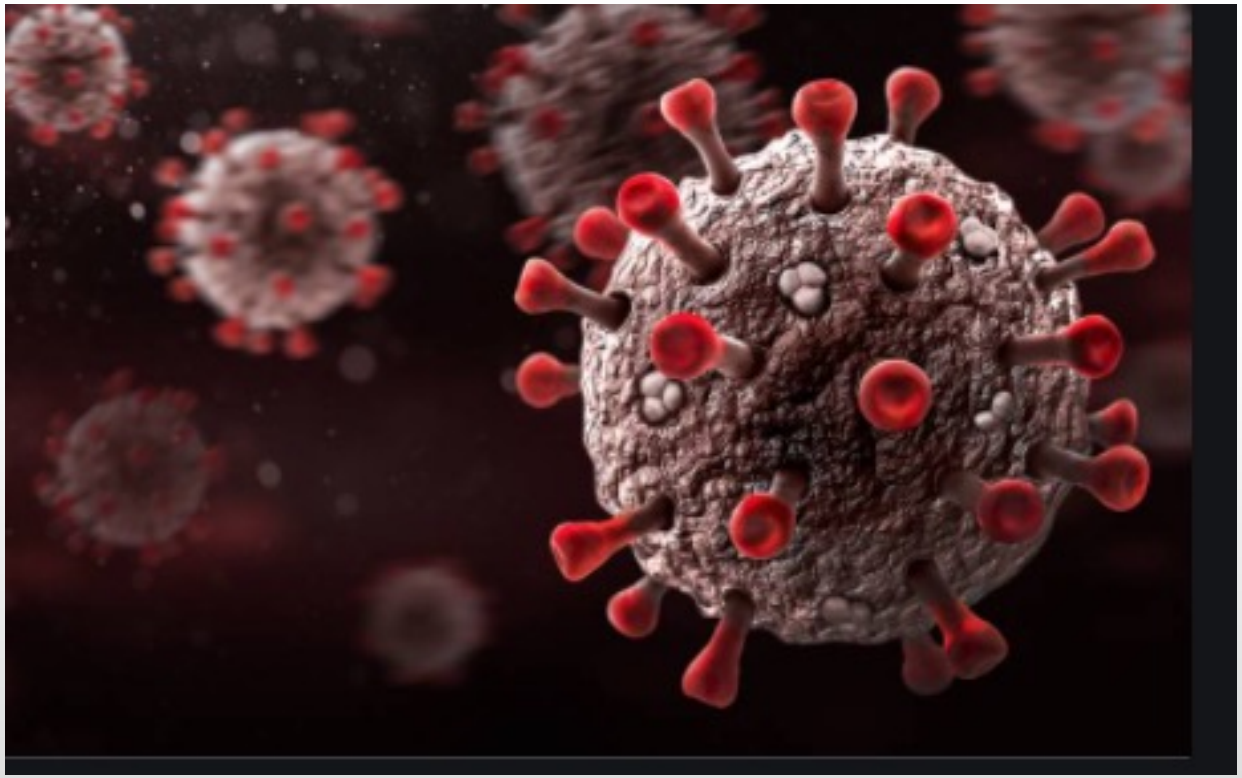
COVID-19 TEXT MINING

Learning

Text Mining

Analysis

Information



- By
- Divya Dongala
- Pragati Gupta

DATA SET

- This data is collected from Kaggle and used for text mining analysis. The data set contain 628319 metadata rows.

- The Data set of covid cases in USA .

- DataSet

- The Data set of covid cases in world .

- DataSet

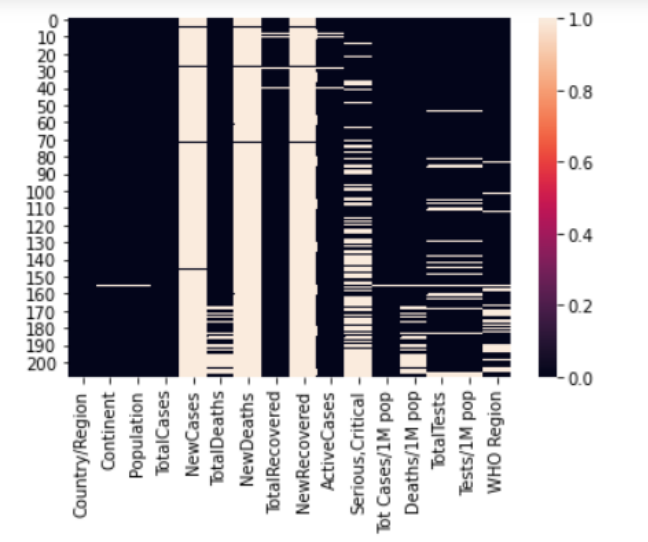

Worldometer dataset


Country wise data

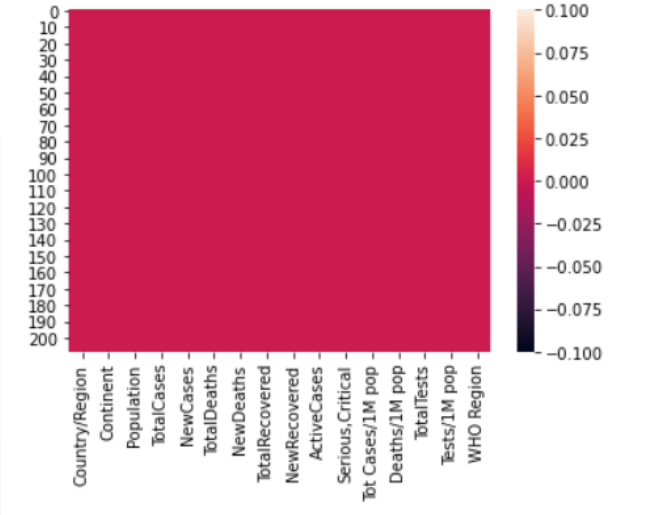

COVID-19 complete data


USA County Wise data

PRE-PROCESSING

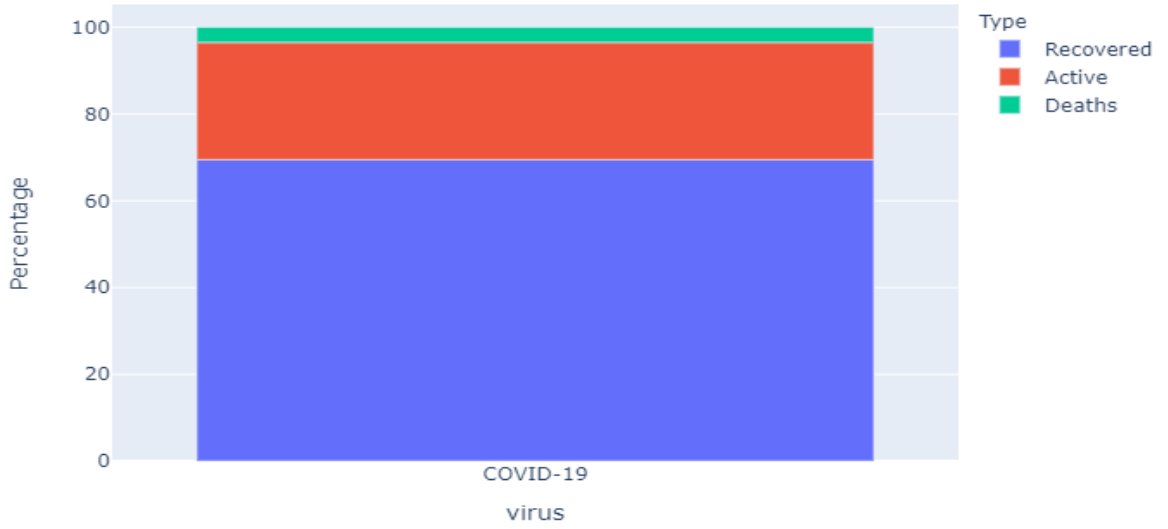


Heat Map before preprocessing



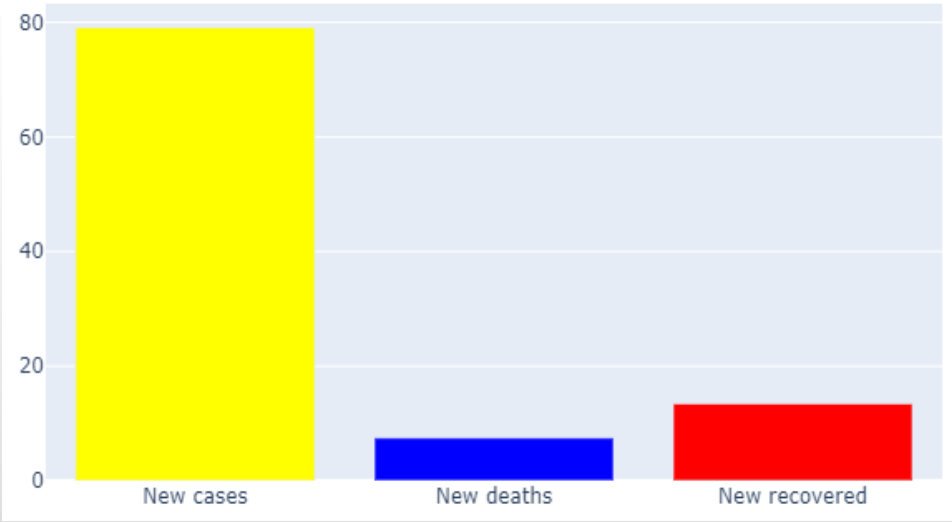
Heat Map after preprocessing

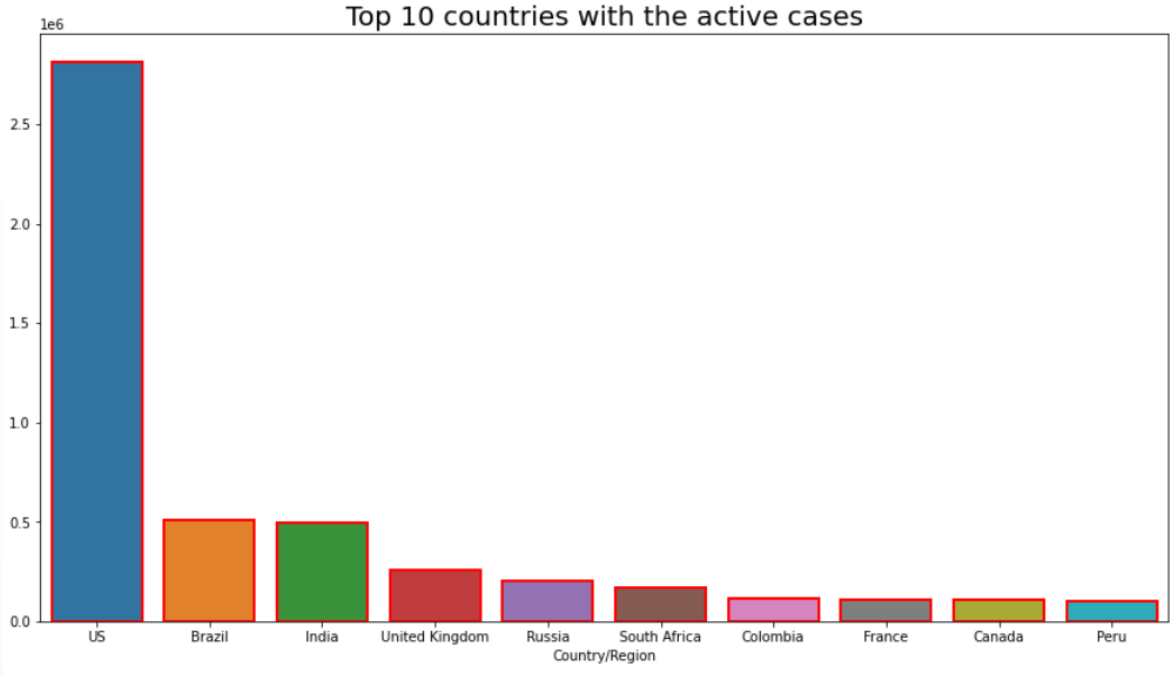
To convert the raw data into the clean dataset in this experiment we removed all the missing values and null values. Pprofile information of the dataset which includes a minimum value, maximum value, mean value and standard deviation of each feature of the dataset. We removed all duplicate rows and the data and finally we start our analysis on the clean data.



We analyzed the worldwide covid-19 data with the number of recovered cases is 69.49pct, the number of active cases 27.01pct and the number of death cases 3.5pct.

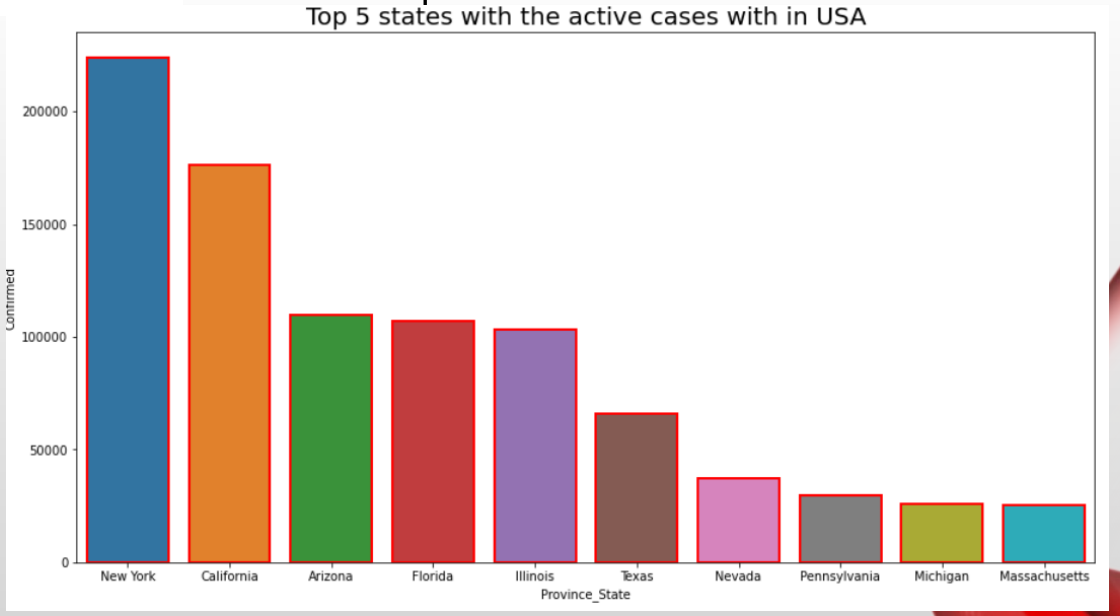
We analyzed the worldwide covid-19 data with the number of new cases is 79.00pct, the number of new deaths 7.00pct and the number of new recovered cases 13.00pct.



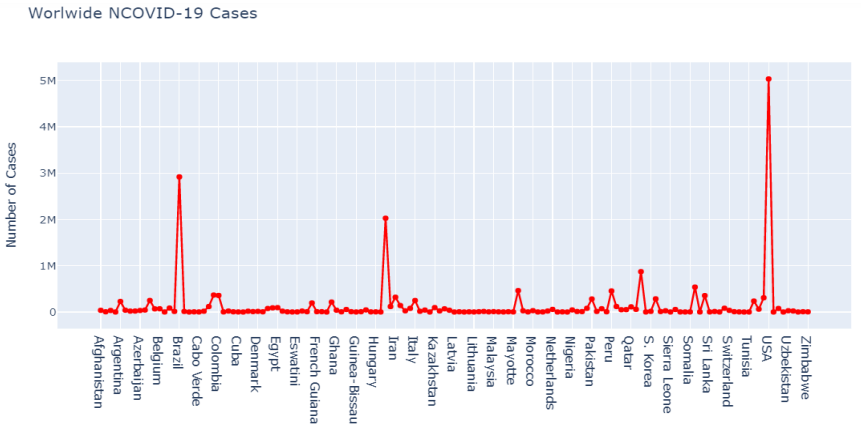


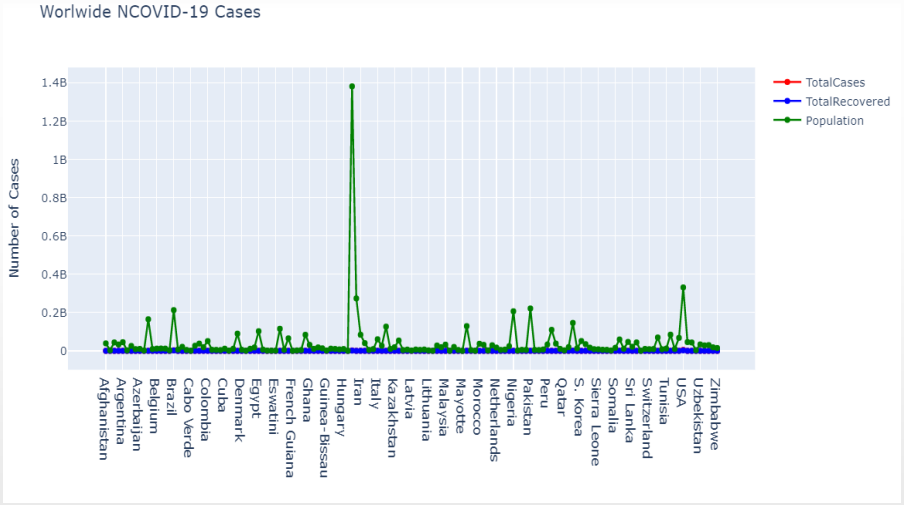
In our world data report, we observed to 10 countries with the highest number of covid-19 cases in US, Brazil, India, United Kingdom, Russia, South Africa, Colombia, France, Canada, and Peru.

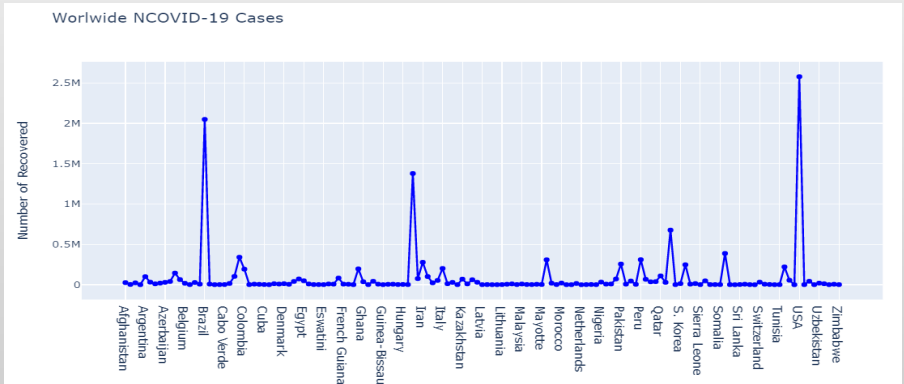
In our US covid data report we observed to 10 states with the highest number of covid-19 cases New York, California, Arizona, Florida, Illinois, Texas, Nevada, Pennsylvania, Michigan, and Massachusetts.

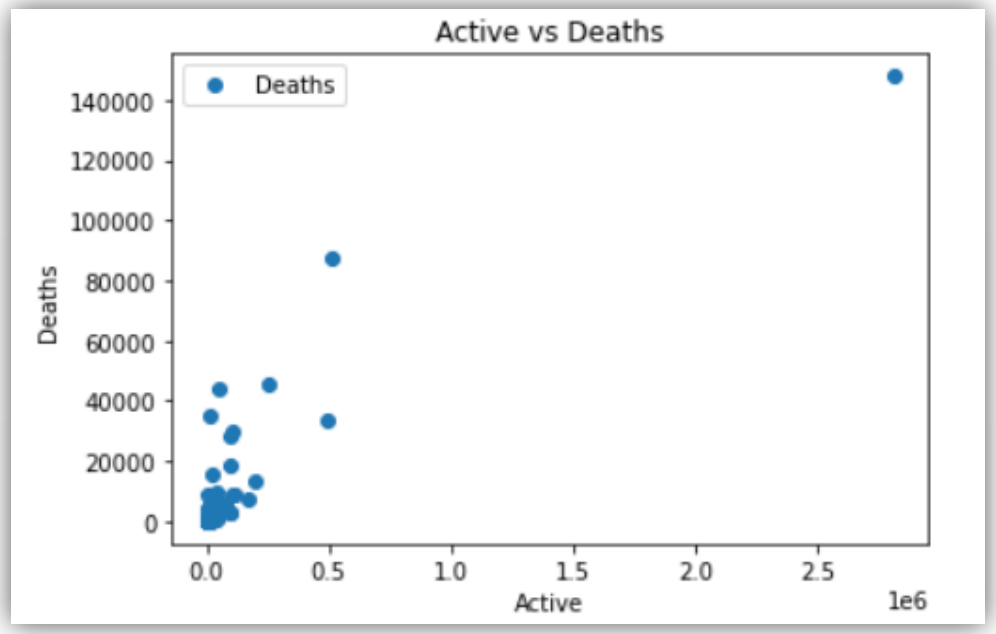


This Analysis provides The total number of cases, the total number of recovered cases over the population in among the countries.

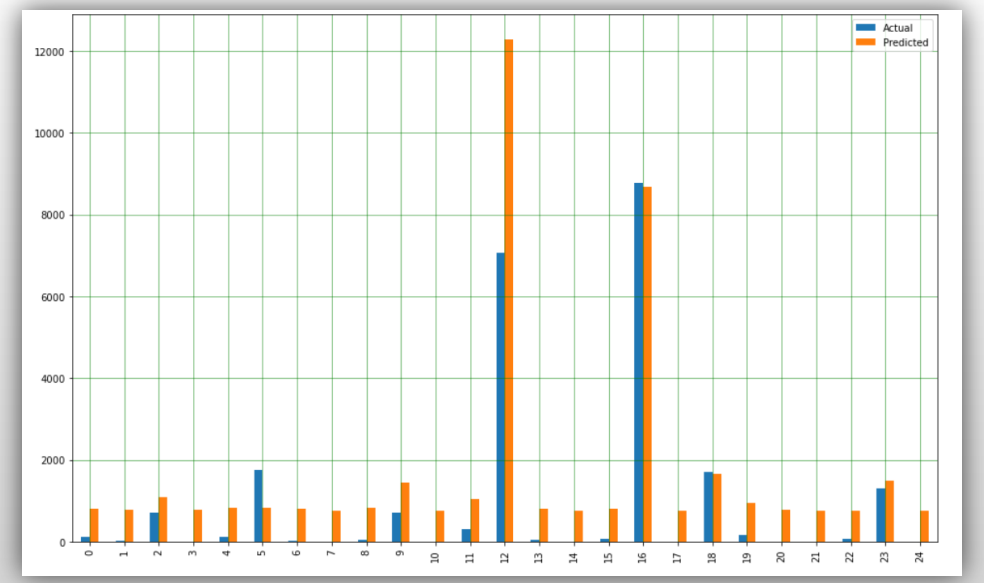








In this paper prediction analysis where the data divided into "attributes" and "labels". Attributes are the independent variables while labels are dependent variables whose values are to be predicted. We split 80% of the data to the training set while 20% of the data to test set. The test size variable is where we specify the proportion of the test set.



To train our algorithm we used the Linear Regression technique which is represented by an equation $Y = a + bX + e$, where a is the intercept, b is the slope of the line and e is the error term.

We will use our test data and see how accurately our algorithm predicts.

OUTCOMES

Q1. Show the number of confirmed deaths and recovered cases in USA.

```
dataset2.groupby('Country/Region').sum()
```

ountry/Region	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop
Afghanistan	39009447.0	36896	0.0	1298.0	0.0	25840.0	0.0	9758.0	31.0	946.0	33.0
Albania	2877470.0	6016	0.0	188.0	0.0	3155.0	0.0	2673.0	23.0	2091.0	65.0
Algeria	43926079.0	33626	0.0	1273.0	0.0	23238.0	0.0	9115.0	57.0	766.0	29.0
Andorra	77278.0	944	0.0	52.0	0.0	828.0	0.0	64.0	1.0	12216.0	673.0
Angola	32956300.0	1483	0.0	64.0	0.0	520.0	0.0	899.0	20.0	45.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...
Vietnam	97425470.0	747	0.0	10.0	0.0	392.0	0.0	345.0	0.0	8.0	0.1
Western Sahara	598682.0	10	0.0	1.0	0.0	8.0	0.0	1.0	0.0	17.0	2.0
Yemen	29886897.0	1768	0.0	508.0	0.0	898.0	0.0	362.0	0.0	59.0	17.0
Zambia	18430129.0	7164	0.0	199.0	0.0	5786.0	0.0	1179.0	0.0	389.0	11.0
Zimbabwe	14883803.0	4339	0.0	84.0	0.0	1264.0	0.0	2991.0	0.0	292.0	6.0

Q2. Show the number of confirmed Deaths and Recovered cases in world

```
dataset2.groupby('Country/Region').sum().head()
```

Country/Region	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop
Afghanistan	39009447.0	36896	0.0	1298.0	0.0	25840.0	0.0	9758.0	31.0	946.0	33.0
Albania	2877470.0	6016	0.0	188.0	0.0	3155.0	0.0	2673.0	23.0	2091.0	65.0
Algeria	43926079.0	33626	0.0	1273.0	0.0	23238.0	0.0	9115.0	57.0	766.0	29.0
Andorra	77278.0	944	0.0	52.0	0.0	828.0	0.0	64.0	1.0	12216.0	673.0
Angola	32956300.0	1483	0.0	64.0	0.0	520.0	0.0	899.0	20.0	45.0	2.0

OUTCOMES

Q3. Remove all the records where the confirmed case is less than 10,000.

```
dataset2=dataset2[~(dataset2.TotalCases<1000)]
```

```
dataset2
```

untry/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop
USA	North America	3.311981e+08	5032179	0.0	162804.0	0.0	2576668.0	0.0	2292707.0	18296.0	15194.0
Brazil	South America	2.127107e+08	2917562	0.0	98644.0	0.0	2047660.0	0.0	771258.0	8318.0	13716.0
India	Asia	1.381345e+09	2025409	0.0	41638.0	0.0	1377384.0	0.0	606387.0	8944.0	1466.0
Russia	Europe	1.459409e+08	871894	0.0	14606.0	0.0	676357.0	0.0	180931.0	2300.0	5974.0
South Africa	Africa	5.938157e+07	538184	0.0	9604.0	0.0	387316.0	0.0	141264.0	539.0	9063.0
...	...	...	...	...	...	...	...	...	...	...	...
Cyprus	Asia	1.208238e+06	1208	0.0	19.0	0.0	856.0	0.0	333.0	0.0	1000.0
Georgia	Asia	3.988368e+06	1206	0.0	17.0	0.0	987.0	0.0	202.0	0.0	302.0
Burkina Faso	Africa	2.095485e+07	1158	0.0	54.0	0.0	961.0	0.0	143.0	0.0	55.0
Niger	Africa	2.428143e+07	1153	0.0	69.0	0.0	1057.0	0.0	27.0	0.0	47.0
Togo	Africa	8.296582e+06	1012	0.0	22.0	0.0	697.0	0.0	293.0	2.0	122.0

Q4. The place where the maximum number of cases are recorded

```
dataset2.groupby('Country/Region').TotalCases.sum().sort_values(ascending=False).head(20)
```

```
Country/Region
USA              5032179
Brazil           2917562
India            2025409
Russia            871894
South Africa      538184
Mexico            462690
Peru              455409
Chile             366671
Colombia          357710
Spain             354530
Iran              320117
UK                308134
```

OUTCOMES

Q5. The place where the minimum number of death cases were recorded.

```
dataset2.groupby('Country/Region').TotalDeaths.sum().sort_values(ascending=True).head(20)
```

```
Country/Region
Uganda              5.0
Rwanda              5.0
Iceland            10.0
Jordan             11.0
Sri Lanka          11.0
Namibia            15.0
Mozambique         15.0
Georgia            17.0
Maldives           19.0
Cyprus             19.0
New Zealand        22.0
```

Q7. Sort the entire data with respect to number of recovered cases in descending order

```
dataset2.sort_values(by=['TotalRecovered'],ascending =False).head(10)
```

	Country/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical
0	USA	North America	3.311981e+08	5032179	0.0	162804.0	0.0	2576668.0	0.0	2292707.0	18296.0
1	Brazil	South America	2.127107e+08	2917562	0.0	98644.0	0.0	2047660.0	0.0	771258.0	8318.0
2	India	Asia	1.381345e+09	2025409	0.0	41638.0	0.0	1377384.0	0.0	606387.0	8944.0
3	Russia	Europe	1.459409e+08	871894	0.0	14606.0	0.0	676357.0	0.0	180931.0	2300.0
4	South Africa	Africa	5.938157e+07	538184	0.0	9604.0	0.0	387316.0	0.0	141264.0	539.0
7	Chile	South America	1.913251e+07	366671	0.0	9889.0	0.0	340168.0	0.0	16614.0	1358.0
6	Peru	South America	3.301632e+07	455409	0.0	20424.0	0.0	310337.0	0.0	124648.0	1426.0
5	Mexico	North America	1.290662e+08	462690	6590.0	50517.0	819.0	308848.0	4140.0	103325.0	3987.0
10	Iran	Asia	8.409762e+07	320117	0.0	17976.0	0.0	277463.0	0.0	24678.0	4156.0
13	Pakistan	Asia	2.212959e+08	281863	0.0	6035.0	0.0	256058.0	0.0	19770.0	809.0

CONCLUSION

Using a text mining approach this study was able to predict the Covid results. There are various data mining techniques used to predict an outbreak. It is hard to deal with this pandemic it requires lots of preparation to manage the unpredicted situation in that way our research helps to predict the future active cases and the number of deaths so that we will the prepared to face the outbreak.

THANK YOU

Team: The Data Scientist

Divya Dongala

Pragati Gupta