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 Kumar 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, PY Spark, Wrangling, Cleaning, Ingestion, Processing, Analyzing Visualizing.

INTRODUCTION

Since the unexpected COVID-19 pandemic that was brought on by the SARS-CoV-2 virus that first appeared in December 2019. The epidemic has dramatically increased the number of deaths worldwide and presents previously unheard-of difficulties for the food, public health, and employment environments. In addition to countries in West and Central Africa, Monkeypox affects nations all over the world, making it a disease of concern for global public health. In the United States in 2003, there occurred the first outbreak of monkeypox outside of Africa. It is primarily the monkeypox virus that causes monkeypox, a viral zoonotic illness with symptoms like smallpox. It spreads both between people and animals as well as between individuals.

In this project we are making an analysis of coronavirus and the monkey of first six months data by following ETL flow and perform analysis which we bring the outbreak analysis in both pandemics. Moreover, we will perform comparative visualizations for both processed datasets by using matplotlib, plotly and seaborn.

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

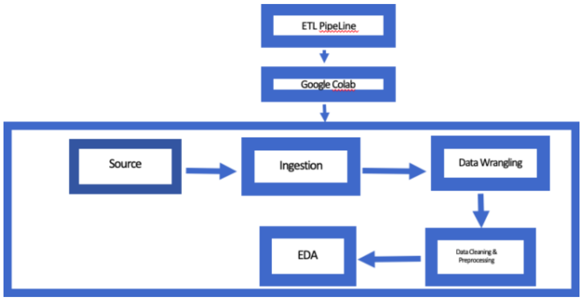


Figure 1.

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 consist 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 Panda's library and then integrating the notebook with ETL pipeline to automate the results. The tools and libraries involve Python (Pyspark framework) notebook, NumPy, pandas libraries for Data Processing, seaborn, matplotlib, plotly for visualizations, Databricks/ Google Collab for notebook creation, ETL pipeline using Azure Data Factory if required.

DATASETS

Covid 19 Data Set:

Time_series_covid_19_confirmed: This data set contains list of confirmed cases in all the counties.

Time_series_covid_19_deaths: This data set contains 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.

covid_19_data.csv: This data set contains all the information of confirmed, dead and recovered cases with state, country, and latest update.

Monkey Pox Data Set:

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.

Covid 19 and Monkey Pox

Covid_processed_data.csv: This data set contains the information about the processed and analyzed information of confirmed, deaths, recovered and observation date of corona virus.

Monkeypox_processed.csv: This data set contains the information about the processed and analyzed information of confirmed, deaths, recovered and observation date of Monkey pox.

BIG DATA TECHNIQUES USED

COVID-19 Outbreak Analysis

Created a "**Corona_Virus_OutBreak_Analysis.ipynb**" notebook where the Kaggle data set has been extracted into the local drive and performed various transformations using complex queries on multiple data sources (data sets). Finally, the Processed data has been loaded as a CSV file for further comparative analysis. Upon the processed data exploratory data analysis has been done and platted a graph where one can easily understand through this visualization.

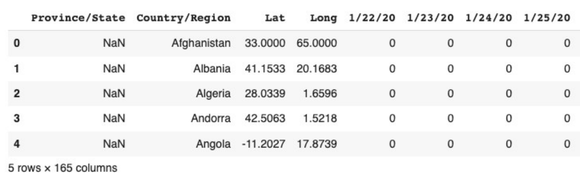
Data Ingestion

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, firstly we have created a notebook for COVID-19 outbreak analysis named that as "**COVID-19_Outbreak_Analysis.ipynb** ", where four datasets has been ingested into google Collab. 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.

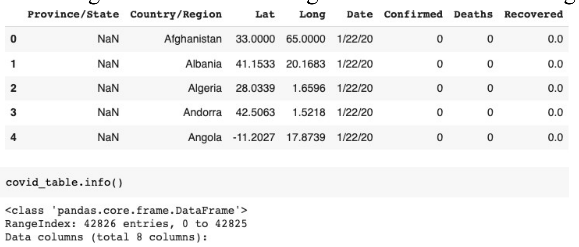


	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20
0	NaN	Afghanistan	33.0000	65.0000	0	0	0	0
1	NaN	Albania	41.1533	20.1683	0	0	0	0
2	NaN	Algeria	28.0339	1.6596	0	0	0	0
3	NaN	Andorra	42.5063	1.5218	0	0	0	0
4	NaN	Angola	-11.2027	17.8739	0	0	0	0

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.

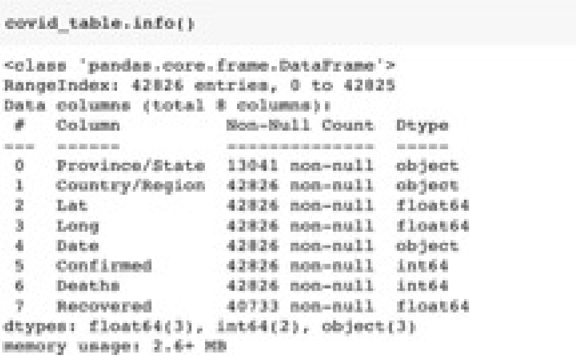
So here after doing the data wrangling, we can see the changes in the above Figure 3 that the column size is reduced to 8 based upon the date field it has been categorized to give the Confirmed, Deaths and Recovered.

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

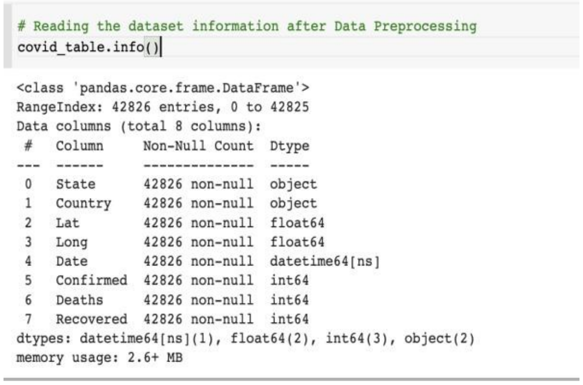


```
covid_table.info()
```

```
<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()
```

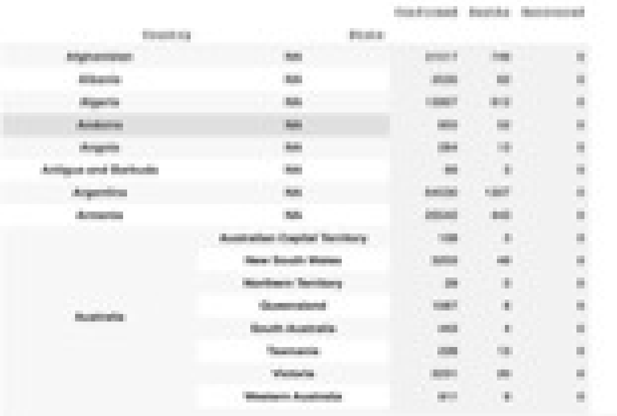
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42826 entries, 0 to 42825
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   State      42826 non-null  object
 1   Country    42826 non-null  object
 2   Lat        42826 non-null  float64
 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

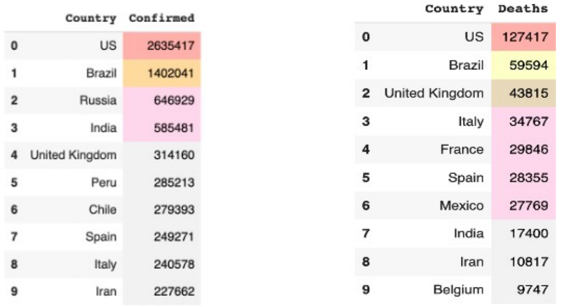
Exploratory Data Analysis is used to analyze data sets and draw insights from the data often using data visualization methods. EDA is an important step in data analysis, or any data science projects often used to determine how to manipulate data sources to discover patterns and anomalies and validate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



Country	State	Confirmed	Deaths	Recovered
Afghanistan	NA	[illegible]	[illegible]	0
Albania	NA	[illegible]	[illegible]	0
Algeria	NA	[illegible]	[illegible]	0
Andorra	NA	[illegible]	[illegible]	0
Angola	NA	[illegible]	[illegible]	0
Antigua and Barbuda	NA	[illegible]	[illegible]	0
Argentina	NA	[illegible]	[illegible]	0
Armenia	NA	[illegible]	[illegible]	0
Australia	Australian Capital Territory	[illegible]	[illegible]	0
	New South Wales	[illegible]	[illegible]	0
	Northern Territory	[illegible]	[illegible]	0
	Queensland	[illegible]	[illegible]	0
	South Australia	[illegible]	[illegible]	0
	Tasmania	[illegible]	[illegible]	0
	Victoria	[illegible]	[illegible]	0
	Western Australia	[illegible]	[illegible]	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.

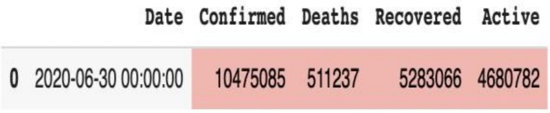


	Country	Confirmed
0	US	2635417
1	Brazil	1402041
2	Russia	646929
3	India	585481
4	United Kingdom	314160
5	Peru	285213
6	Chile	279393
7	Spain	249271
8	Italy	240578
9	Iran	227662

	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

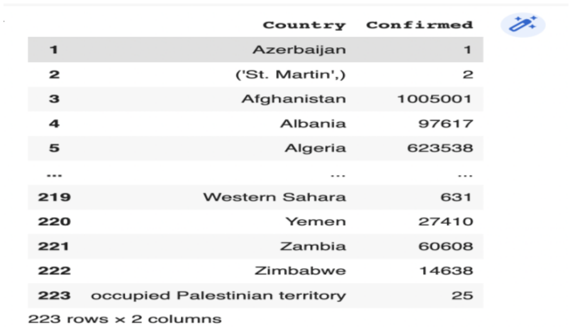
- Most of the positive cases were found in USA followed by Brazil and Russia
- Few countries are infecting the neighboring countries of China.
- India has a positive case of 585481 which is 4th highest among all countries.
- Outside China, particularly more positive cases were found in Italy and Spain.
- Similar to positive cases, most of the deaths cases were found in USA followed by Brazil and UK which is 3rd highest in deaths reported.
- India has a total death case of 17400 which is 7th highest among all countries.
- Outside China, particularly more death cases were found in Italy and Spain.



	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.



	Country	Confirmed
1	Azerbaijan	1
2	('St. Martin',)	2
3	Afghanistan	1005001
4	Albania	97617
5	Algeria	623538
...	...	...
219	Western Sahara	631
220	Yemen	27410
221	Zambia	60608
222	Zimbabwe	14638
223	occupied Palestinian territory	25

223 rows × 2 columns

Figure 9.

In figure 9, We conducted research to indicate the number of Covid 19 cases that have been officially confirmed in each nation as of the most recent date.

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.

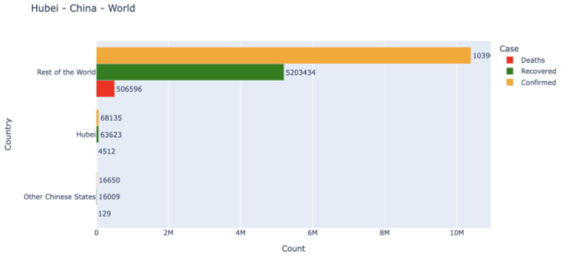


Figure 10.

In Figure 10, 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.

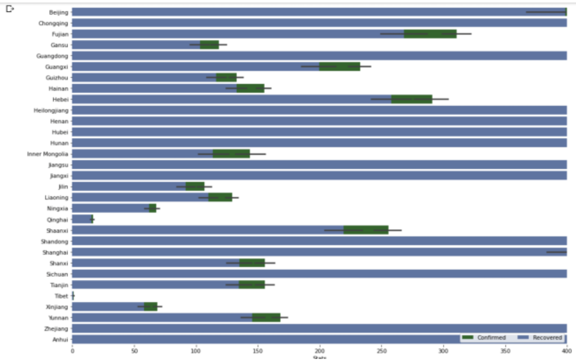
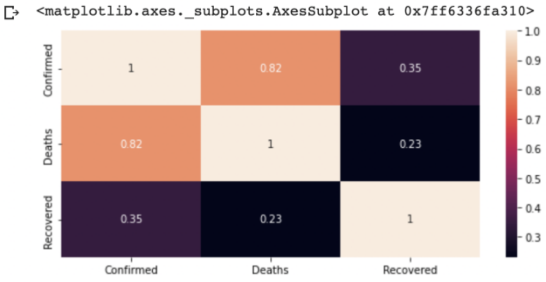


Figure 11.

In Figure 11, From the above plot, if you take "Yunnan", there are nearly 170 confirmed cases out of which only 20 were recovered. Similarly, the plot indicates the same for the rest of the states



```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff6336fa310>
```

Figure 12. Correlation heat map between Confirmed, Death and Recovered Cases

Correlation is a factor to find relationship between each attribute with the target feature. It ranges from 0-1. Higher the correlation factor higher the dependency. A good correlation ranges from 0.5 and vice-versa to the opposite.

In Figure 12, A correlation matrix between three discrete dimensions is displayed in a heatmap, and we can see that Deaths and Confirmed Cases are linearly connected, according to the correlation heat map. High fatality rates in the early phases were caused by the lack of vaccine, effective treatment, or knowledge.

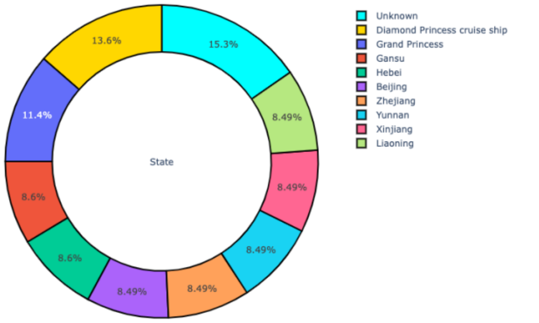


Figure 13. Percentage of confirmed covid cases for ten States

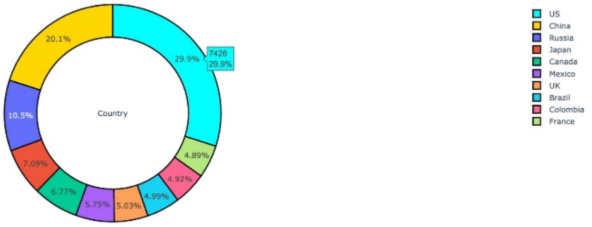
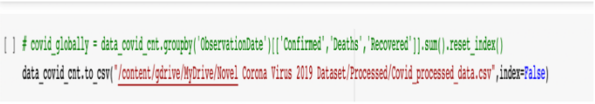


Figure 14. Percentage of confirmed covid cases for top ten countries

confirmed cases

The above Pie Plots representation shows the number of times that event i.e., Confirmed, Deaths, Recovered Cases etc.. has been occurred in the dataset which indicates the count corresponding to its percentage.

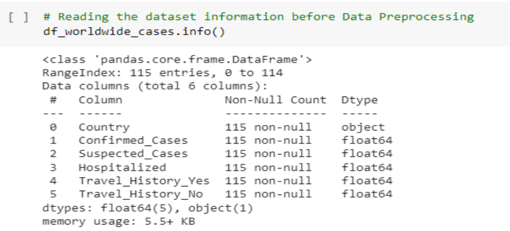


```
[ ] # covid_globally = data_covid_cnt.groupby('ObservationDate')[['Confirmed','Deaths','Recovered']].sum().reset_index()
    data_covid_cnt.to_csv("/content/gdrive/MyDrive/Novel Corona Virus 2019 Dataset/Processed/Covid_processed_data.csv",index=False)
```

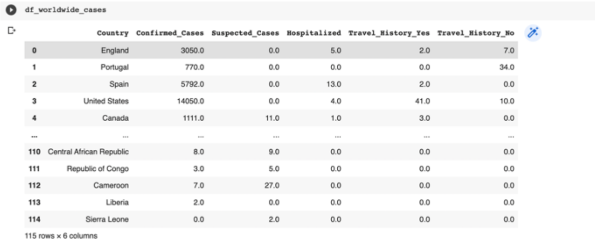
Figure 15. Write the Processed Data to a CSV file

Monkeypox Outbreak Analysis:

For Monkeypox Analysis, created notebook named "**Monkeypox_Outbreak_Analysis.ipynb**" like Covid where data ingestion, cleaning and preprocessing has been Done through ETL and finally the processed has been loaded as CSV file. Upon the transformed data, data analysis and visualizations has been made.



```
[ ] # 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
```



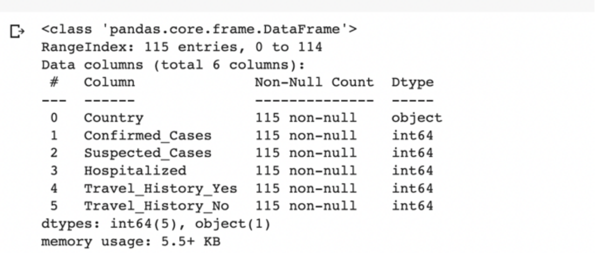
```
df_worldwide_cases
```

	Country	Confirmed_Cases	Suspected_Cases	Hospitalized	Travel_History_Yes	Travel_History_No
0	England	3050.0	0.0	5.0	2.0	7.0
1	Portugal	770.0	0.0	0.0	0.0	34.0
2	Spain	5792.0	0.0	13.0	2.0	0.0
3	United States	14050.0	0.0	4.0	41.0	10.0
4	Canada	1111.0	11.0	1.0	3.0	0.0
...	...	...	...	...	...	...
110	Central African Republic	8.0	9.0	0.0	0.0	0.0
111	Republic of Congo	3.0	5.0	0.0	0.0	0.0
112	Cameroon	7.0	27.0	0.0	0.0	0.0
113	Liberia	2.0	0.0	0.0	0.0	0.0
114	Sierra Leone	0.0	2.0	0.0	0.0	0.0

115 rows × 6 columns

Figure 16. Monkeypox_Worldwide_cases before Data Processing

- From the above dataset, we can see invalid date types for the columns.
- So, Data Cleaning and Data Pre-processing need to be performed on the dataset.



```
<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
```

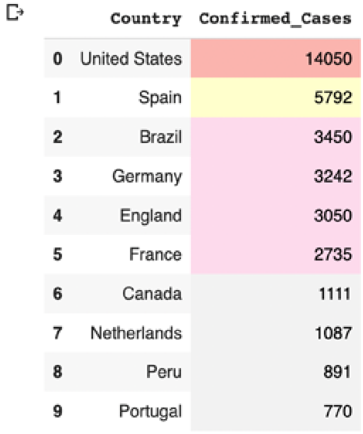
Figure 17. Monkeypox_Worldwide_cases after Data Processing

From Figure 16 and 17, you can see the modifications i.e. datatypes has been changed.

EXPLORATORY DATA ANALYSIS

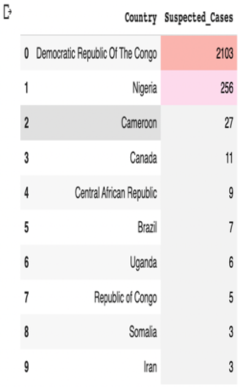
Data sets are analyzed using exploratory data analysis, which frequently employs data visualization techniques to draw conclusions from the data. EDA is a crucial stage in data analysis and is frequently used to find out how to modify data sources to find patterns and anomalies and validate hypotheses by comprehending the data.

Top 10 Countries with most no. of Confirmed cases

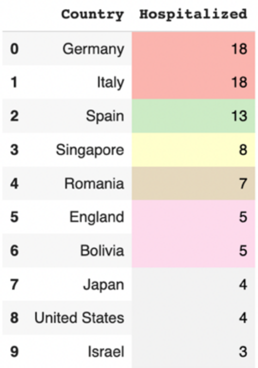


	Country	Confirmed_Cases
0	United States	14050
1	Spain	5792
2	Brazil	3450
3	Germany	3242
4	England	3050
5	France	2735
6	Canada	1111
7	Netherlands	1087
8	Peru	891
9	Portugal	770

Top 10 Countries with most no. of Suspected and no. of Hospitalized cases



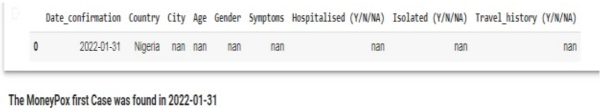
	Country	Suspected_Cases
0	Democratic Republic Of The Congo	2103
1	Nigeria	256
2	Cameroon	27
3	Canada	11
4	Central African Republic	9
5	Brazil	7
6	Uganda	6
7	Republic of Congo	5
8	Somalia	3
9	Iran	3



	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

The above plots show the following:

- The greatest number of Confirmed Cases can be found in United States with 1450 cases followed by Spain, Brazil etc. The Confirmed Cases of USA is almost the double to that of Spain.
- The greatest number of suspected Cases can be found in Democratic Republic of The Congo with 2103 cases. The only country with more than 1000 cases.
- From the Hospitalized plot only, few cases have been identified with Germany and Italy being the top with 18 cases followed by Spain.



	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 18. Monkeypox First Case

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.

DATA VISUALIZATION:

For the data visualization since we have reparented the corona virus in pie and bar chart representations for monkey pox we have used bar, geometrical and bar plot representations.

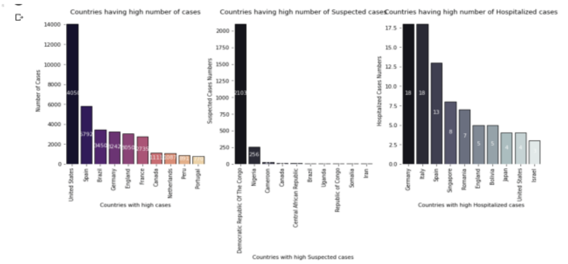


Figure 19. Top 10 countries with highest number of cases Worldwide

In figure 19 Here, we can see a case analysis. There are three graphs that we can see, and the first one shows the countries that have cases, the second shows the countries that have suspected cases, and the third shows the countries that have hospitalized cases.

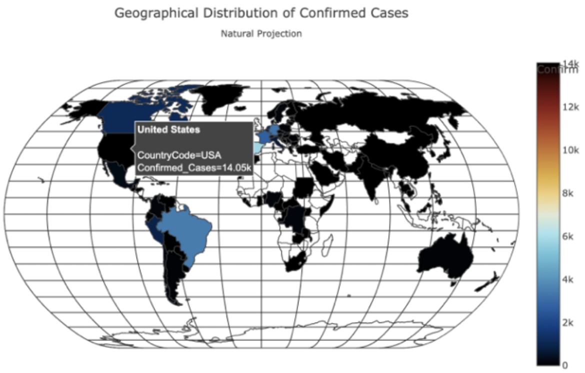


Figure 20. Geographical Distribution of Confirmed Cases Natural Projection

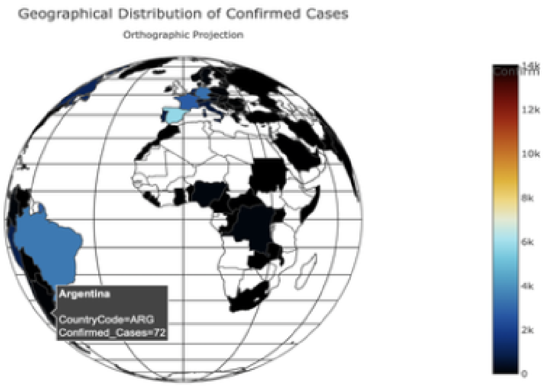


Figure 21. Geographical Distribution of Confirmed Cases Orthographic Projection

In figure 24 & 25 above, we can see the visualization of the Confirmed cases across all the countries in a Geographical Distribution Natural Projection and Orthographic Projection. When we move the cursor on each country you can see the count of confirmed cases.

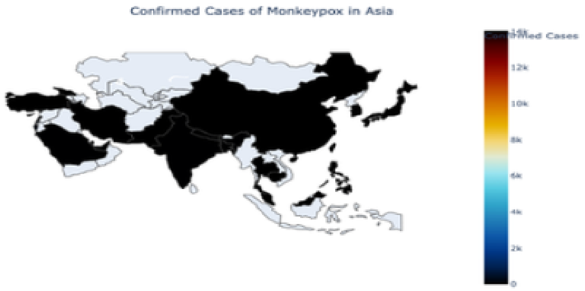


Figure 22. Geographical Distribution of Confirmed Cases for Asia

As shown in Figure 22, we can see the visualization of the Confirmed cases for Asia in a Geographical Distribution. When we move the cursor on each country you can see the count of confirmed cases. And as shown in Figure 21, we can see the Confirmed Cases for Top 3 Continents in Map representation.

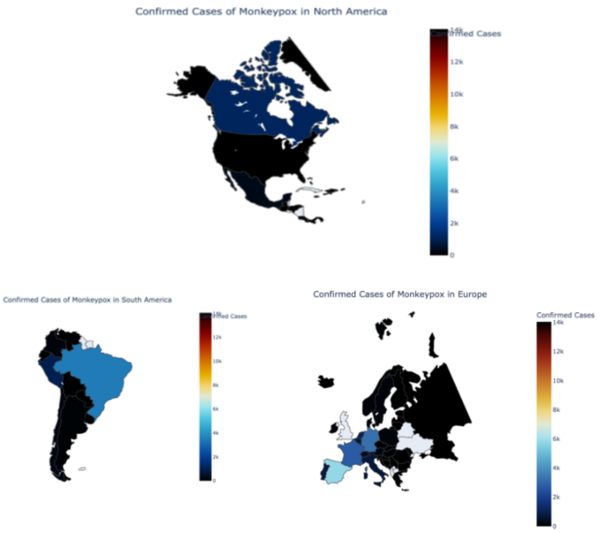

Confirmed Cases of Monkeypox in North America
Confirmed Cases
Confirmed Cases of Monkeypox in South America
Confirmed Cases of Monkeypox in Europe
Confirmed Cases

Figure 23. Map Representation of Confirmed Cases for Top 3 Continents

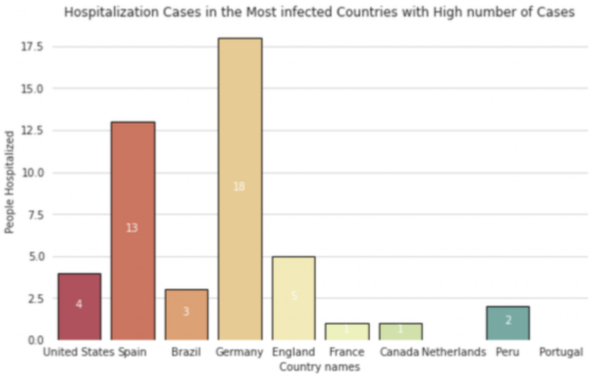


Figure 24. Bar plot representation of Hospitalized Cases for Top 10 most infected Countries.

Hospitalized cases in the most infected Countries with the high number of cases which are represented as bar graph, in which Germany is in the Top1 are shown in Figure 24.

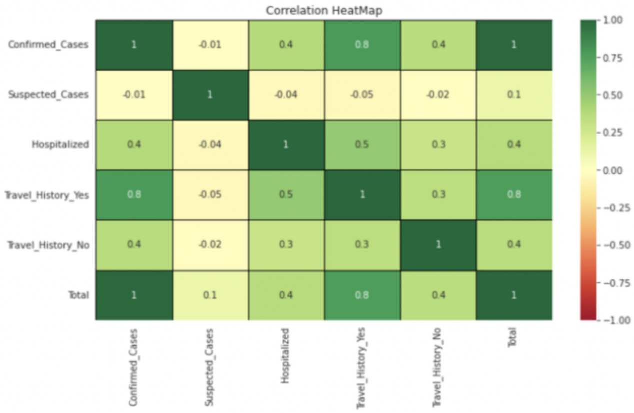
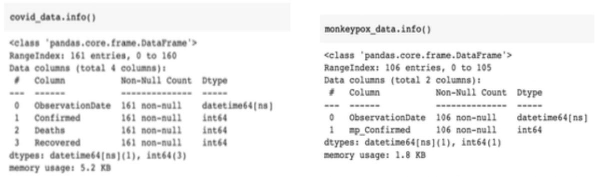


Figure 25. Correlation Heat Map

As shown in Figure25, from Correlation heat map, Travel history is highly correlated to Confirmed Cases i.e., Confirmed Cases are highly dependent upon Traveling history where traveling has a high impact on spreading of virus.

Covid-19 V/S Monkey Pox Outbreak Analysis

We have performed the Covid Vs Monkeypox analysis with the processed data which we have in the form of CSV in both covid, and Monkey pox were ingested into the third notebook and two data sets are mentioned below.
In the next step we are going to merge the two data sets and perform the comparative analysis on both.

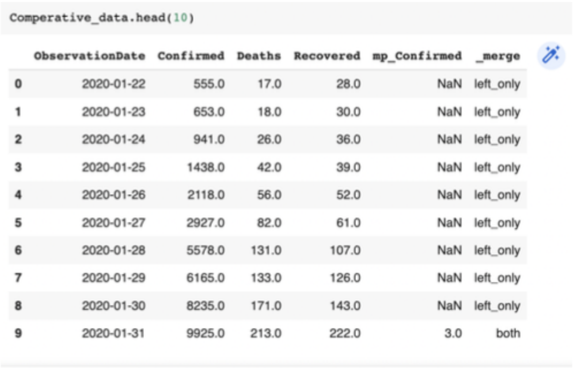


```
covid_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 161 entries, 0 to 160
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ObservationDate  161 non-null    datetime64[ns]
 1   Confirmed        161 non-null    int64
 2   Deaths           161 non-null    int64
 3   Recovered        161 non-null    int64
dtypes: datetime64[ns](1), int64(3)
memory usage: 5.2 KB
```

```
monkeypox_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 106 entries, 0 to 105
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ObservationDate  106 non-null    datetime64[ns]
 1   mp_Confirmed     106 non-null    int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 1.8 KB
```

Figure 26: Covid-19 and Monkeypox processed data information

The Covid -19 data and the Monkeypox data are processed as shown in Figure 26 and the merging of the Covid and Monkeypox data is processed as shown in Figure 27.



```
Comperative_data.head(10)
```

	ObservationDate	Confirmed	Deaths	Recovered	mp_Confirmed	_merge
0	2020-01-22	555.0	17.0	28.0	NaN	left_only
1	2020-01-23	653.0	18.0	30.0	NaN	left_only
2	2020-01-24	941.0	26.0	36.0	NaN	left_only
3	2020-01-25	1438.0	42.0	39.0	NaN	left_only
4	2020-01-26	2118.0	56.0	52.0	NaN	left_only
5	2020-01-27	2927.0	82.0	61.0	NaN	left_only
6	2020-01-28	5578.0	131.0	107.0	NaN	left_only
7	2020-01-29	6165.0	133.0	126.0	NaN	left_only
8	2020-01-30	8235.0	171.0	143.0	NaN	left_only
9	2020-01-31	9925.0	213.0	222.0	3.0	both

Figure 27: Merged data frame of Covid-19 and Monkeypox processed data

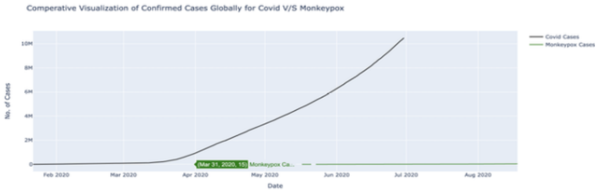


Figure 28: Comparative Visualization of Confirmed Case Globally for Covid V/S Monkeypox

The cumulative number of confirmed cases for Covid V/S Monkeypox are shown as comparative visualization in Figure 28, the black line indicates the Covid Cases, and the green line indicates Monkeypox cases.

From the above comparative Visualization,

- Covid Cases were in the range of Milionis, but Monkeypox data was in the range of thousands.
- There is continuous increase in Covid cases. The Covid bar can be viewed clearly but not Monkeypox bar.
- Hence the visualizations is not effective

In the next visualization plot, lets investigate Monkeypox bar to understand clearly

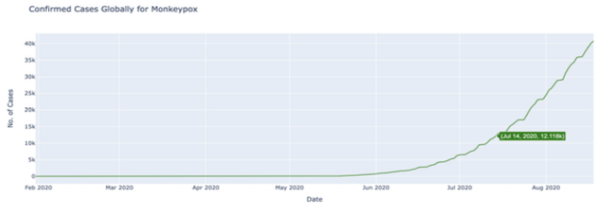


Figure 29: Cumulative Number of Confirmed Cases over time for Monkeypox

The cumulative number of confirmed cases as overtime for Monkeypox are shown in Figure 29, as highest number of confirmed cases globally for Monkeypox has reached to Forty Thousand.

- Here we can see the clear visualization for Monkeypox data when compared to the Comparative Visualizations.

- The Cases count is constant at the initial months whereas there's a high increase during the months of July to august.

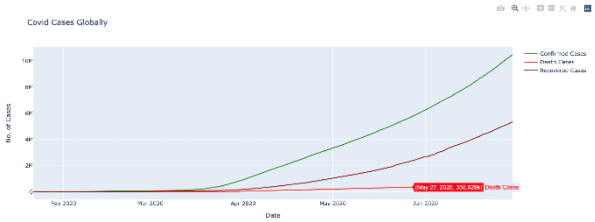


Figure 30. Spread of the Coronavirus Over Time in World

From the above plot,

- There's no huge increase in the initial two months, but there is continuous increase in all the three cases.
- The death Cases shows the constant plot with no much increase, but when you plot for only Death Cases, you can see clear visualization as the count is in the range of Lakhs (Ten Thousands).

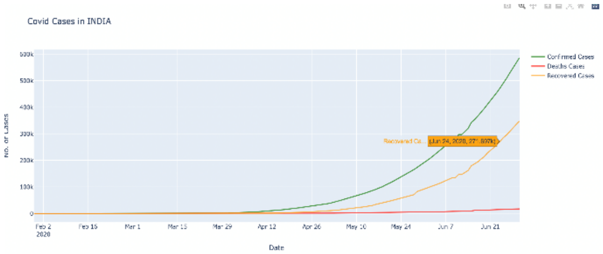


Figure 31. Spread of the Coronavirus Over Time In India

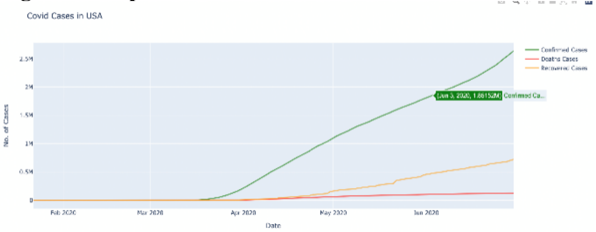


Figure 32. Spread of the Coronavirus Over Time In USA

- In figures 30, 31 and 32 We have analysed corona virus cases in the globe, India, and the USA. The results are shown in graphs, where green denotes confirmed cases, red, death cases, and orange, recovered cases.

- Like the Global Covid Cases Visualization, the two visualizations for India & USA remains same.

CONCLUSION

Consequently, we have seen two distinct analyses here, both separately and in combination. According to the visualization of the combined Covid 19 and monkeypox data, Covid 19 was more significantly affected. We have performed data cleaning, data wrangling and exploratory analysis to each data sets and brought down all the columns into the required form like converting data types, dates, changing column names and filling null values. We have done analysis for most effected, death countries, sates, places, geographically and continentally etc., We performed different visualizations for easy understanding. By comparing the cumulative number of confirmed, death and recovered cases for Covid and monkeypox, we can see that the impact of Covid 19 is significantly higher than those of monkeypox due of its higher effect rate.

FUTURE SCOPE

Future study will concentrate on how Machine Learning Models function, and we will be able to perform an additional Covid-19 analysis on airlines and education how those are affected. We can conduct research on additional pandemics. Analyze the effects of online commerce compared to the Covid-19 or Money Pox. Additionally, by combining data from the COVID 19 and monkeypox pandemics, we may visualize data on other pandemics.

ACKNOWLEDGEMENT

We would like to express our gratitude to Prof. Md Yusuf Sarwar Uddin for teaching us the material that serves as the foundation for comprehending and working on this topic, as well as for aiding us by offering his insightful opinions at various points to better our work. Additionally, we appreciate the advice from the other students in our class and our TA's Vijaya Kumari Yeruva and Rohith Potluri who helped us see things from a different angle.

CONTRIBUTION

Although each team member gave equally to the project, they each paid close attention to their implementations. Mohammad has implemented the project's design and architecture as well as the tools and libraries required. Mohammad contributed by creating the notebooks, which includes data ingestion, data wrangling, data cleaning and preprocessing, and analysis. Consequently, different sources have been combined into a single processed dataset (the target dataset). Rajesh performed EDA on the prepared data and produced analysis. Rajesh has also conducted extensive study to finalize the datasets. Kranthi worked along with Mohammad which involved data ingest to processing while working on creating monkeypox datasets and focusing heavily on the full monkeypox outbreak study. 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.

References:

1.Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CCY, Poon RWS, Tsoi HW, Lo SKF, Chan KH, Poon VKM, Chan WM, Ip JD, Cai JP, Cheng VCC, Chen H, Hui CKM, Yuen KY. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020;395(10223):514–23.

2. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, Wang M. Presumed asymptomatic carrier transmission of COVID-19. JAMA. 2020. https://doi.org/10.1001/ jama.2020.1585.

3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020. https://doi.org/10.1038/ s41586-020-2012-7.

4.Monkeypox: background information. UK Health Security Agency, 2018 (https:// www.gov.uk/guidance/monkeypox# transmission).

5 Mbala PK, Huggins JW, Riu-Rovira T, et al. Maternal and fetal outcomes among pregnant women with human monkeypox infection in the Democratic Republic of Congo. J Infect Dis 2017; 216:824-8. 14. Monkeypox. World Health Organization, May 19, 2022 (https://www.who.int/ newsroom/fact-sheets/detail/ monkeypox). 15. Ogoina D, Iroezindu M, James HI, et al. Clinical course and outcome of human monkeypox in Nigeria. Clin Infect Dis 2020;71(8):e210-e214.

6. Monkeypox outbreak toolbox. World Health Organization, 2022 (https://www.who.int/emergencies/outbreak-toolkit/ disease-outbreak-toolboxes/monkeypox -outbreak-toolbox).

https://journals.sagepub.com/doi/full/10.1177/2347631120983481

https://www.sciencedirect.com/science/article/pii/S0969699721000454

https://www.sciencedirect.com/science/article/pii/S0268401220314869

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9352646/

https://www.kaggle.com/datasets/imdevskp/corona-virus-report

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

https://www.kaggle.com/datasets/deepcontractor/monkeypox-dataset-dailupdated?select=Daily_Country_Wise_Confirmed_Cases.csv

https://www.kaggle.com/code/andrewmvd/monkeypox-cases-analysis/data