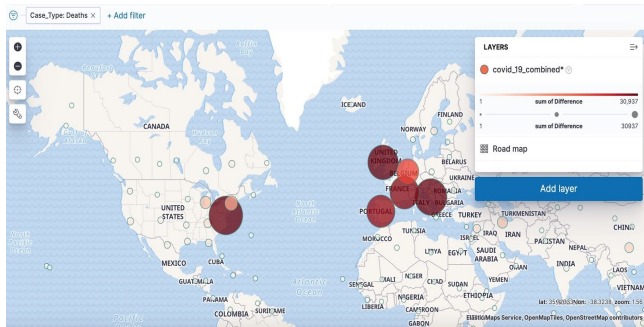


Figure 6. Mapping COVID-19 deaths.

In this case, the larger and darker the circle on the map indicates more COVID-19 related deaths in that area. As you can see, The majority of COVID-19 related deaths seem to be concentrated in Western Europe and the East Coast of the United States.

4.2 Tracking the rise of COVID-19 cases per day

We are also able to create a histogram of the rise of COVID-19 cases globally. The time field allows us to get all records and sort them by date.

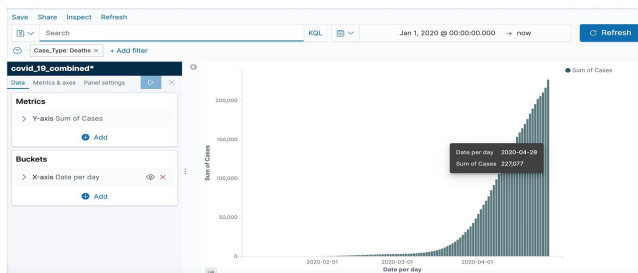


Figure 7. Histogram showing exponential rise of cases.

In this graph, the y-axis represents the number of cases and the x-axis represents time in days. As time goes forward the number of cases is increasing which poses a grave danger for the global population.

4.3 Sorting global cases by country

We are also able to sort all global cases by country to be able to see what countries are affected.

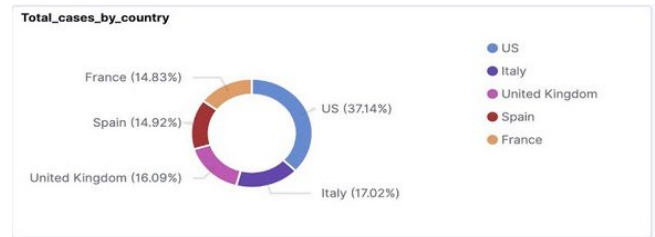


Figure 8. Cases by country

As shown above, out of our sample size, the United States has the most cases of COVID-19 at 37.14% followed by Italy at 17.02%. This can imply different things, such as the lack of preparedness on behalf of the governments of those countries.

4.4 Calculating ratio of total cases to deaths using Azure ML

We were able to graph the ratio cases to deaths by training a K-Means machine learning model on Azure ML. By uploading the dataset to Azure ML and sanitizing it using an R script to remove null values and SQL to select only the data that we want to train our model with.

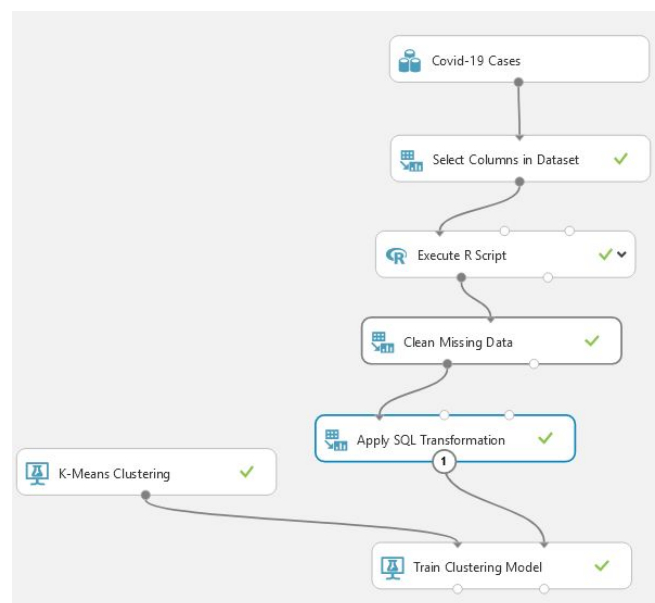


Figure 9. Azure ML experiment

We begin by loading our dataset and excluding columns that we do not need using an R script and removing rows with missing values. We also use a SQL script to select only records where the patient has deceased.

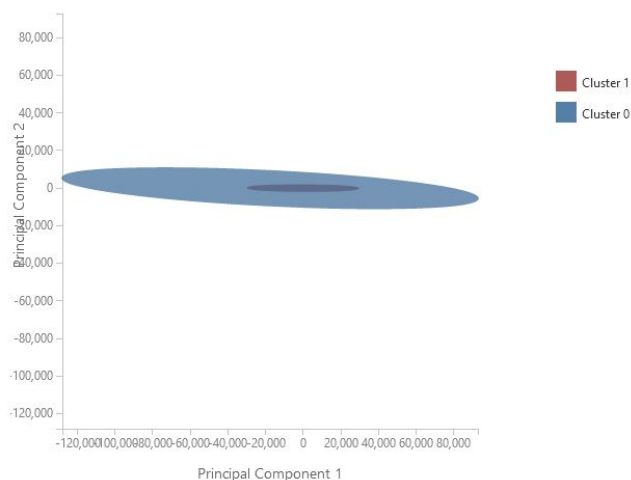


Figure 10. Results of model.

We trained this model with 205,156 entries. We can see that the blue ellipse represents the total number of cases and the smaller red ellipse represents deaths. As of the time of this publication the death rate for our subsample is 21.6%.

5. Conclusions

We have been able to determine the areas of higher infection through visualization and have been able to determine the death rate of these areas using machine learning. We believe that data analysis/visualization and machine learning will help us to predict the direction that this pandemic will take and better prepare us to address this pandemic.

References

- [1] Coronavirus Update (Live): 3,822,951 Cases and 265,084 Deaths from COVID-19 Virus Pandemic - Worldometer
<https://www.worldometers.info/coronavirus/>
- [2] Elasticsearch Retrieved from
<https://github.com/elastic/elasticsearch>
- [3] Five Reasons to Upgrade to Kibana 4 Retrieved from
<https://www.theserverside.com/discussions/thread/80828.html>
- [4] What is Machine Learning Studio (classic)? Retrieved from
<https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio>
- [5] Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE Retrieved from
<https://github.com/CSSEGISandData/COVID-19>