# DATA ANALYSIS: COVID -19 VS MONKEYPOX

## PRINCIPLES OF BIG DATA MANAGEMENT

## GROUP 20

Mohammad Shaik
School of Computing and Engineering
University of Missouri-Kansas City
ms6bz@umsystem.edu

Rajesh Tummala
School of Computing and Engineering
University of Missouri-Kansas City
rt9cd@umsystem.edu

Kranthi Mangalagiri
School of Computing and Engineering
University of Missouri-Kansas City
kmq3v@umsystem.edu

Niharika Thakur
School of Computing and Engineering
University of Missouri-Kansas City
ntf7t@umsystem.edu

## ABSTRACT

A Comparative Analysis of Global pandemic COVID 19 and Monkeypox and its Initiatives and the Evolution of Global Transmission examining data to discover the association. Monkeypox and Corona have a significant impact on global population. In this project, we are going to work with the COVID19 & Monkeypox dataset, published by kaggle, which consists of the data related to the increased number of confirmed, recovered & death cases, per day, in each Country. By using Big Data tools, we are going perform data Ingestion, Processing, Analyzing and Visualizing the reports.

## KEYWORDS

Covid-19, Monkeypox, Pyspark, Wrangling, Cleaning, Ingestion, Processing, Analyzing Visualizing.

## INTRODUCTION

Since the sudden pandemic of the infectious COVID-19 virus caused by the SARS-CoV-2 virus, which first emerged in December 2019. The pandemic has caused dramatic loss of life around the world and poses unprecedented challenges to workplaces, public health and food systems. The pandemic has disproportionately harmed the poor and vulnerable, and it threatened to push millions more into poverty.

Monkeypox is a disease of global public health concern because it affects not only countries in West and Central Africa, but also countries around the world. In 2003, the first monkeypox outbreak outside of Africa was in the United States. Monkeypox is a viral zoonotic disease with symptoms similar to smallpox and it is mainly caused by the monkeypox virus. It is transmitted from animals to humans and also from person to person.

## RELATED WORK

In this project, we are going to work with the COVID19 & Monkeypox dataset, published by Kaggle and by creating the notebooks where one can gain Insights from reading COVID19 & Monkeypox data, pivoting the data and preparing it for the analysis by dropping columns and aggregating rows. Performing EDA by plotting correlation between attributes and find attributes that are more related to each other. Deciding on and calculating a good measure for our analysis. Finally, on the processed data of both Covid and Monkeypox, merging two datasets for Visualizing our analysis results. Design and creating ETL Pipeline. Integrating Databricks notebook to ETL Pipeline to automate the Data Preprocessing.

## PROPOSED TECHNIQUE

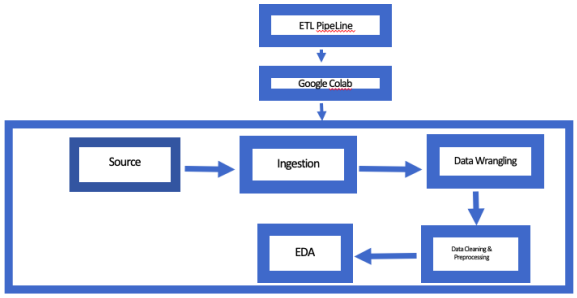



**Figure 1.**

First, we will be creating 4 notebooks each for Covid outbreak analysis, Monkeypox outbreak analysis, Processed data of both Covid and Monkeypox for comparative analysis, Covid outbreak analysis for US counties. The Source datasets which consists of four datasets for Covid-19 for the year of 2020 from January to June and three datasets for Monkeypox which is six months of data. All the notebook follows 3 tier ETL Architecture which involves data ingestion, data cleaning and preprocessing, data

wrangling, complex transformations, merge or load the processed data. By using processed data, EDA has been done and finally, generates reports and insights where one can understand easily through visualization. Finally, On the top that we will be creating Databricks notebooks using Pyspark framework/ Pandas library and then integrating the notebook with ETL pipeline to automate the results. The tools and libraries involves Python(Pyspark framework) notebook, NumPy, pandas libraries for Data Processing, seaborn, matplotlib, plotly for visualizations, Databricks/ Google Colab for notebook creation, ETL pipeline using Azure Data Factory if required.

## DATASETS

***Time_series_covid_19_confirmed***: This dataset contains of list of confirmed cases of all the countries

***Time_series_covid_19_deaths***: This dataset contains of list of deaths occurred in all the countries

***Time_series_covid_19_Recovered:*** This dataset contains of list of recovered cases in all the countries.

***Covid_19_clean_complete***: This dataset contains information related to confirmed, deaths, recovered cases all over the world.

***Monkey_Pox_Cases_Worldwide***: This dataset has information related confirmed and suspected cases all over the world.

***Worldwide_Case_Detection_Timeline***: This dataset contains information about the confirmed cases. It also includes the date and time and other details about each reported case.

***Daily_Country_Wise_Confirmed_Cases:*** This dataset has information about the confirmed cases all over the world daily.

## BIG DATA TECHNIQUES USED

### Data Ingestion

The process of retrieving data from one or more sources/databases and importing data into one consistent database to further process and analyze the data is known as data ingestion. Prioritizing data sources is essential in an efficient data ingestion process.

Here in our project, created notebook for Monkeypox analysis named "**Monkeypox_Outbreak_Analysis.ipynb** ", where four datasets has been ingested into google colab. Imported all the required libraries in a notebook, where we can perform EDA on the top of the processed data. Hence, Ingestion has been done by mounting the datasets.

### Data Wrangling

Data wrangling is the process of transforming wide data into narrow data which makes the user to understand the data and easy to analyze. It is a part of data cleaning or sometimes data wrangling refers to data cleaning or data munging. Moreover, it is used to deal complex data, produce more accurate results, and make better decisions in less time.

| | Province/State | Country/Region | Lat | Long | 1/22/20 | 1/23/20 | 1/24/20 | 1/25/20 |
|---|---|---|---|---|---|---|---|---|
| 0 | NaN | Afghanistan | 33.0000 | 65.0000 | 0 | 0 | 0 | 0 |
| 1 | NaN | Albania | 41.1533 | 20.1683 | 0 | 0 | 0 | 0 |
| 2 | NaN | Algeria | 28.0339 | 1.6596 | 0 | 0 | 0 | 0 |
| 3 | NaN | Andorra | 42.5063 | 1.5218 | 0 | 0 | 0 | 0 |
| 4 | NaN | Angola | -11.2027 | 17.8739 | 0 | 0 | 0 | 0 |

5 rows × 165 columns

**Figure 2.**

As shown in the above Figure 2 there are 5 rows and 165 columns, so by doing data wrangling we reduce the column size to organize the data and to get the better understanding on the data.

| | Province/State | Country/Region | Lat | Long | Date | Confirmed | Deaths | Recovered |
|---|---|---|---|---|---|---|---|---|
| 0 | NaN | Afghanistan | 33.0000 | 65.0000 | 1/22/20 | 0 | 0 | 0.0 |
| 1 | NaN | Albania | 41.1533 | 20.1683 | 1/22/20 | 0 | 0 | 0.0 |
| 2 | NaN | Algeria | 28.0339 | 1.6596 | 1/22/20 | 0 | 0 | 0.0 |
| 3 | NaN | Andorra | 42.5063 | 1.5218 | 1/22/20 | 0 | 0 | 0.0 |
| 4 | NaN | Angola | -11.2027 | 17.8739 | 1/22/20 | 0 | 0 | 0.0 |

```
covid_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42826 entries, 0 to 42825
Data columns (total 8 columns):
```

**Figure 3.**

So here after doing the data wrangling, we can see the changes in the above Figure 3 that the column size is reduced to 8 based upon the date field it has been categorized to give the Confirmed, Deaths and Recovered.

### Data Cleaning and Preprocessing

The process of fixing missing data, removing incorrect, duplicate or irrelevant data from a data set is known as data cleaning. Data preprocessing is the process of converting a raw dataset into an understandable format. Like data cleaning, it ensures that your data is ready for use in the future.

The following are the few cleaning and preprocessing steps performed in notebook:

- Renaming column names.
- Check null values in the dataset and fill those null or missing values as per analysis.
- Modifying the datatypes of a dataset.
- Removing Uncertainty or Inconsistence in data

```
covid_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42826 entries, 0 to 42825
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Province/State  13041 non-null  object 
 1   Country/Region  42826 non-null  object 
 2   Lat             42826 non-null  float64
 3   Long            42826 non-null  float64
 4   Date            42826 non-null  object 
 5   Confirmed       42826 non-null  int64  
 6   Deaths          42826 non-null  int64  
 7   Recovered       40733 non-null  float64
dtypes: float64(3), int64(2), object(3)
memory usage: 2.6+ MB
```

**Figure 4. Covid_Table_Info**

As shown in the Figure 4, the dataset consists of 42826 entries and 8 columns. We need to rename the column names for some of them and we can see null values, invalid datatypes for the columns those must be modified.

```
# Reading the dataset information after Data Preprocessing
covid_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42826 entries, 0 to 42825
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype         
---  ------     --------------  -----         
 0   State      42826 non-null  object        
 1   Country    42826 non-null  object        
 2   Lat        42826 non-null  float64       
 3   Long       42826 non-null  float64       
 4   Date       42826 non-null  datetime64[ns]
 5   Confirmed  42826 non-null  int64         
 6   Deaths     42826 non-null  int64         
 7   Recovered  42826 non-null  int64         
dtypes: datetime64[ns](1), float64(2), int64(3), object(2)
memory usage: 2.6+ MB
```

**Figure 5.**

In Figure 5, here by doing data cleaning and preprocessing we changed the column names Province/State to State, Country/Region to Country, Null values are fixed for all the columns, for the field Date the datatype was object now it has been changed to datetime64 and for the Recovered field datatype is changed to int64.

## EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is used to analyze data sets and draw insights from the data often using data visualization methods. EDA is an important step in data analysis or any data science projects often used to determine how to manipulate data sources to discover patterns and anomalies and validate hypothesis by understanding the data.

| Country | State | Confirmed | Deaths | Recovered |
|---|---|---|---|---|
| Afghanistan | NA | 31517 | 746 | 0 |
| Albania | NA | 2535 | 62 | 0 |
| Algeria | NA | 13907 | 912 | 0 |
| Andorra | NA | 855 | 52 | 0 |
| Angola | NA | 284 | 13 | 0 |
| Antigua and Barbuda | NA | 69 | 3 | 0 |
| Argentina | NA | 64530 | 1307 | 0 |
| Armenia | NA | 25542 | 443 | 0 |
| Australia | Australian Capital Territory | 108 | 3 | 0 |
| | New South Wales | 3203 | 49 | 0 |
| | Northern Territory | 29 | 0 | 0 |
| | Queensland | 1067 | 6 | 0 |
| | South Australia | 443 | 4 | 0 |
| | Tasmania | 228 | 13 | 0 |
| | Victoria | 2231 | 20 | 0 |
| | Western Australia | 611 | 9 | 0 |

**Figure 6.**

In Figure 6, we are merging all the datasets of Confirmed cases, Deaths cases and Recovered cases datasets into the one final data frame to analyze the number of cases for each country for Confirmed, Deaths and Recovered.

| | Country | Confirmed |
|---|---|---|
| 0 | US | 2635417 |
| 1 | Brazil | 1402041 |
| 2 | Russia | 646929 |
| 3 | India | 585481 |
| 4 | United Kingdom | 314160 |
| 5 | Peru | 285213 |
| 6 | Chile | 279393 |
| 7 | Spain | 249271 |
| 8 | Italy | 240578 |
| 9 | Iran | 227662 |

| | Country | Deaths |
|---|---|---|
| 0 | US | 127417 |
| 1 | Brazil | 59594 |
| 2 | United Kingdom | 43815 |
| 3 | Italy | 34767 |
| 4 | France | 29846 |
| 5 | Spain | 28355 |
| 6 | Mexico | 27769 |
| 7 | India | 17400 |
| 8 | Iran | 10817 |
| 9 | Belgium | 9747 |

**Figure 7. Covid-19 Top 10 Confirmed and Deaths**

In Figure 7, we have done an analysis to show the top 10 Countries which are being affected by the Covid-19 with most number of the Confirmed cases and Death cases.

| | Date | Confirmed | Deaths | Recovered | Active |
|---|---|---|---|---|---|
| 0 | 2020-06-30 00:00:00 | 10475085 | 511237 | 5283066 | 4680782 |

**Figure 8.**

Similarly, data ingestion, data cleaning and preprocessing, EDA has been done for covid_19_clean_complete which is fourth dataset to analyze and to show the total number of Confirmed, Deaths & Recovered Cases as per the date which is for june2020 from Figure 8.

## DATA VISUALIZATION

Visualizing the data either in charts or graphs where the user or business stakeholder can easily understand the data. This process is known as data visualization. There are various tools and libraries to visualize the results. For our results, we used libraries.

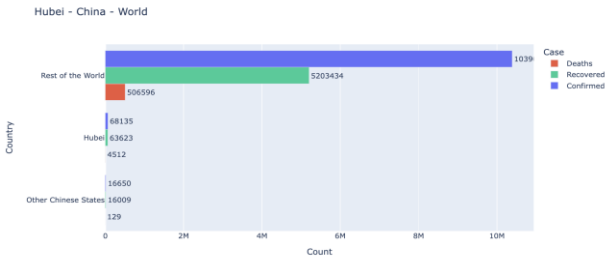



**Figure 9.**

In Figure 9, here by performing data visualization, As the first case was found in Hubei, Wuhan. We are plotting the number of cases by comparing the cases with respect to the Wuhan state, other Chinese states as well as rest of the world for the month of June2020.

For Monkeypox Analysis, created notebook named "**Monkeypox_Outbreak_Analysis.ipynb**" similar to Covid where data ingestion, cleaning and preprocessing has been done as follows:

```
[ ]  # Reading the dataset information before Data Preprocessing
     df_worldwide_cases.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 115 entries, 0 to 114
     Data columns (total 6 columns):
      #   Column              Non-Null Count  Dtype
     ---  ------              --------------  -----
      0   Country             115 non-null    object
      1   Confirmed_Cases     115 non-null    float64
      2   Suspected_Cases     115 non-null    float64
      3   Hospitalized        115 non-null    float64
      4   Travel_History_Yes  115 non-null    float64
      5   Travel_History_No   115 non-null    float64
     dtypes: float64(5), object(1)
     memory usage: 5.5+ KB
```

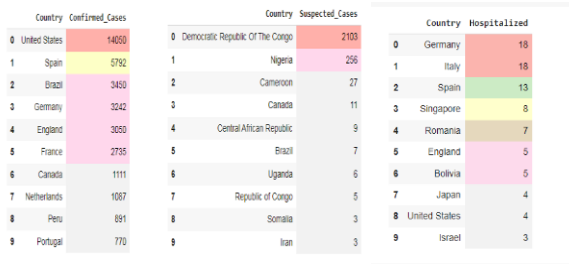
**Figure 10. Monkeypox_Worldwide_cases before Data Proccessing**

```
[10] # Reading the dataset information after Data Preprocessing
     df_worldwide_cases.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 115 entries, 0 to 114
     Data columns (total 6 columns):
      #   Column              Non-Null Count  Dtype
     ---  ------              --------------  -----
      0   Country             115 non-null    object
      1   Confirmed_Cases     115 non-null    int64
      2   Suspected_Cases     115 non-null    int64
      3   Hospitalized        115 non-null    int64
      4   Travel_History_Yes  115 non-null    int64
      5   Travel_History_No   115 non-null    int64
     dtypes: int64(5), object(1)
     memory usage: 5.5+ KB
```

**Figure 11. Monkeypox_Worldwide_cases after Data Proccessing**

From Figure 10 and 11, you can see the modifications i.e. datatypes has been changed.



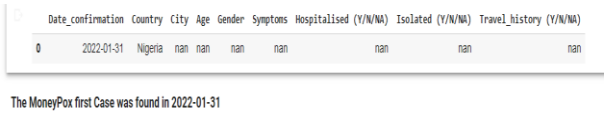
| | Country | Confirmed_Cases |
|---|---|---|
| 0 | United States | 14050 |
| 1 | Spain | 5792 |
| 2 | Brazil | 3450 |
| 3 | Germany | 3242 |
| 4 | England | 3050 |
| 5 | France | 2735 |
| 6 | Canada | 1111 |
| 7 | Netherlands | 1087 |
| 8 | Peru | 891 |
| 9 | Portugal | 770 |

| | Country | Suspected_Cases |
|---|---|---|
| 0 | Democratic Republic Of The Congo | 2103 |
| 1 | Nigeria | 256 |
| 2 | Cameroon | 27 |
| 3 | Canada | 11 |
| 4 | Central African Republic | 9 |
| 5 | Brazil | 7 |
| 6 | Uganda | 6 |
| 7 | Republic of Congo | 5 |
| 8 | Somalia | 3 |
| 9 | Iran | 3 |

| | Country | Hospitalized |
|---|---|---|
| 0 | Germany | 18 |
| 1 | Italy | 18 |
| 2 | Spain | 13 |
| 3 | Singapore | 8 |
| 4 | Romania | 7 |
| 5 | England | 5 |
| 6 | Bolivia | 5 |
| 7 | Japan | 4 |
| 8 | United States | 4 |
| 9 | Israel | 3 |

**Figure 12. Monkeypox Top10 Countries for Confirmed, Suspected and Hospitalized cases**

From Figure 10, data analysis has been done on the processed datasets for Confirmed, Suspected and Hospitalized cases where you can see the top 10 countries which are affected by Monkeypox by grouping the country column.



| | Date_confirmation | Country | City | Age | Gender | Symptoms | Hospitalised (Y/N/NA) | Isolated (Y/N/NA) | Travel_history (Y/N/NA) |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2022-01-31 | Nigeria | nan | nan | nan | nan | nan | nan | nan |

The MoneyPox first Case was found in 2022-01-31

**Figure 13.**

From other data "Daily_Country_Wise_Conformed_Cases" , after preprocessing, then EDA has been done. Through this analysis, we found the first case in the world.

## CONTRIBUTION

Equal contribution has been done by all the group members, but each of the team members took major focus on each implementation individually. Design and Architecture has been implemented by Mohammad as well as tools and libraries needed for the project. Created "Corona_Virus_OutBreak_Analysis.ipynb" notebook where ingestion, data wrangling, data cleaning and preprocessing, analysis is a part of Mohammad's contribution. Therefore, various sources have been transformed into one processed(target) dataset. EDA on the processed data has been done by Rajesh and reported analysis. Moreover, deep research work has been done by the Rajesh for finalizing the datasets. Kranthi had worked on generating Monkeypox datasets and took major focus on entire Monkeypox outbreak analysis where he created "Monkeypox_Outbreak_Analysis.ipynb" which involved from Ingestion to processed data. Further, EDA on the processed data has been done by Niharika along with Visualization and reports. Moreover, her major focus is on presentation and documentation.