**Covid-19 Pandemic Case Statistics and Analysis**

A Project Report   
Presented to    
DATA-228   
Fall, 2021

**By**

**Deep Arvind Bambharoliya**

**Ha Phan**

**Shrey Bishnoi**

**Vaibhav Yalla**

**13-12-2021**

**Acknowledgment**

We would like to express our deep gratitude to our supervisor Professor Andrew Bond, San Jose State University, San Jose, California, for all the knowledge he has been giving us and the opportunity to work on this project. It enables us to apply what we learn to solve a practical problem and gain more experience working in a team.

We are very thankful to Pravallika Vallapuri TA, Data-228 Big data technologies and applications, San Jose State University, San Jose, California, for her encouragement and support to complete this project.

We also acknowledge all the authors mentioned in the reference for their inspiration to work on this problem.

**Contents**

[**1** **Abstract:** 3](#_Toc90322389)

[**2** **Introduction:** 4](#_Toc90322390)

[**2.1** **Project goals and objectives** 4](#_Toc90322391)

[**2.2** **Problem and motivation** 4](#_Toc90322392)

[**2.3** **Project application and impact** 5](#_Toc90322393)

[**2.4** **Project results and deliverables** 5](#_Toc90322394)

[**3** **Literature Survey:** 6](#_Toc90322395)

[**4** **Data Description:** 6](#_Toc90322396)

[**4.1** **Data Source:** 6](#_Toc90322397)

[**4.2** **Data Modeling:** 7](#_Toc90322398)

[**5** **Design and Implementation:** 8](#_Toc90322399)

[**5.1** **Overview:** 8](#_Toc90322400)

[**5.2** **S3 Buckets:** 9](#_Toc90322401)

[**5.3** **AWS Glue Crawler:** 9](#_Toc90322402)

[**5.4** **AWS Glue Job 1:** 10](#_Toc90322403)

[**5.5** **AWS Glue Job 2:** 11](#_Toc90322404)

[**6** **Analytical queries:** 12](#_Toc90322405)

[**6.1** **Redshift:** 12](#_Toc90322406)

[**6.2** **Redshift queries:** 12](#_Toc90322407)

[**7** **Visualization:** 15](#_Toc90322408)

[**8** **Github:** 20](#_Toc90322409)

[**9** **Conclusion:** 20](#_Toc90322410)

[**10** **References:** 21](#_Toc90322411)

# **Abstract:**

Since the first Covid-19 case was reported in November 2019, the world has been experiencing this deadly disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that attacks the human respiratory system. This deadly disease, Covid-19 (Coronavirus), had killed 4.5M people (about twice the population of New Mexico), and there are still 219M cases present across the world that are infected. Collecting and examining a large amount of Covid-19 data is challenging, and processing and visualizing it statistically needs more processing power. We believe that through an open and collaborative effort that combines data, science, and technology, we can inspire insights and foster breakthroughs necessary to limit COVID-19 cases by finding the trends. In this project, we will use datasets like “COVID-19 Data Lake”, “COVID-19 Foot Traffic Data”, and Kaggle datasets which are used to collate daily COVID-19 cases, deaths, and various other trends. Some supplement datasets are also added to improve the accuracy of our projections. Using this tremendous amount of data and AWS Services, we analyze how this global pandemic has affected human lives worldwide and analyze the vaccination policies of different nations to examine the best feasible way to help mitigate fatality rates. It also discusses how some policies (social distancing, face mask coverings, stay-at-home restrictions) prevent Covid-19 cases. This project will help understand and capture the trends to make people assess the situation quantitatively and take necessary preventive actions. This project can be used by researchers, biomedical engineering, government, and NGOs to make people aware and provide the right support (vaccine, hygiene protocols) and common people to understand the data in a simpler format.

# **Introduction:**

## **Project goals and objectives**

The project goal is to analyze the huge trending covid-19 dataset. Our objective is to understand the patterns of the covid cases, and how other attributes like vaccinations, lockdowns, and handwashing help reduce the number of deaths and cases. Using tools, we have created a pipeline that will help clean the data and analyze it.

## **Problem and motivation**

Covid-19 pandemic had shaken the entire world. Many lost their loved ones, their jobs, career, businesses, and many more. This had an unbelievably harmful impact on our society. There were more than 270 million people who were affected by covid and more than 5.7 million who died due to covid. There were many vaccines supplied, new policies were added like lockdown, handwashing, and social distancing just to reduce the risk of getting affected by Covid-19.

The problem is if the above policies and vaccines are even helping reduce the total number of covid cases in the entire world. To understand the pattern, we have analyzed the huge data from the year 2020 February till December 2021. With this data, we can find out some patterns which might help reduce covid.

This information should be present to the entire world, for that we can use some visualization tools which will help users easily understand complex data. So, making this information public will help citizens, researchers, Doctors, students, and Government be cautious about their health and the future.

## **Project application and impact**

This project analysis can help people understand the rise and fall of covid cases, deaths, vaccines supplies, fully vaccinated people, and many more all over the help. Users all over the world can see the analysis and be cautious about the situation. This will not only help common people, but also the government, doctors, and researchers to advance their techniques

## **Project results and deliverables**

In our project, we applied knowledge about big data analytic tools on AWS to process and store a huge amount of Covid data worldwide.

We perform analytical queries on AWS and visualization processes on Tableau on to give a whole picture about covid19 in the world in terms of the total cases, deaths, hospitalization country by country, continent by continent, daily, monthly, the numbers of people vaccinated and gain insights how policies and regulations have a significant impact on slowing the spread of this virus.

# **Literature Survey:**

Many scientific articles have been conducted based on statistical data on the number of people infected with Covid in different countries to help scientists and epidemiologists build predictive models to help scientists capture the nature of the virus the direction and extent of infection in the community. This helps the government and legislators to make appropriate regulations to reduce the possibility of disease spread in each region as well as to support the right resources. Tosi D (2020) analyzed data collected globally to build models which can predict the waves of infection and when the pandemic is under control in Italia. They also emphasized the effectiveness of big data and data analytics in collecting and analyzing data so that scientists can build models that predict the development of Covid 19. Alsunaidi (2021) presented the importance of big data analytics in processing and studying the huge volume of covid 19 so that scientists can gain insights that contribute to defeating this pandemic. Farhadi (2021) concluded that accurate data sets will be effective tools for scientists and governments to evaluate the effectiveness of methods and policies to reduce the spread of this disease.

It has been two years since the first case was confirmed, the world is still learning and discovering the possibilities of evolutionary of the disease caused by Covid 19. Tools and applications for data collection, storage, and analysis always make a significant contribution in the battle to control the pandemic.

# **Data Description:**

## **Data Source:**

The data is collected and updated by Amazon Open Repository. The team has been collecting data from many sources such as John Hopkins University, European Centre for Disease Prevention and Control (ECDC), and some other countries like the United Kingdom, United States, Canada. The parameters in the data include the number of confirmed cases, deaths, hospitalizations, and testing, as well as other related information such as the degree of policies are imposed on different regions, countries, GDP, or population which are helpful for the analysis process.

## **Data Modeling:**

Data modeling was a challenging part for us. As we took data from multiple sources, we had to design a model as per our requirements. Our main dataset contained 67 columns and around 200,000 records. So, from this collective dataset, we had to partition it into smaller related entities.

Diagram

Description automatically generated

Figure 1. Data Model

We partitioned the dataset into 11 different tables and created primary key foreign keys to connect them. We created tables like Vaccinations, which contained all the information about the vaccination status for all the countries. For the table confirmed cases we put all the columns containing data about the cases in all the countries. Similarly, we created a total of 11 tables which were connected by using foreign keys

# **Design and Implementation:**

## **Overview:**

We have performed ETL operations using AWS. The services include S3 bucket, Glue, Athena, and Redshift. There was also a third-party integration, which is Tableau for visualization.

We have started importing Covid-19 data from the AWS open-data registry to the S3 bucket. S3 buckets are good for storing data, as they are scalable with superior performance and security. After importing, this data is sent to the Data Catalog using AWS crawlers. Crawlers are used to crawl the data and store the metadata in the database as a tabular form in Data Catalog.

Diagram

Description automatically generated

Figure 2. Architecture

We have created a custom Glue code to fill the NULL values to zero and another job that will help to divide the main table into smaller dimensions. Each dimension has its unique attributes. Once it is divided, we can check the data using SQL queries in ATHENA. All the files are sent to Redshift for analytical queries.

Integration with redshift was done for visualizing the data. Visualization of the data will help in understanding this complex data more easily.

## **S3 Buckets:**

We need to store this data in simple storage in AWS, and one of the best options we have is AWS S3. The data we have contains covid-data from all countries all over the world. To access this data in AWS, we first need to store this data in the S3 bucket. We have created two buckets, one for the input and another for the Gl7ue job output files.

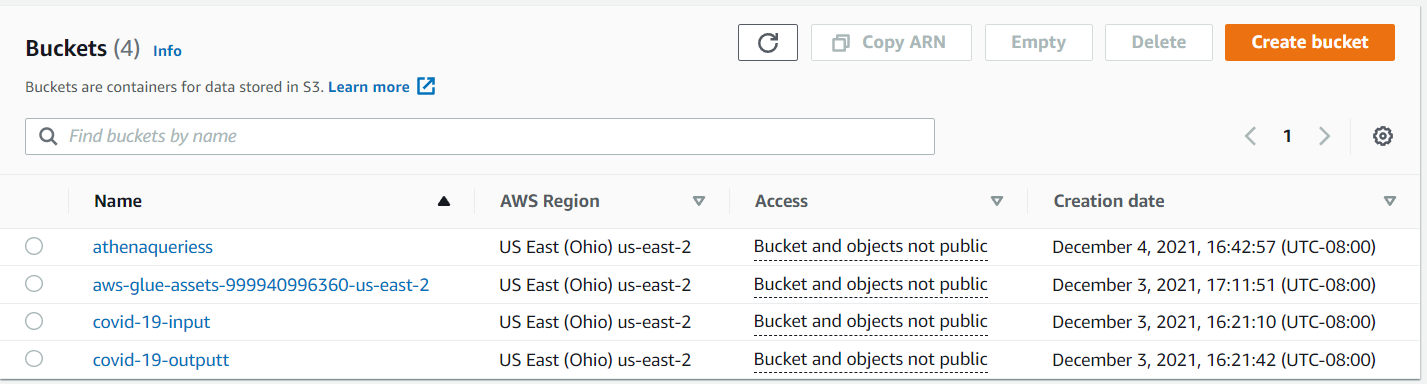


Figure 3. S3 Bucket

## **AWS Glue Crawler:**

Before using crawlers, we need to create a database in the Data Catalog. This will help to store the tables which were created by the crawler in the database.

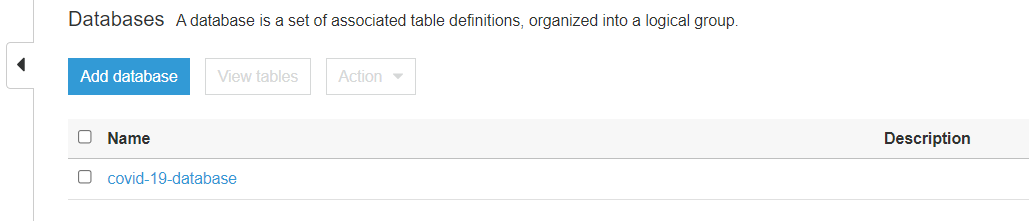


Figure 4. Creating a database for Crawler step

After creating the database, we have created a crawler with Run on-demand specifications. The source was from the S3 input bucket, and the output will be stored in a table in the database.

Once we run the crawler, the crawler will create a table that will be stored in the Database. We can run it hourly or daily if our data is streaming. The crawler will help recognize your data store, which in turn is used to create a schema for the table.

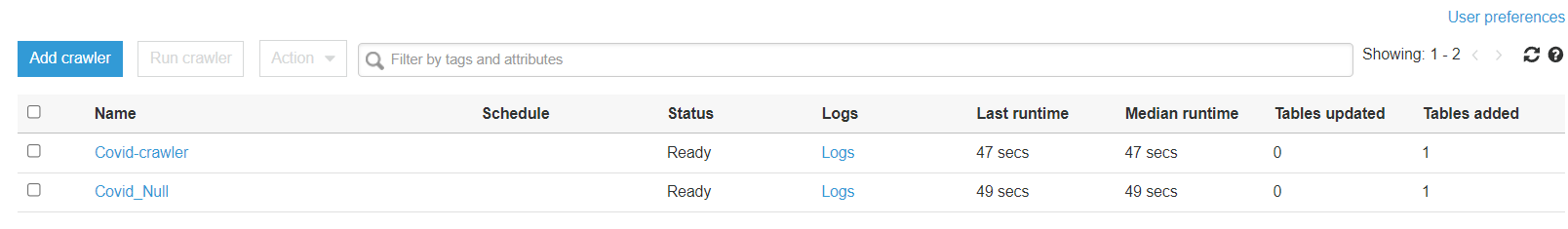


Figure 5. Creating Crawler

The table with the metadata is provided by the crawler as shown below:

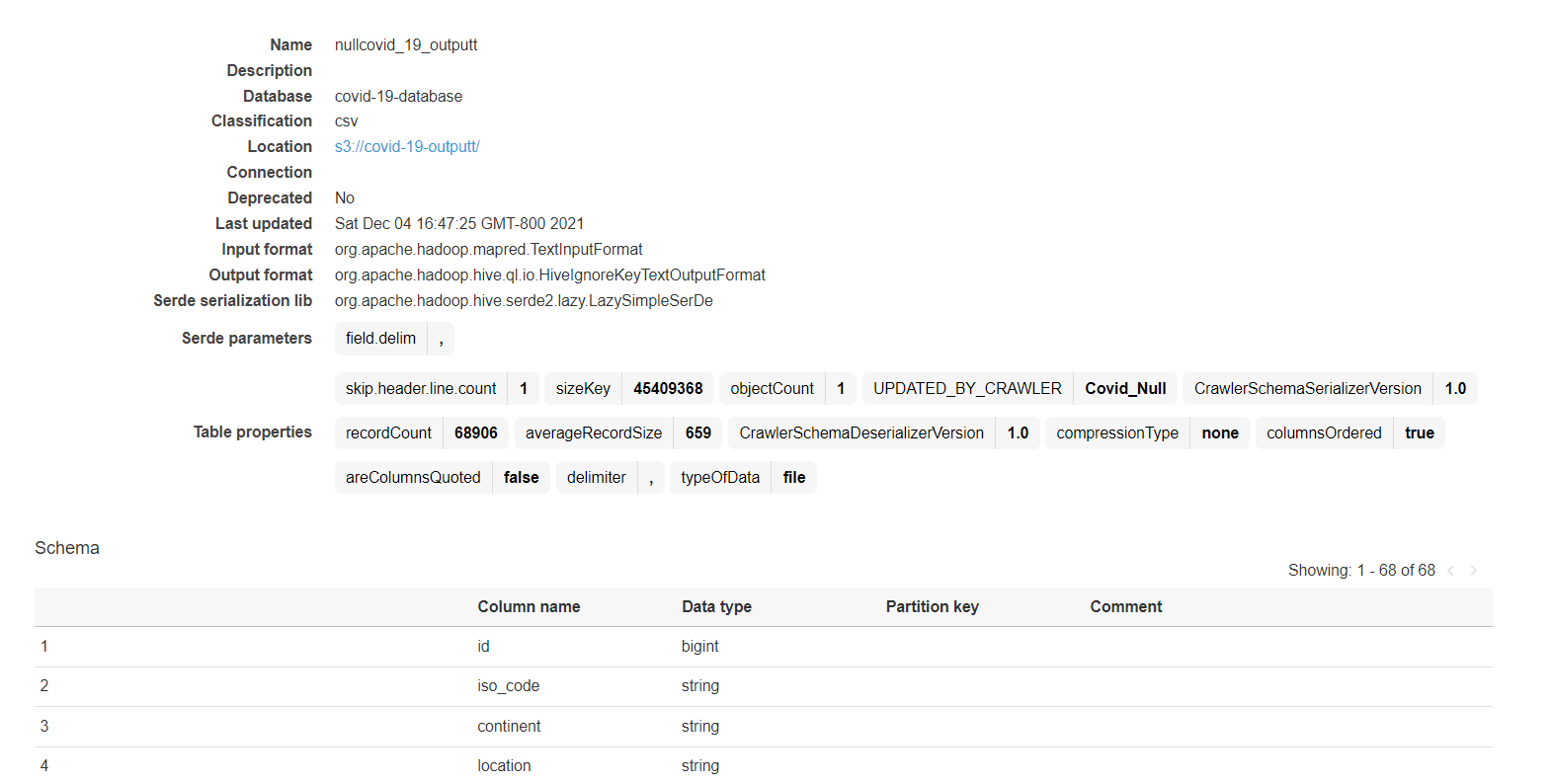


Figure 6. Data loaded by Crawler

## **AWS Glue Job 1:**

The data we have contains 67 columns and more than a hundred thousand rows, where most of the parts are empty.

For instance, from 24th February 2020 to 23rd March 2020 no one dies due to covid in Afghanistan, so the values are empty.

Without filling in these Null values, there would be a problem while querying or visualization. So, one of the best approaches is to clean the data by filling the null values with 0. By using PYSPARK we had written a custom code that will help fill the zeros in the empty places.

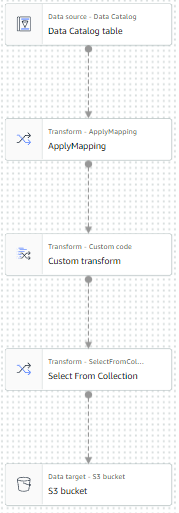


Figure 7. Glue Job 1

Once filling in the null values code is written in the custom transform, we need to export the output in S3 or even in the database. We have stored the output in the S3 output bucket and the database.

After the output is generated, we have checked the data using SQL queries in ATHENA. Once the output shown in Athena is as we expected, we have created another job.

## **AWS Glue Job 2:**

As we have designed a data model, we have divided the dimensions in Glue Job. The first job output is sent to the second job. This second job will help to divide the mail table into different dimensions, that is 11 tables with unique attributes and having surrogate keys.

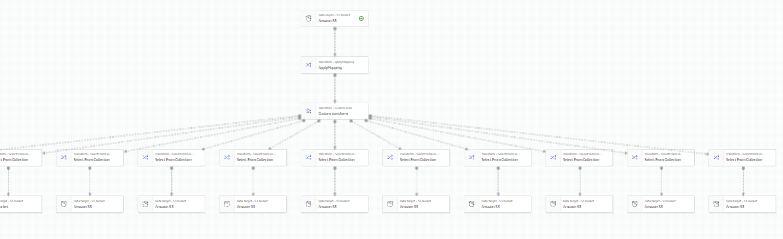


Figure 8. Glue Job 2

For this, we have written a custom code where we used the data frames concept in PySpark. We have divided the main table into 11 different tables, each having its attributes. For instance, the vaccination table has attributes like total vaccination till that day, Number of people vaccinated, and many more. We have separated it into tables for better analytical queries.

The output is sent to the S3 bucket and the database. There will be many files created as there are many tables present.

# **Analytical queries:**

## **Redshift:**

Redshift is a data warehousing tool used for analytics. For this, we need to integrate with the glue. This will help send the data from Glue to redshift. We need to have an IAM role which is having both Glue and Redshift policies. The other way we can do this is by sending the files present in the S3 output bucket to Redshift.

We had created tables in Redshift and inserted the data which came from Job 2. Now after inserting the data into their respective tables, we have started writing analytical queries.

## **Redshift queries:**

This figure shows the output of the query which we executed to find out the top countries with the greatest number of Covid-19 infections.

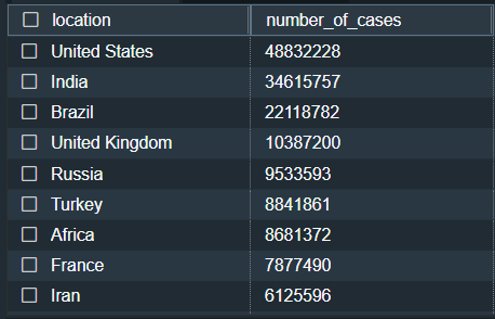


Figure 9. Top 10 countries having most cases

This figure shows the status of vaccination in different countries. Here we have displayed the top 10 countries by the number of people who are not yet fully vaccinated. India has the maximum number of people who are yet not fully vaccinated. India has a long way in front to complete their vaccination.

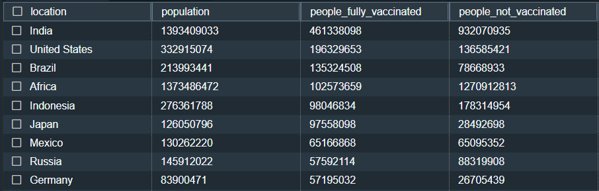


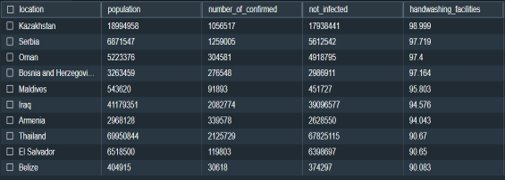
Figure 10. Top 10 countries having most people not fully vaccinated

This figure shows the output of the query to find the top 10 countries with the greatest number of deaths in the Asian continent. Here it can be inferred that India is the worst affected by Covid-19 in the Asian region.



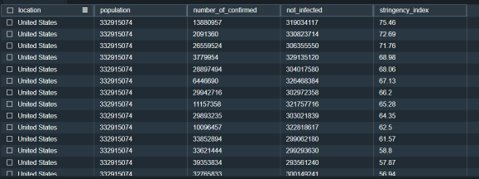
Figure 11. Top 10 countries having most cases in Asia

This tabular data shows the relation between handwashing facilities in a country and the number of confirmed cases. It was observed that countries having a high number of handwashing facilities has a small number of Covid infections compared to countries where handwashing facilities are not adequate.



*Figure 12. Top 10 countries having most handwashing facilities*

This query result shows the relation between the stringency index and the number of cases in the United States. It is observed that as the value of the stringency index increases the number of new confirmed cases decreases. Thus, a country having a high stringency index can help them in reducing the number of new infections.



*Figure 13. Stringency index versus new cases in the United States*

# **Visualization:**

Visualization is an effective way to represent information. An image can represent a piece of information way better than text. So, we created some visualizations to represent some of the trends that we discovered through our analysis.

Map

Description automatically generated

Figure 14. Number cases across the world

A picture containing table

Description automatically generated

The above visualization shows us the number of Covid-19 infections throughout the world. We created a world map and indicated all the confirmed cases for all the countries. For the same, we created a bar chart and from the bar chart, we can see that the United States has the greatest number of confirmed cases with 48 million infections and is followed by India with 34 million cases.

Map

Description automatically generated

Figure 15. Number of deaths

This world map depicts the number of deaths throughout the world. The larger the circle is, the greater the number of deaths is there. We can see that countries like the United States, Brazil, and India are the worst affected countries.

Chart, bar chart

Description automatically generated

Figure 16. Top 10 countries by Death per Millions

This bar graph indicates the top 10 countries with the highest number of deaths per million. Here we can see that smaller countries like Peru and Bulgaria have a high number of deaths per million because while they have a low population, the number of deaths resulting from Covid-19 for them is extremely high. For Peru, they faced 6000 deaths for every 1 million population.

Chart, histogram

Description automatically generated

Figure 17. Number of cases and Stringency index

This dual-axis graph represents the data of stringency index versus new daily infections in the United States. Stringency index represents the measure of response to government policies like lockdown, social distancing, school closures from the public. We can say that overall, the response from the public is good in the United States. As the value of the stringency index increases, the number of new infections is decreasing. So, from this government can decide to strengthen the measures and imply strict rules so that the public follows them and as a result, it will help in limiting the spread of Covid-19.

Chart, bar chart

Description automatically generated

Figure 18. Top 10 countries with number of booster shots

This bar graph indicates the top 10 countries which have administered the maximum number of boosters to the public. Booster shots are given to individuals which are already fully vaccinated. These shots are given to maintain the efficacy of the vaccines in the individuals. It can be observed that China as a country has administered the maximum number of boosters with 65 million shots followed by the United States with 42 million shots followed by the United Kingdom with 19 million shots.

# **Github:**

https://github.com/yallavaibhav/Covid-19\_DataAnalysis-\_AWS

# **Conclusion:**

The analysis of this dataset helped us to understand the basics of COVID-19. It gave us information about the factors that affect the number of cases. This analysis will be useful to make strategies to prevent the worst situations. With the help of this exploratory analysis, people can unders7tand the data about Covid-19 and can make better decisions. From our analysis, we can conclude that by implementing measures like lockdown, social distancing, and proper sanitization we can limit the spread of the COVID-19 virus. It will also help the government to make better policies which could help them tackle such problems in the future.

# **References:**

1. Alsunaidi SJ, Almuhaideb AM, Ibrahim NM, et al. Applications of Big Data Analytics to Control COVID-19 Pandemic. Sensors (Basel). 2021;21(7):2282. Published 2021 Mar 24. doi:10.3390/s21072282
2. Farhadi N, Lahooti H. Are COVID-19 Data Reliable? A Quantitative Analysis of Pandemic Data from 182 Countries. COVID. 2021; 1(1):137-152. https://doi.org/10.3390/covid1010013
3. https://ourworldindata.org/coronavirus
4. Tosi D, Campi A. How Data Analytics and Big Data Can Help Scientists in Managing COVID-19 Diffusion: Modeling Study to Predict the COVID-19 Diffusion in Italy and the Lombardy Region. J Med Internet Res 2020;22(10): e21081. URL: https://www.jmir.org/2020/10/e21081. DOI: 10.2196/21081