1. INTRODUCTION:

The whole world is fighting against the **CoVID-19** pandemic, and in these tough times data analytics is playing a great role in better understanding the situation as to how to tackle it and is even promoting in finding a cure.

So, I also thought of using the knowledge of data analytics that I have gained so far in making a project that focuses on the CoVID scenario for a state in ***India*** in which I live i.e. '**JHARKHAND**'.

This project aims to find out the **safest district** to live among all the districts in the state, if you are in any case coming to **Jharkhand**, based on a variety of features and demographics.

2. DATA USED:

In order to perform the analysis according to the proposed problem,I decided to collect the following demographics :

- The list of the **districts in the state**
- Current CoVID-19 scenario of the state, i.e.:
    - The **no. of active cases**
    - The **no. of deaths**
    - The **no. of recovered cases**
    - The **total cases** that have happened since the start
- The latitude and longitudes of the districts in order to visualise them.
- And, finally the no. of **medical facilities available** in the state (*note: I couldn't find the exact list of the hospitals, so I counted the no. of hospital management societies in each district - though consider the no. of hospitals to be lesser than those no.s*)

So, after web scraping, geocoding and manual counting from some pdfs I got a dataframe which can be further preprocessed according to the requirements and I had the following as the head of the dataframe :

s_%of_total_cases_per_district	active_%of_total_cases_per_district	no_of_hospital_managing_societies_per_active_cases	death_per_recovery	lat	lng
1	43	62.790698	1.818182	23.779985	85.969322
0	39	20.512821	0.000000	24.206200	84.871300
0	50	28.000000	0.000000	24.491533	86.691254
2	59	74.576271	5.263158	23.794043	86.426170
0	30	106.666667	0.000000	24.265244	87.250191

	district	total_cases_%of_total_cases_in_jharkhand	recovered_%of_total_cases_per_district	deaths_%of_total_cases_per_district	active_%of_total_cases_per_di
0	Bokaro	1	55	1	
1	Chatra	3	60	0	
2	Deoghar	1	49	0	
3	Dhanbad	6	38	2	
4	Dumka	0	69	0	

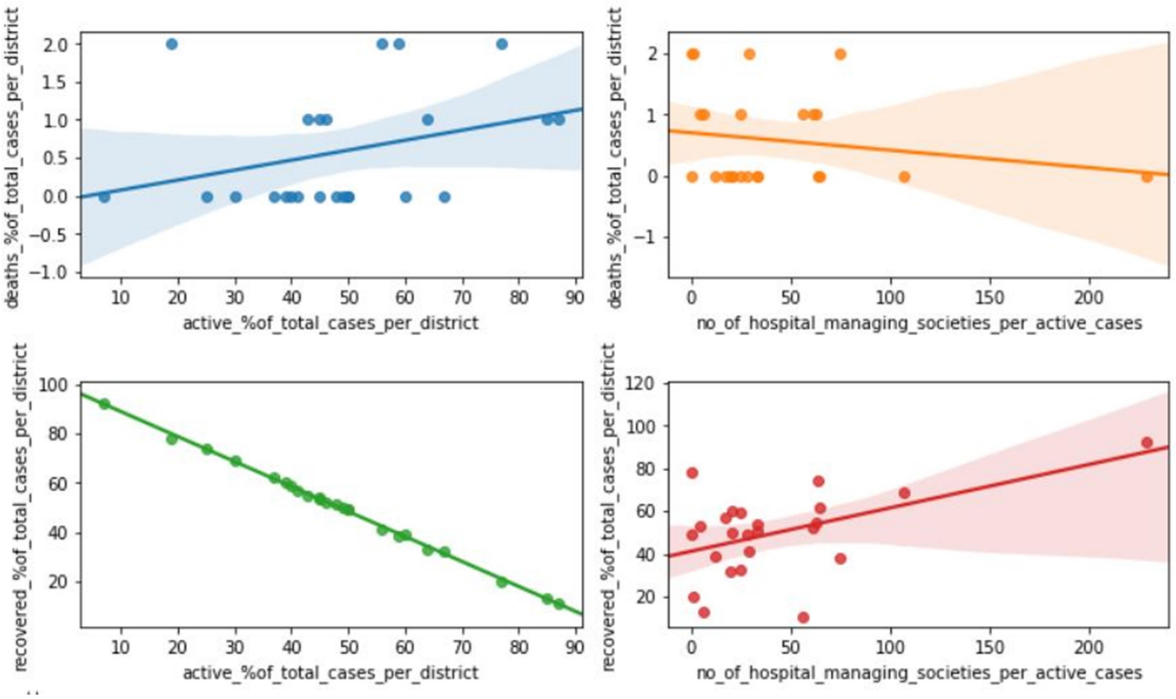
I converted the total cases, recovered cases, deaths and active cases features to %'s relative to the total cases in Jharkhand ,and as per each district for some features. I also introduced 'death_per_recovery' feature which gives us a quantitative idea of the no. of deaths for each newly recovered patient.

Here are various sources from which I gathered the required features. They include:

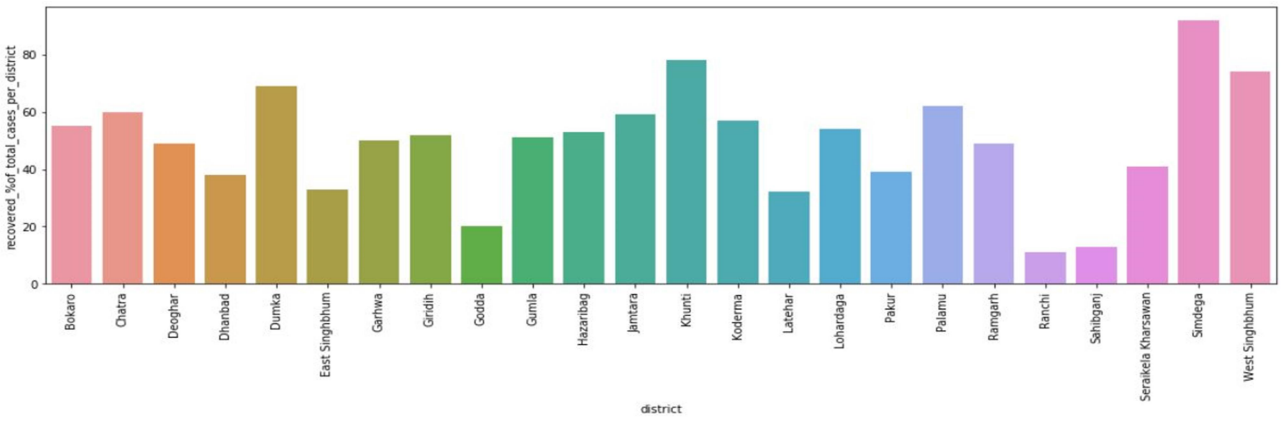
- **Foursquare API (***couldn't use it as the server is down in my region for several days***)**
- **https://covidindia.org/jharkhand/** *(for the no. of active case, deaths, total cases, etc...)*
- **https://www.nhm.gov.in/images/pdf/communitisation/rogi-kalyan-samiti/jharkhand_rks.pdf** *(to get the no. of hospital managing societies in each district)*
- **geocoding from geocodefarm server.** *(to get the latitude and longitude for each district)*

3. METHODOLOGY:

On performing initial exploratory data analysis it was found that the **recovered %** showed a **decreasing trend** with the **% increase of active cases** and an **increasing trend** with **the no. of increasing hospital management societies** for the respective districts, which is an obvious inference.Also, the **death %** seemed to show an **opposite trend** to that of the recovered % for the respective features as expected. The following graph gives us an idea about the trends.



As far as the distribution of the recovered % over the various districts was concerned, it was varied significantly over the districts may be due to the **population variation** or due to the **variation in distribution of the features** mentioned in the above plot. The plot below shows us the relative % distribution over the districts of Jharkhand.

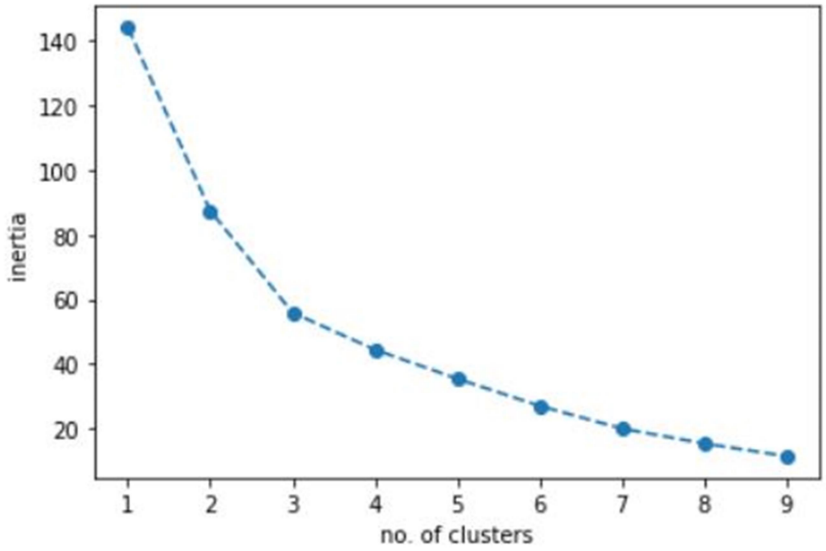


The above plot indicated the districts- **Simdega, West Singhbhum, Khunti** and **Dumka** to have high recovered %, they may turn out to be one of the safe zones in the district.

And finally, the machine learning model that I used to determine the safe zones in the state was a, **K-Means Clustering** model as I wanted to cluster the different districts spread throughout the states and to know what features are responsible for distinctly separating one cluster from the other.

4. RESULTS:

In order to find the ideal cluster for my model I decided to use the **elbow method** by plotting the respective inertias for each clustering model (*by variating the no. of clusters*) and according to the graph below I found the elbow point to be at no. of clusters = 3. (*The elbow point may not be promising due to less amount of data).*



After passing scaled and selective features through the clustering model , assigning the labels to the clusters and calculating mean over the labelled dataframe I got the following:

labels	total_cases_%of_total_cases_in_jharkhand	recovered_%of_total_cases_per_district	deaths_%of_total_cases_per_district	active_%of_total_cases_per_district
0	46.500	83.0000	0.0000	16.000000
1	19.500	26.0000	1.5000	71.333333
2	4.875	54.3125	0.3125	44.312500

ses_per_district	deaths_%of_total_cases_per_district	active_%of_total_cases_per_district	no_of_hospital_managing_societies_per_active_cases	death_per_recovery
83.0000	0.0000	16.000000	146.285714	0.000000
26.0000	1.5000	71.333333	31.941766	6.659121
54.3125	0.3125	44.312500	31.772919	0.512010

5. DISCUSSION / CONCLUSION:

On observing the labelled dataframe I was able to infer the following:

- **Cluster 0** indicated **high recovery** and **low (death/recovery)** so I concluded it to be the **safest** of the lot.
- **Cluster 1** indicated **low recovery, high death%** and **high (death/recovery)** so I concluded it to be the **most unsafe** of the lot.
- **Cluster 2** though indicated **low total case% relative to jharkhand** and **low death%** but it also showed **low (hospital societies/case)** and also **greater active% of cases per district** so I concluded it to be **considerable to live** after the districts in cluster 0.

Plotting the districts after color coding them according to the clusters ,using Folium:

My model labelled only two districts as green zones which include- **West Singhbhum** and **Simdega.** Which, as a matter of fact we thought of being as one of the safe zones during the early exploratory analysis as they had the great recovered % values.

So, this marks the end of my report , however I would like to make some final remarks:

- There was not a lot of data to work with hence the results can be enhanced more.
- Also due to the less no. of districts and population variation among them, there might be skewness to some extent ,to the results.
- Also, the CoVID scenario will change along with the data used at the time of making this report ,hence the conclusions here may differ from a future perspective of the situation.
- Finally , being a resident of the **East Singhbhum (*red zone*)** district I might be considering to relocate. (not to be taken seriously ;P)

THANK YOU !