

**BIA-678 A: Big Data Technologies**

**Project Report on COVID-19 Analysis**

**Table of Contents**

* Introduction
* Data Insights
* Exploratory Data Analysis
* Data Visualization
* Model
* Measurements of Performance
* Conclusion & Future Recommendations
* Appendix

# **Introduction:**

Since the first case was reported on December 31st, 2019 in Wuhan, China, the novel coronavirus disease (COVID-19) swept the world into a global pandemic. Virtually every country, organization, and person was affected by this virus in some capacity. As of the end of November 2020, there have been 62.1 million cases, with 39.8 million recovered and 1.45 million deaths, marking that global mortality rate at 2.3%.

In the US alone, there have been over 328,000 cases since January. In response, the White House, private universities, and a coalition of leading research groups have created many datasets that are free to use for the public. Our objective is to conduct an exploratory data analysis and cluster analysis on which areas were most affected by COVID-19.

Some key questions we want to know is how the virus spread over time, what demographics were most affected, and what geographical clusters exist, if any.Some technologies that will help with ourmodel are PySpark, AWS EMR, Microsoft Excel, Tableau, and AWS S3. These findings will be significant to figure out how the virus spread and how to prevent it from spreading to more vulnerable communities, as well as which areas may need priority access to the vaccine. In this paper, we will discuss the dataset, the big data visualization techniques, and the models used to answer our key issues.

Ultimately, the global pandemic is one of the most important issues affecting the world today, and we hope to shed some light on the state of the virus.

# **Data Insights:**

The dataset we used was a collection of datasets compiled by Johns Hopkins University. The data sources include renown research establishments including the WHO, ECDC, CDC, BNO News, and so on. There are 8 files and 1608 columns, and the largest dataset, which provides geographical and daily information on COVID-19 cases, has over 156,000 rows.

The dataset spans from January 22nd, 2020, to November 16th, 2020, providing almost 11 months of daily data of confirmed, recovered, and death COVID-19 cases from over 100 countries. A sample of the dataset can be found in the Appendix under Sample Dataset section.

This dataset also includes time-series information of confirmed, death, and recovered on the international and US level, as well as line lists that include characteristics of each case such as age, gender, location, exposure date, and symptoms.

The time-series data pivot each date as the column and the country on the international data and states on the US level as the row. Some of the variables that we used for our exploratory data analysis were Last Updated, Total Confirmed, Total Recovered, and Total Deaths.

# **Exploratory Data Analysis:**

To begin our analysis, we started with the daily data to see the timeline of COVID-19. The tables can be seen in the appendix as Table A and Table B. We saw an exponential increase of the number of confirmed cases across the world. At the end of January, there were 10,151 confirmed increases, but by November 16th, there were 54,370,186 cases, with the biggest increase happening between September and October. We saw that in January there were about the same deaths as recoveries, as doctors scrambled to assess all the symptoms and how to manage them.

However, as doctors had more time and researchers found treatments and medications that mitigate symptoms, we saw recoveries followed a very similar exponential curve to the confirmed cases, while total deaths are increasing much more slowly in comparison. This confirms that many COVID-19 cases do end up with recovery.

We also looked into the mortality rate, total deaths over total confirmed, to see how that has changed over time. The lowest mortality rate was in January at 2%, and this could be due to the dataset starting on January 22nd, and doctors were unclear if patients were dying from COVID-19 or another illness. The highest mortality rate was at 7% in April, which was when COVID-19 cases started to exponentially increase all across the world.

Now, mortality rate is at 2.4% and is slowly decreasing as doctors are able to successfully treat patients, and with a new vaccine on the horizon, hopefully it will decrease even further.

# **Data Visualization:**

For big data visualization, we have created several Tableau tables to visually capture the spread of coronavirus from January to November, separated by quarters. We used the COVID-19 daily international file as the input. The 150,000+ rows csv file is overwhelming to assess on a spreadsheet, but on a geographical map, it is easier to analyze the highly impacted areas. The tables can be found in the Appendix section. Coronavirus cases insights in the entire world:

1. *For the first quarter of 2020:*

We observe that in Quarter 1 of 2020, China comprised the highest number of COVID-19 cases i.e., about 22% of the total cases across the world were from China. Also, Italy did not stand far behind with 13.71% cases of the world being from Italy. Italy being a relatively small country in Europe had a huge number of per capita cases as opposed to China, which is approximately 30 times the size of Italy!

1. *For the second quarter of 2020:*

We observed that China no longer has a maximum number of COVID cases. Instead, the United States now has the highest share in terms of new COVID cases i.e., 27.63% of the cases worldwide to be precise. Southern America and in particular Brazil were also impacted by the virus during the second quarter having about 10% of total cases worldwide.

1. *For the third quarter of 2020:*

The United States again led in terms of total COVID cases, but the total percentage of cases worldwide decreased to 23.08% from 27.63%. Also, India had a boom in the number of cases in Quarter 3 and comprised 15.19% of the world’s share. Brazil was not far behind with 14.78% of the cases. We also observe that China and Italy which were major constituents in Quarter 1 have done fairly well in recovering from COVID-19.

1. *For the fourth quarter of 2020:*

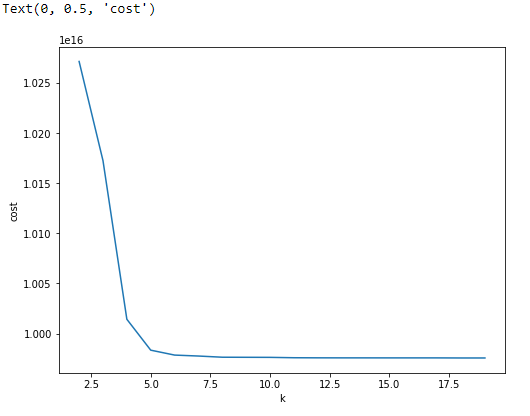
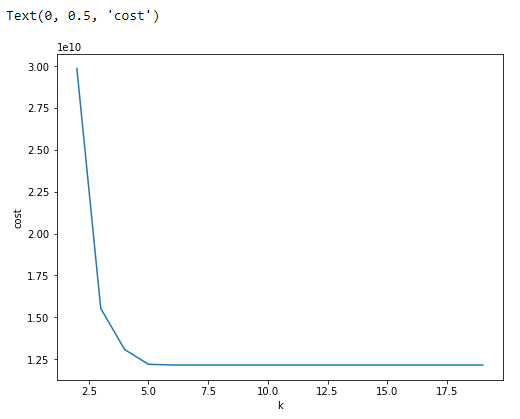
We observe a decrease in percentage share of COVID cases for U.S. and Brazil but the cases are still rising and yet to reach their peak in India with India comprising approximately 17% cases alone worldwide.

# **Model: “K-Means Clustering Algorithm”**

The data taken into consideration for building the model here is the time series data of both the confirmed cases in the world and the United States [Refer appendix for the snapshot of the data]. As mentioned above the data ranges from Jan, 22 to November 16, 2020. Here the individual time columns of the dataset are our input features based on which the clusters would be computed.

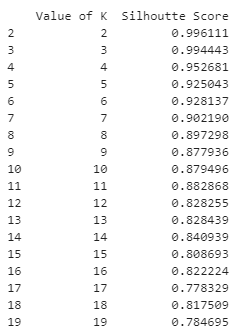
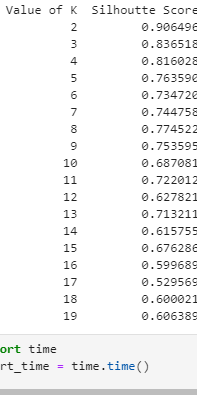
To evaluate the optimal number of clusters for our data we will be using two types of metrics. The silhouette scores and the elbow method. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. It is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The higher the score the better. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k.

The following are the elbow graphs obtained for the World Dataset and the US dataset:

** **

***World Dataset USA Dataset***

The following are the Silhouette Scores obtained for the World Dataset and the US dataset:

***World Dataset USA Dataset***

From the *elbow graph* of the world dataset we can see that there is an inflection point or bend at around k=4. Though the silhouette score for k = 2 and 3 is highest, it is not feasible to have a small number of clusters for more than 200 countries and therefore in our analysis we have chosen k as 4 for the world dataset. The same observations can be made for the USA dataset, k =3 and 4 are close contenders but based on the elbow graph we choose our k as 4 again here.

# **Measurements of Performance:**

* **Based on time and parallel computation**

The first measurement of performance we tested was time performance (Based on simple execution of k-mean based on the chosen k values). The first time we ran the model was on a Docker container in a Jupyter notebook run in PySpark on a local Windows computer. The model ran with no errors and in decent time, with the whole code taking a few minutes and running the model took 33 seconds.

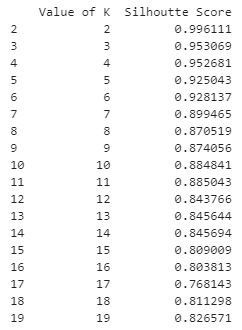
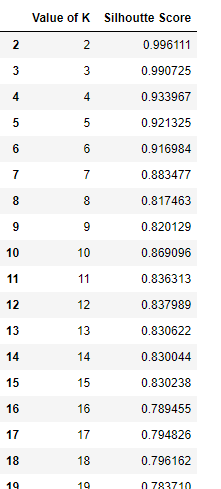
Then, we switched to the AWS ec2 service, opting for a Linux virtual machine with a higher memory capacity. We saw a substantial reduction in the time, with the code running the same data, but twice as fast. The model ran in 18.2 seconds. This showed that using a machine with higher memory, bandwidth, and CPU core improved the time performance of the model and had an overall increase in the quality of analysis.

Next we checked the performance of Local Machine based Jupyter Notebook with AWS EMR running on clusters size of 3 (1 master and two nodes). This comparison is based on the computation of the Silhouette scores for different values of k (2,20) on the two datasets [Refer appendix [5]]. We particularly picked out this block of code for comparison to show the substantial difference in the speed based on a single machine and parallel machines. The following comparisons were noted:

|  |  |  |
| --- | --- | --- |
| ***Time taken for computing Silhouette scores ( k in range (2,20)*** | ***Local Machine***  ***(sec)*** | ***AWS EMR (cluster of 3)***  ***(sec)*** |
| World Dataset | 723.32 | 55.96 |
| USA Dataset | 702.48 | 27 |

* **Based on quality of analysis**

Our major comparison metric over here was the value of the Silhouette scores. This score compares how well the elements of clusters are classified and we compare the performance for this on a local single machine with the scores given by AWS EMR notebook. Here is the comparison below for the World Dataset.

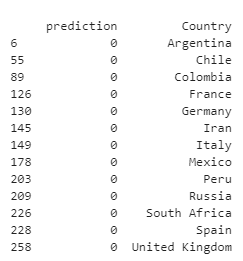
 

**AWS EMR Local Machine**

As seen above, the AWS EMR makes or forms clusters in a better way compared to when the code is run on a single machine. For k = 4 on which our cluster is based there is a difference of 2 percent in the silhouette score. Similar results are observed for the US dataset where the parallel computation of AWS EMR outperforms single machines.

**Model Clustering Results:**

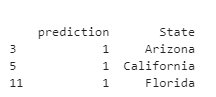
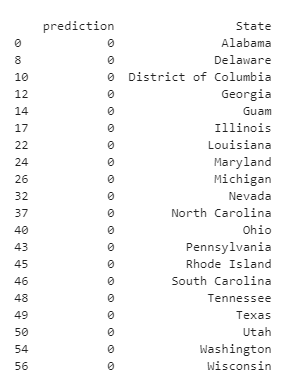
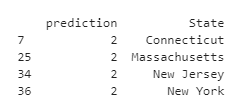
1. *World DataSet:*

  ****

\*The rest of the countries in the dataset are in the last cluster.

1. *USA DataSet:*

K =4, the states not seen below are in the last cluster.

# **Conclusion & Future Recommendations:**

Ultimately, we have found several key observations in our COVID-19 analysis. Starting with exploratory data analysis, we found the timeline of the COVID-19 spread. This showed that although the mortality rate has significantly decreased since April, the confirmed cases rate is still going up, therefore increasing the transmission rate. This shows that the pandemic is still far from over, and precautions need to be taken to avoid overwhelming the healthcare systems across the world. Next, we completed a data visualization analysis using Tableau. We used data to find the geographical spread of the virus since January. We found that in each quarter the severity of each country differed, from China and Italy having the worst outbreak in the 1st quarter, to the United States currently experiencing the worst. These graphs made it much less complex to analyze the big dataset, rather than sifting through the locations column on an Excel document. This map showed which countries were successful in combating the virus, such as China, and which are still struggling today, such as the US.

Next, we built a clustering model for both the World and the US states. Based on the clusters formed for the World dataset, we identified the observations mentioned above. The USA being the country with highest cases is associated with a single cluster. The next affected countries Brazil and India are associated with another cluster. Countries which are still hit by the virus but are recovering are in another cluster (European Countries mainly have formed one basket, indicating the geographical based cases clustering). For the USA dataset, we again see a similar pattern. The tristate area has been clustered together, again indicating a geographical significance. The states with the highest cases California and Florida belong to another cluster.

A similar clustering model has been developed by us where we cluster based on the provinces in the states. That part of the model can be seen in our submitted code. Over here the major part this model would play is in understanding the risk factors based on provinces or counties. This would help the states in planning out the rolling of the forthcoming vaccines based on the rates of the cases. This is one of the use cases of our project.

As seen above, the model had one cluster which contained a huge number of states and countries. One of the future enhancements of this model would be to form the clusters of this particular large cluster. Such a one step below iteration of this model would help us in identifying perfect clusters.

Finally, there are also some other future enhancements for this analysis. This dataset is updated almost everyday as new numbers come from the WHO and CDC, so it would be interesting to keep tweaking the analysis and model as well as adding more analysis as the United States contemplates how to navigate the pandemic. Primarily, more pandemic-related shutdowns likely lie ahead in different parts of the world. We hope that people review their actions and decisions during this first shutdown that they will not only focus on successes but also identify and learn from their deficiencies. In our quest to make additions to our project of covid analysis, there are several recommendations.

First, we can add on dynamic data stories with more automated and consumerized experiences in our analysis. Here, we can replace visual, point-and-click authoring and exploration performed on Tableau. This leads to eliminating time spent by users on predefined dashboards. This shift to in-context data stories will show/stream the most relevant insights to users. These dynamic insights will make use of technologies like augmented analytics, NLP, streaming anomaly detection and collaboration.

Next, as we know, it is critical in combing through thousands of research papers, news sources, social media posts and clinical trials data. These sources are very helpful for medical and public health experts to make predictions on disease spread, capacity-plan, find new treatments and identify vulnerable populations. So, in our analysis, we can add graph analytics to assist us in identifying, predicting and planning for upcoming covid disaster in the future. Being a data analyst, we can use a data variable for a range of different structured and unstructured content such as text analytics, video analytics, audio analytics, etc. We tend to make use of this variable and contribute in solving challenges in pandemic.

Third, in terms of large scale, Graph analytics comes into light. It is a set of analytic techniques that allows for the exploration of relationships between entities of interest such as organizations, people and transactions. This can help us find unknown relationships in data and review which is not easily analyzed with traditional analytics.

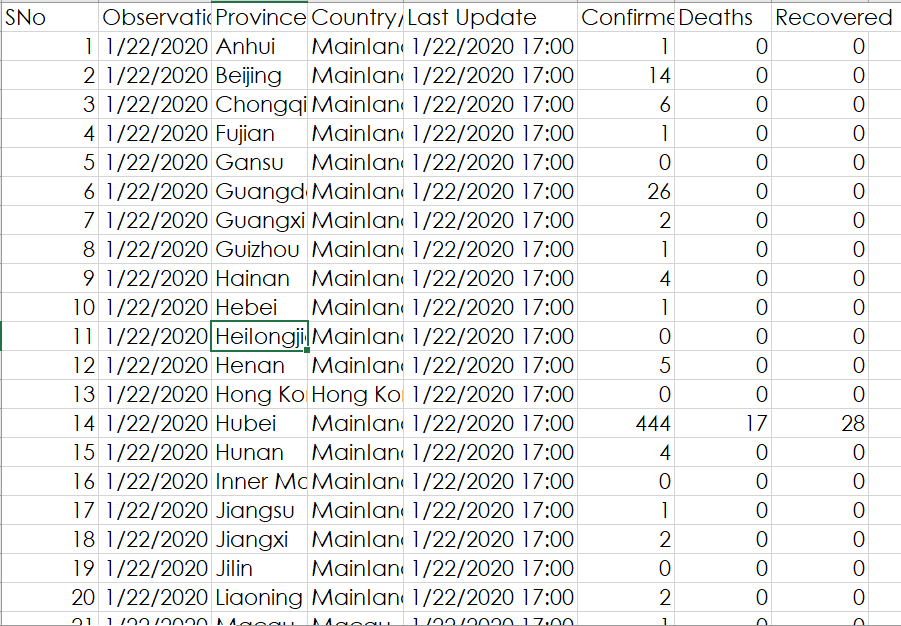
When combined with ML algorithms, graph analytics technologies can be used to comb through thousands of data sources and documents. This helps us discover new possible treatments or factors that contribute to more negative outcomes for some infected people. We can evaluate opportunities to incorporate graph analytics into analytics portfolios and applications to uncover hidden patterns and relationships. In addition, consider investigating how graph algorithms and technologies can improve our AI and ML initiatives.

Finally, in our COVID 19 analysis, we can introduce augmented data management.

This makes use of ML and AI techniques to optimize and improve operations, converts metadata from being used in auditing, lineage and reporting to powering dynamic systems, examines large samples of operational data, including actual queries, performance data and schemas. Using the existing usage and workload data, an augmented engine which tunes operations and optimizes configuration, security and performance can be added to our project. This will help us enable active metadata to simplify and consolidate architectures, and increase automation in redundant data management tasks.

# **Appendix:**

*Sample Dataset:*



*Exploratory Data Analysis:*

Table A:

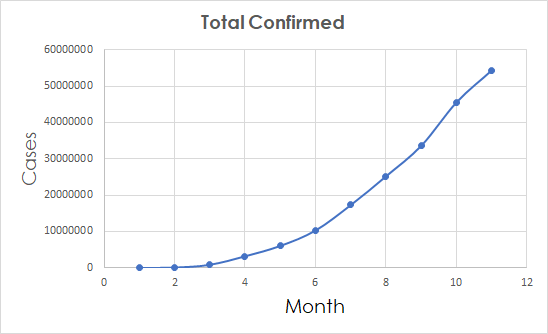
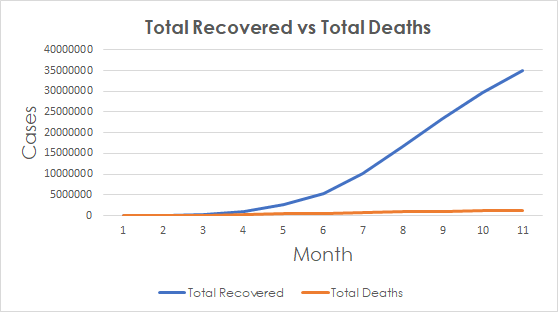
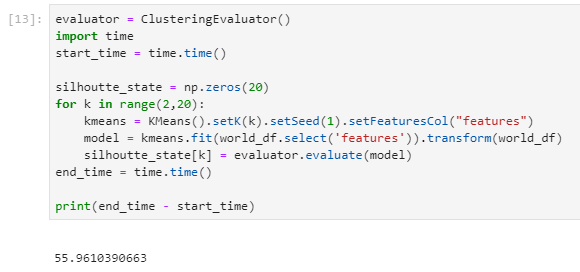


Table B:

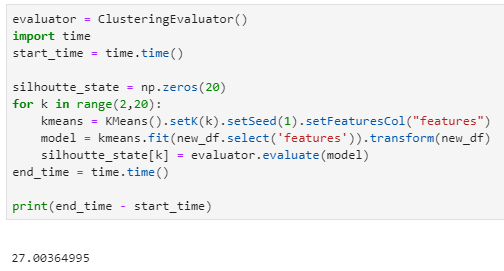


***Silhouette Score Computation:* AWS EMR, cluster size = 3**

***World Dataset****:*

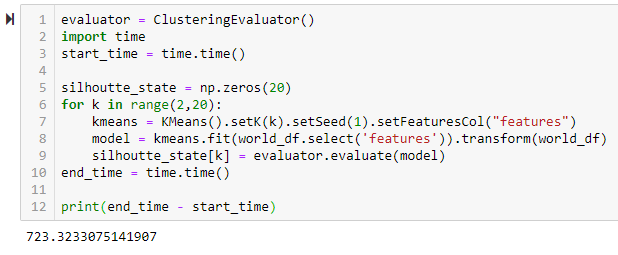
****

***USA Dataset:***

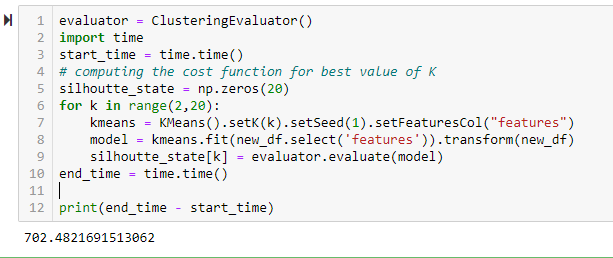
****

***Local Machine (Jupyter NoteBook)***

***World Dataset:***

****

***USA Dataset:***

****

**Sources:**

1. *COVID-19 global numbers:*

<https://tinyurl.com/y6lcjnng>

https://tinyurl.com/y6lcjnng

1. *Johns Hopkins Dashboard:*

https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

1. *Dataset:*

<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>

*4. Project Code:*

https://drive.google.com/drive/folders/18ok4OWUCtvq6q68BX9Sw9a7GOyhrNTeA?usp=sharing