International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 04 | April -2020 ISSN: 2582-3930

DATA ANALYSIS ON COVID-19 PANDEMIC

SUJAY J

Co-author: - Professor. **ALAMMA B H**
PG SCHOLAR, DEPT of MCA, DSCE
CA - ASST. PROF, DEPT of MCA, DSCE

Abstract

COVID-19 is a disease which was recently discovered and caused by Coronavirus, this disease was unknown until the outbreak began in china, with this case study we can get to know which all countries have been affected by coronavirus and how many members have been confirmed that they are affected by coronavirus and how many people have lost their lives from getting affected by coronavirus and how many people have recovered from coronavirus as on 26 MARCH 2020. Day to day there are many people getting affected by coronavirus and people have no clear idea about how many people are getting affected throughout the world because of different media showing different data and people are making their own assumptions because of that there are many fake news spread all around on the data of coronavirus. in this paper the data has been collected from trusted websites such as (w h o) world health organisation etc, which gives the clear information about the Corona virus spread.

Introduction

Coronavirus are type of virus which spread the disease called covid-19 it affects the respiratory tract infections which causes cold and slight fever, COVID-19 is the new virus which was found in 2019 as it was never identified in humans in past history. Covid-19 has created a huge fear among the people all around the world.

can the coronavirus spread

The studies have been proven that the virus doesn't travel through air but it spreads though the person who has been infected from it. when the infected person cough or sneeze, the virus is spread in and around the place of that person and whenever another person comes into that zone even, he gets infected

Spreading of Coronavirus

When the person is infected with Covid-19, he doesn't know that he has got infected by coronavirus up to 7-14 days based on his immune power only after that the person falls sick or other symptoms by this time i.e. within those 7 to 14 days the person would meet so many people and spread the virus to them and those people further spread the virus many other people.

Panic on pandemic

There is no vaccine found for this virus but there are many people who have recovered from this case but it takes too much time in incentive care to recover. But there are many people are getting diseased and very less beds in the countries which is why people are afraid. people are getting more afraid to the news of spreading of the virus because there are many fake news or fake

data spreading all over the internet and social media. For this reason, ***this paper is made from collecting the data from reliable resource and plotting the graph for easy understanding of data***

Literature Survey

COVID-19 was first identified in Wuhan, Hubie Province, china in December 2019. This disease rapidly spread over many countries with in a small period of time. there were many people suffering from these disease and many had already lost their lives. the WHO(world health organisation)declared this disease as a Public Health Emergency of International Concern .

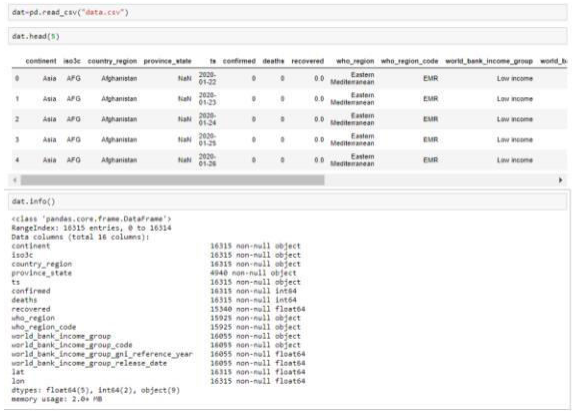
Methodology

For the analysis of COVID-19 pandemic I am going to be using python as a programming language ,python is a very powerful object oriented programming language where we can import packages which are very useful for our analysis such numpy pandas, matplotlib, pyplot, where numpy is used for large and multidimensional array and also for statistical operations, pandas used to perform operation on large datasets as dataframes, matplotlib is used to plot and visualise the graphs in the analysis. Anaconda is used as environment and jupyter Notebook is used as IDE for the data analysis

Analysis

For the analysis of coronavirus, the data has been collected from various reliable resources such as COVID-19 Open Research Dataset (CORD-19) ,COVID-19 Epidemiological Data Repository by Johns Hopkins University Centre for Systems Science & Engineering (JHU CCSE), WHO COVID-19 Data, World Bank Indicators relevant to COVID-19 etc this data has been processed using Anaconda Navigator and Jupyter Notebook. And many packages have been imported such as numpy , plotly, matplotlib , pandas etc

First lets look at the dataset, I have imported the data using read_csv command. Now the data has been inserted we need to look at the data and its contents, that is done by the head()



```
dat=pd.read_csv("data.csv")
dat.head(5)
```

	continent	iso3c	country_region	province_state	ts	confirmed	deaths	recovered	who_region	who_region_code	world_bank_income_group	world_b
0	Asia	AFG	Afghanistan	NaN	2020-01-22	0	0	0.0	Eastern Mediterranean	EMR	Low income	
1	Asia	AFG	Afghanistan	NaN	2020-01-23	0	0	0.0	Eastern Mediterranean	EMR	Low income	
2	Asia	AFG	Afghanistan	NaN	2020-01-24	0	0	0.0	Eastern Mediterranean	EMR	Low income	
3	Asia	AFG	Afghanistan	NaN	2020-01-25	0	0	0.0	Eastern Mediterranean	EMR	Low income	
4	Asia	AFG	Afghanistan	NaN	2020-01-26	0	0	0.0	Eastern Mediterranean	EMR	Low income	

```
dat.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16315 entries, 0 to 16314
Data columns (total 16 columns):
continent                              16315 non-null object
iso3c                                  16315 non-null object
country_region                         16315 non-null object
province_state                         4040 non-null object
ts                                     16315 non-null object
confirmed                              16315 non-null int64
deaths                                 16315 non-null int64
recovered                              15340 non-null float64
who_region                             15925 non-null object
who_region_code                        15925 non-null object
world_bank_income_group                16055 non-null object
world_bank_income_group_code           16055 non-null object
world_bank_income_group_gni_reference_year 16055 non-null float64
world_bank_income_group_release_date   16055 non-null float64
lat                                    16315 non-null float64
lon                                    16315 non-null float64
dtypes: float64(5), int64(2), object(9)
memory usage: 2.0+ MB
```

[fig.1]

In the image [fig.1]we can get to know the columns that are in the data and the first 5 rows present in dataset, well now we have imported the data we need to clean the dataset that is we need keep only the essential data and remove all the other data in this dataset I have kept continent,country_region,province_state,timestamp,confirmed,deaths,recovered,latitude and longitude, and dropped the remaining columns.

```
dat.drop("iso3c",axis=1,inplace=True)
dat.drop("who_region",axis=1,inplace=True)
dat.drop("who_region_code",axis=1,inplace=True)
dat.drop("world_bank_income_group",axis=1,inplace=True)
dat.drop("world_bank_income_group_code",axis=1,inplace=True)
dat.drop("world_bank_income_group_gni_reference_year",axis=1,inplace=True)
dat.drop("world_bank_income_group_release_date",axis=1,inplace=True)
```

```
dat.head(5)
```

	continent	country_region	province_state	ts	confirmed	deaths	recovered	lat	lon
0	Asia	Afghanistan	NaN	2020-01-22	0	0	0.0	33.0	65.0
1	Asia	Afghanistan	NaN	2020-01-23	0	0	0.0	33.0	65.0
2	Asia	Afghanistan	NaN	2020-01-24	0	0	0.0	33.0	65.0
3	Asia	Afghanistan	NaN	2020-01-25	0	0	0.0	33.0	65.0
4	Asia	Afghanistan	NaN	2020-01-26	0	0	0.0	33.0	65.0

[fig.2]

The next process of data cleaning is we need to check for any null values present in

the dataset and how are we handling the null values. Well in this dataset there were few null values in the recovered column which means if there is no value means there are no people have recovered in that country on that day so I filled it with '0'

```
dat.isna().sum().to_frame().sort_values(0).style.background_gradient(cma
```

	0
continent	0
country_region	0
ts	0
confirmed	0
deaths	0
lat	0
lon	0
recovered	975
province_state	11375

[fig.3]

Now the cleaning part is done ,we are going to do the visualization part first thing is we are going to group the data according to time stamp with the columns confirmed ,deaths, recovered so that we can figure out how many people have got confirmed with the virus and how many people have died and how many have recovered on daily basis.

```
In [8]: da=dat.groupby('ts')['confirmed','deaths','recovered'].sum(
        da
```

Out[8]:

ts	confirmed	deaths	recovered
2020-01-22	555	17	28.0
2020-01-23	654	18	30.0
2020-01-24	941	26	36.0
2020-01-25	1434	42	39.0
2020-01-26	2118	56	52.0
...	...	...	...
2020-03-22	336956	14651	97889.0
2020-03-23	378238	16505	98341.0
2020-03-24	418050	18625	107890.0
2020-03-25	467664	21181	113604.0
2020-03-26	529604	23970	121966.0

65 rows × 3 columns

```
In [9]: da[-1:]
```

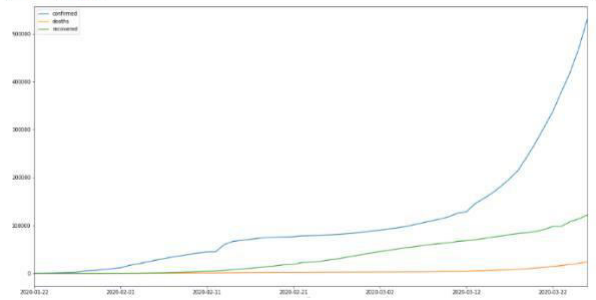
Out[9]:

ts	confirmed	deaths	recovered
2020-03-26	529604	23970	121966.0

[fig.4]

From the image[fig.4] we can get to know that on the day 22 Jan 2020 there were 555 cases confirmed,17 cases where dead and 28 cases where recovered and as on for each day. On 26 march 2020 the total confirmed cases raised to 529604 and around 23970 people were dead and 121966 people were recovered. This data can be very useful to run test cases on patient for finding vaccines. Now we plot the same data in graph for better visuality and understanding

```
dat.groupby('ts')['confirmed','deaths','recovered'].sum().plot(kind = 'line',figsize=(20,10))
plt.savefig('ts.png')
```



[fig.5]

From the image[fig.5] we can easily understand the growth of the disease throughout the time of 22 Jan to 26 march it has constantly growing and rapidly all over the world where as the recovered rate has a growth but it is not enough when looking at confirmed cases and at the same time death rate has not grown or not gone down .

Up to now we have been grouping the data as per the day count from which we got a clear picture of data changes from day to day but now we are going to look through the data from different perspective that is from the view point of each country from which we can get to know which country has suffered the most

.

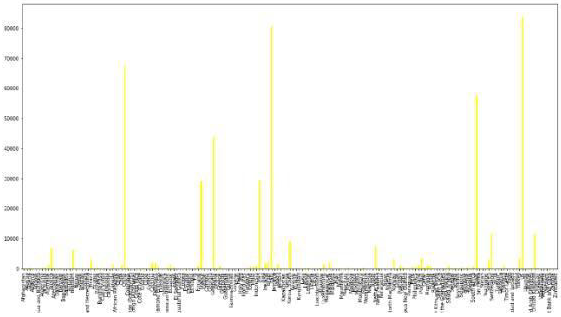
```
xy=dat.groupby('country_region')['confirmed','deaths','recovered'].max()
ab=xy.sort_values("confirmed", axis = 0, ascending = False)
print(ab)
```

```
                             confirmed  deaths  recovered
country_region
US                               83836    1209      681.0
Italy                            80589    8215    10361.0
China                            67801    3169    61201.0
Spain                            57786    4365     7015.0
Germany                          43938     267     5673.0
...                                ...     ...        ...
Saint Kitts and Nevis                2       0        0.0
Saint Vincent and the Grenadines     1       0        0.0
Libya                                1       0        0.0
Timor-Leste                          1       0        0.0
Papua New Guinea                     1       0        0.0

[175 rows x 3 columns]
```

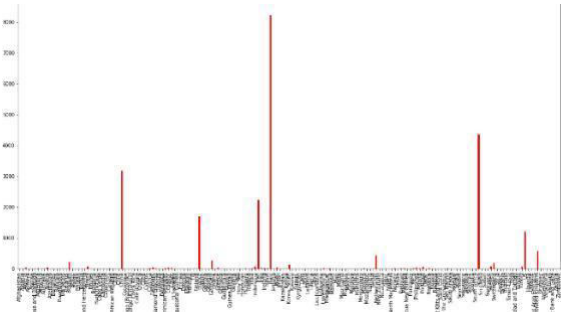
[fig.6]

According to the stats from the image [fig.6]we can get to know that currently **U**nited **S**tates has more Confirmed cases as on 26 march but whereas **I**taly has suffered more from this coronavirus because they faced more deaths than United States but this not the end as the stats are growing for day to day in every country which is making the condition more worse and worse. We can understand these stats easily by plotting a graph for each. let's look at those one by one.



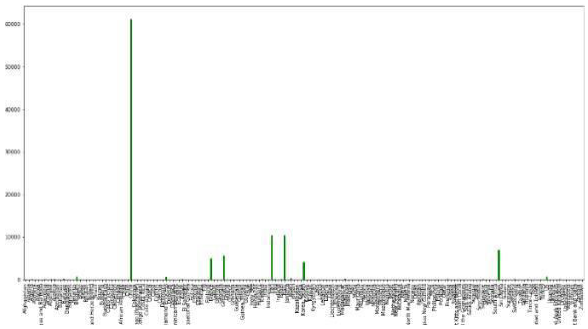
[fig.7]

Well in bar graph[fig.7] we can know that there are many confirmed cases in United States followed by Italy and china that is, in these countries the virus has spread drastically and these countries are trying very hard to control the situation



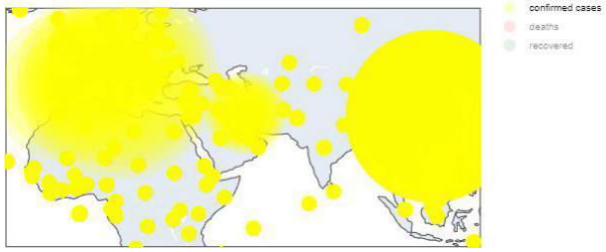
[fig.8]

In this graph[fig.8] we are looking at the Death rate at every country which shows Italy has nearly 8k deaths followed by Spain with more than 4k deaths and then china with 3k death toll in their country.

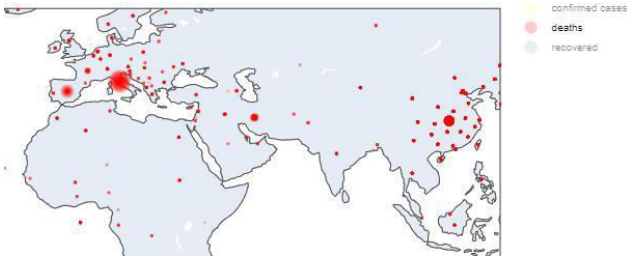


[fig.9]

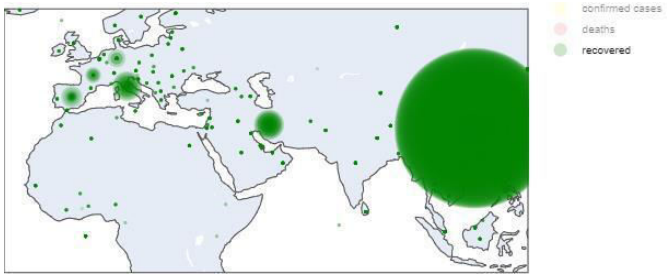
I guess this chart[fig.9] shows how humans are fighting against the coronavirus, these charts represent the number of people that have recovered from coronavirus in each country. In this chart we can see clearly that china has recovered many people and the happy part is that Italy is the second place for recovering the patients which has the highest death rate among all other countries. Now that all the countries have trying to recover their patients all that they have to do is to contain the disease and stop further spreading of it.



[fig.10]



[fig.11]



[fig.12]

The images (fig.10,11and 12) are the best way to display the geographical spread of disease, how vast the virus has spread all around the world. the image (fig.10) the yellow scatter plots are the number of confirmed cases of coronavirus all around the world from which we can say US, China and Italy have the greatest number of cases. In the image (fig.11) the red scatter plots are the number of people dead because of coronavirus. In the image (fig.12) there are green scatter plots on the map which indicates number of people who are recovered from the coronavirus.

Conclusion

The corona virus have infected many more people than expected many country couldn't control the outbreak of the virus even though many precautions were taken by planning ahead to be approx. around more than 5 lakh people where confirmed with coronavirus and nearly 24 thousand people have lost their lives for this coronavirus. United States, Italy, China and Spain are the counties which are most infected and suffered from the coronavirus. Few of you mite feel it is difficult to read the charts/images so I have given my github link where you can get all the data, code and the images regarding this paper and get additional information or you can work from them. The following link is

https://github.com/Sujay-j/Coronavirus26-3.git .

References

1. WHO COVID-19 Data,
2. COVID-19 Open Research Dataset (CORD-19) ,
3. COVID-19 Epidemiological Data Repository by Johns Hopkins University Centre for Systems Science & Engineering (JHU CCSE),
4. World Bank Indicators relevant to COVID-19