

DATA MANAGEMENT PROJECT REPORT

(Project Semester: August-December 2021)



L OVELY
P ROFESSIONAL
U NIVERSITY

COVID DATA ANALYSIS

Submitted by

ARIJIT BERA

11901674

Programme and Section: B.Tech (CSE), KM002

Course Code: INT217

Under the Guidance of

Maneet Kaur - 15709

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that ARIJIT BERA bearing Registration no. 11901674 has completed INT217 project titled, “**COVID DATA ANALYSIS**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Maneet Kaur

**School of Computer Science & Engineering
Lovely Professional University Phagwara,
Punjab.**

Date: 30/12/2021

DECLARATION

I, ARIJIT BERA, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 30/12/2021

Registration No.: 11901674

ARIJIT BERA

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Mrs. Maneet Kaur who gave me the golden opportunity to do this wonderful project of analysis of the data of a superstore namely “COVID DATA ANALYSIS” which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them. Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

TABLE OF CONTENT

1. Introduction	6
2. Scope of the Analysis	8
3. Source of dataset	9
4. ETL process	10
5. Analysis on dataset (for each analysis)	15-19
i. Introduction	
ii. General Description	
iii. Specific Requirements, functions and formulas	
iv. Analysis results	
v. Visualization	
6. List of Analysis with results	20
7. References	23
8. Bibliography	24

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China, and has resulted in an ongoing pandemic. The first case may be traced back to 17 November 2019. As of 8 June 2020, more than 6.98 million cases have been reported across 188 countries and territories, resulting in more than 401,000 deaths. More than 3.13 million people have recovered.

The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms.

The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms.

PANDEMIC :

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was first identified in Wuhan, China, in December 2019. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January, and a pandemic on 11 March.

A global coordinated effort is needed to stop the further spread of the virus. A pandemic is defined as “occurring over a wide geographic area and affecting an exceptionally high proportion of the population.” The last pandemic reported in the world was the H1N1 flu pandemic in 2009.

Coronaviruses are important human and animal pathogens. At the end of 2019, a novel coronavirus was identified as the cause of a cluster of pneumonia cases in Wuhan, a city in the Hubei Province of

China. It rapidly spread, resulting in an epidemic throughout China, followed by an increasing number of cases in other countries throughout the world. On 30th January 2020 India recorded its first COVID-19 case in state of Kerala. It was a student who had travel history to china. And till the start of June India has over 200 thousand confirmed cases.

Problem Statement:

In this project we dived deep into ‘What does data say about Covid-19 situation in India?’. And with available data we came up with some observations and conclusions.

This analysis mainly focuses on:

- ✓ What is the current COVID-19 situation in India?
- ✓ State-wise comparison.
- ✓ What could be the reasons behind cases clusters found in India.
- ✓ Is lockdown in India successful or not?

SCOPE OF ANALYSIS

This project on Covid data Statistics of India provides the overall Statistics details of the matches of IPL and teams progress in various aspects from the year 2020 - 2021.

Objectives of this project:

- To good hand on excel.
- To use different feature and get friendly with the excel.
- To learn the ETL process in the Tableau Prep.
- How to link one sheet to another and traverse between different sheets.
- How to use pivot table and pivot chart.
- Learn to make dashboard in excel.
- To make different type of graphs in excel.
- To learn how to fetch data from other source to excel in different formats.

Aim of this project is to answer the above objectives in the form of visualization by creating a dashboard to convey the answers effectively and efficiently.

SOURCE OF DATASET

The data is being taken from the Kaggle. Kaggle is an AirBnB for Data Scientists – this is where they spend their nights and weekends. It’s a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars from the likes of Hal Varian (Chief Economist at Google), Max Levchin (Paypal), Index and Khosla Ventures and then ultimately been acquired by the Google in March of 2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only 94 Kaggle Grandmasters in the world to this date.

ETL PROCESS

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Precisely, ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL stands for Extract, Transform and Load.

Before ETL, the dataset looked like this. **This data is taken from Kaggle.**

<

AYUSHHospitals.csv (1.69 kB)

Download

Fullscreen

Detail

Compact

Column

9 of 9 columns

About this file

Number of beds and Hospitals in Ayush Hospitals

A Srl no.		A State / UT		A Number of Hospit...		A		A
[null]	7%	[null]	12%	0	12%	0	79%	2
-1	2%	-2	2%	4	9%	35	5%	5
Other (39)	91%	Other (37)	86%	Other (34)	79%	Other (7)	16%	0
				Govt .		Local Body		Others
-1		-2		-3		-4		-5
1		Andhra Pradesh		8		0		0
2		Arunachal Pradesh		11		0		1
3		Assam		4		0		0
4		Bihar		8		0		0
5		Chhattisgarh		7		0		8

< Employees State Insurance Corporation .csv (779 B)



Detail Compact Column

3 of 3 columns ▾

About this file

Number of beds and Hospitals maintained by Employees state Insurance corporation

▲ Employees State I...	▲	▲	▲
[null] 6%	[null] 6%	0 18%	
S. No. 3%	States /UTs 3%	1 15%	
Other (30) 91%	Other (30) 91%	Other (22) 67%	
S. No.	States /UTs	Total No. of Hospital	
1	Andhra Pradesh	5	
2	Assam	1	
3	Bihar	3	
4	Chandigarh [Adm.]	1	
5	Chhattisgarh	0	
6	Delhi	4	

< Hospitals and Beds maintained by Ministry of Defence.csv (703 B)





Detail Compact Column

4 of 4 columns ▾

About this file

Number of beds and Hospitals maintained by Ministry of defence

▲ S. No.	▲ Name of State	# No. of Hospitals	# No. of beds
31 unique values	31 unique values		
1	Assam	8	2357
2	Andhra Pradesh	1	306
3	Andaman & Nicobar Islands	1	107
4	Arunachal Pradesh	1	198
5	Bihar	2	348
6	Delhi	2	1993

< Hospitals and beds maintained by Railways.csv (937 B)



Detail Compact Column

3 of 3 columns ▾

About this file

Number of beds and Hospitals maintained by Railways

▲ Number of Hospit... ▾	▲ ▾	▲ ▾
27 unique values	27 unique values	1 9 Other (15) 33% 11% 56%
S.No.	Zone / PU	Total No. of Hospitals
1	Central Railway	11
2	Eastern Railway	8
3	East central Railway	9
4	East coast Railway	4
5	Northern Railway	9
6	North Central Railway	5
7	North East Railway	6

< Hospitals_and_Beds_statewise.csv (1.31 kB)







Detail Compact Column

6 of 6 columns ▾

About this file

Statewise number of beds and Hospitals (Overall)

▲ ▾	# PHC ▾	# CHC ▾	# SDH ▾	# DH ▾	# ▾
37 unique values	 4 29.9k	 2 5.57k	 1 1.25k	 1 1k	8
Andaman & Nicobar Islands	27	4		3	34
Andhra Pradesh	1417	198	31	20	16
Arunachal Pradesh	122	62		15	19
Assam	1007	166	14	33	12
Bihar	2007	63	33	43	21
Chandigarh	40	2	1	4	47
Chhattisgarh	813	166	12	32	16

< Number of Government Hospitals and Beds in Rural and Urban Areas .csv (1... ⬇ ⌕

Detail Compact Column

6 of 6 columns ▾

About this file

Number of beds and Hospitals in Government hospitals

▲ States/UTs	▲ Rural hospitals	▲	▲ Urban hospitals	▲	▲
38 unique values	56 0 Other (34) 5% 5% 89%	0 Beds Other (35) 5% 3% 92%	50 0 Other (34) 5% 5% 89%	0 Beds Other (35) 5% 3% 92%	31 31 0
	No .	Beds	No .	Beds	
Andhra Pradesh	193	6480	65	16658	0
Arunachal Pradesh*	208	2136	10	268	31
Assam	1176	10944	50	6198	31
Bihar	930	6083	103	5936	31
Chhattisgarh	169	5070	45	4342	0
Goa*	17	1405	25	1608	31
Gujarat	364	11715	122	20565	31
Haryana*	609	6690	59	4550	31

Through the process of ETL, we are going to clean the dataset and bring all the entities to their proper data format.

Step 1: Removing the blank cells from the dataset.

For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will remove any rows with blank cells.

Step 2: Removing columns which are not properly defined or not crucial to our analysis.

For this we will columns which are redundant like the column with just the index numbers. For this we will select that particular column and then go to delete option in the home tag and then select Delete Columns from the drop-down menu.

Step 3: Giving proper and appropriate column names.

The dataset does not have proper columns so our next step would be to give proper column names to the columns wherever required.

Step 4: Excluding the NULL values from the data.

We'll be using Tableau prep for this work as it'll make the work simple and faster because we might not know how many null values could be there in this huge data set. Tableau helps us doing one step cleaning with ease.

Step 5: Improvising Proper Data Formatting

Without proper Data Formatting, proper analysis will not take place. So, we will bring down certain columns to their proper format. For example, the dates should be in the date format and price and sales should be in currency format for better results.

Step 6: Removing Duplicate Values

It might be possible that our data may be containing duplicate values which may hinder in precise analysis. So, our last task in ETL will be removing duplicate values and making our data perfect for analysis.

ANALYSIS OF DATASET

1. Total numbers of hospitals in India :

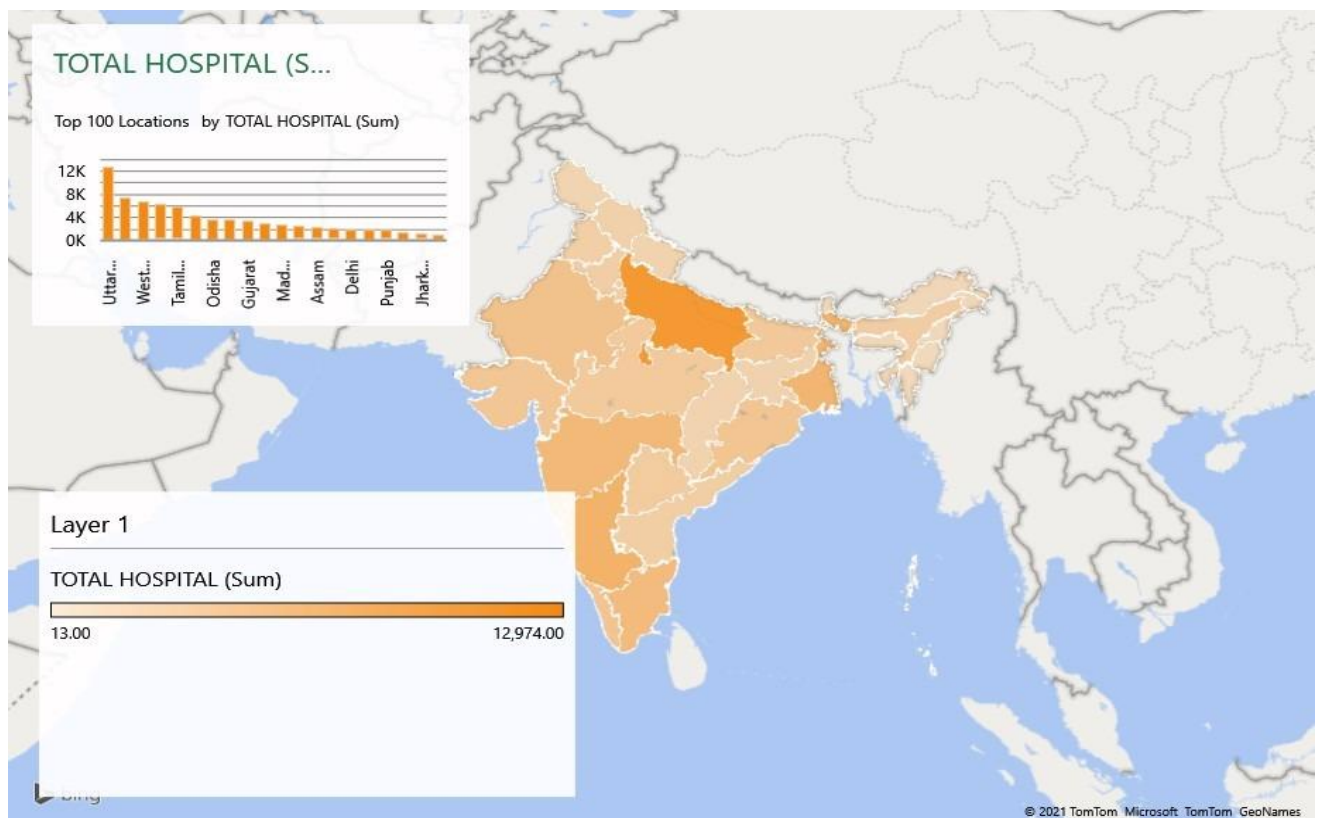
Description:

In this objective we will be finding the number of hospitals that are present in India by taking the data of states.

Specific function and requirements

We have to create a pivot table to determine the difference in goals and then visualize it on graph.

Results:



Visualization:

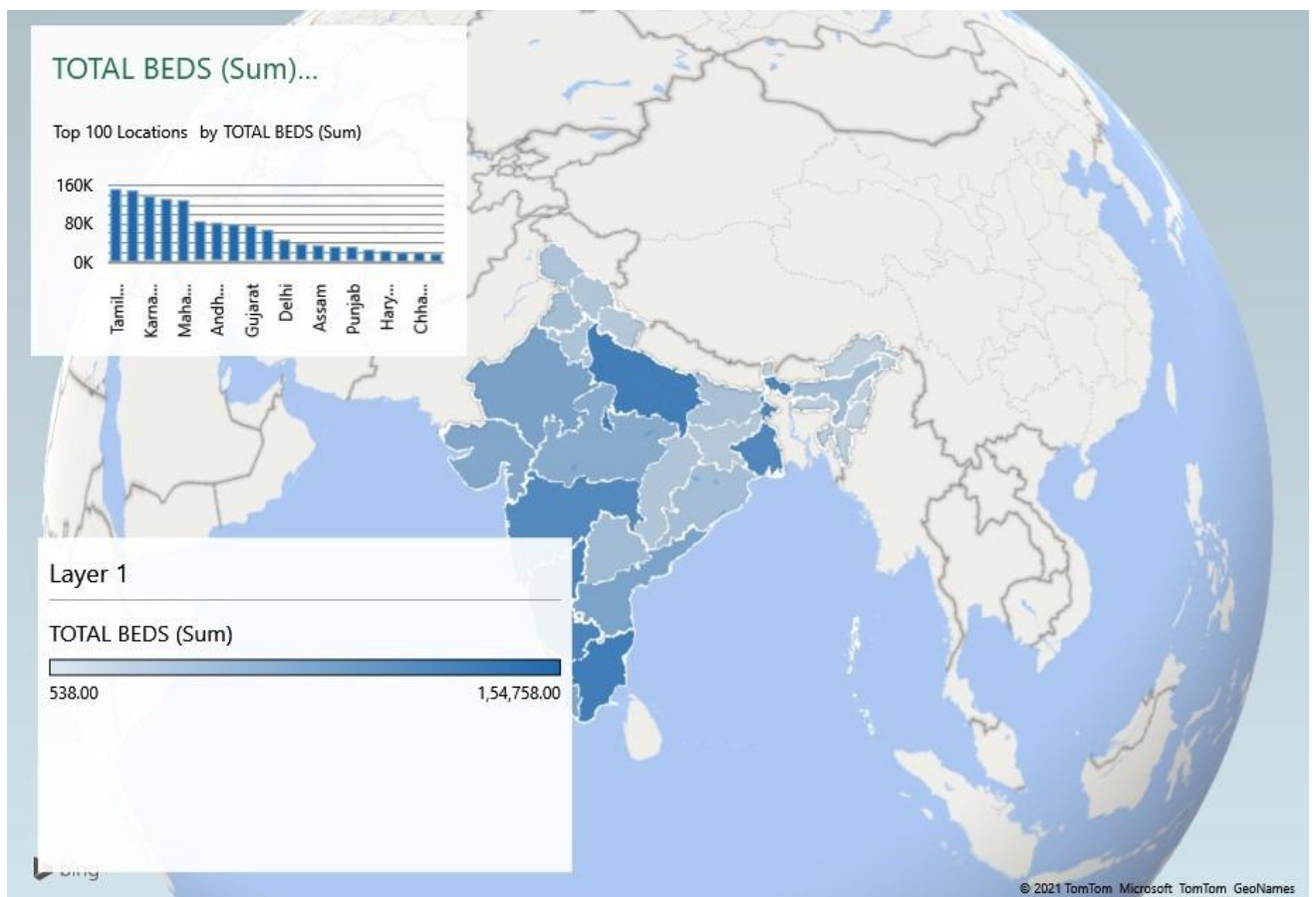
The data is been shown in this manner.

2. Number of bed that has been in the hospital:

Desc:

In this we will be checking out the number of bed that has been in the different hospital across the country.

Results:



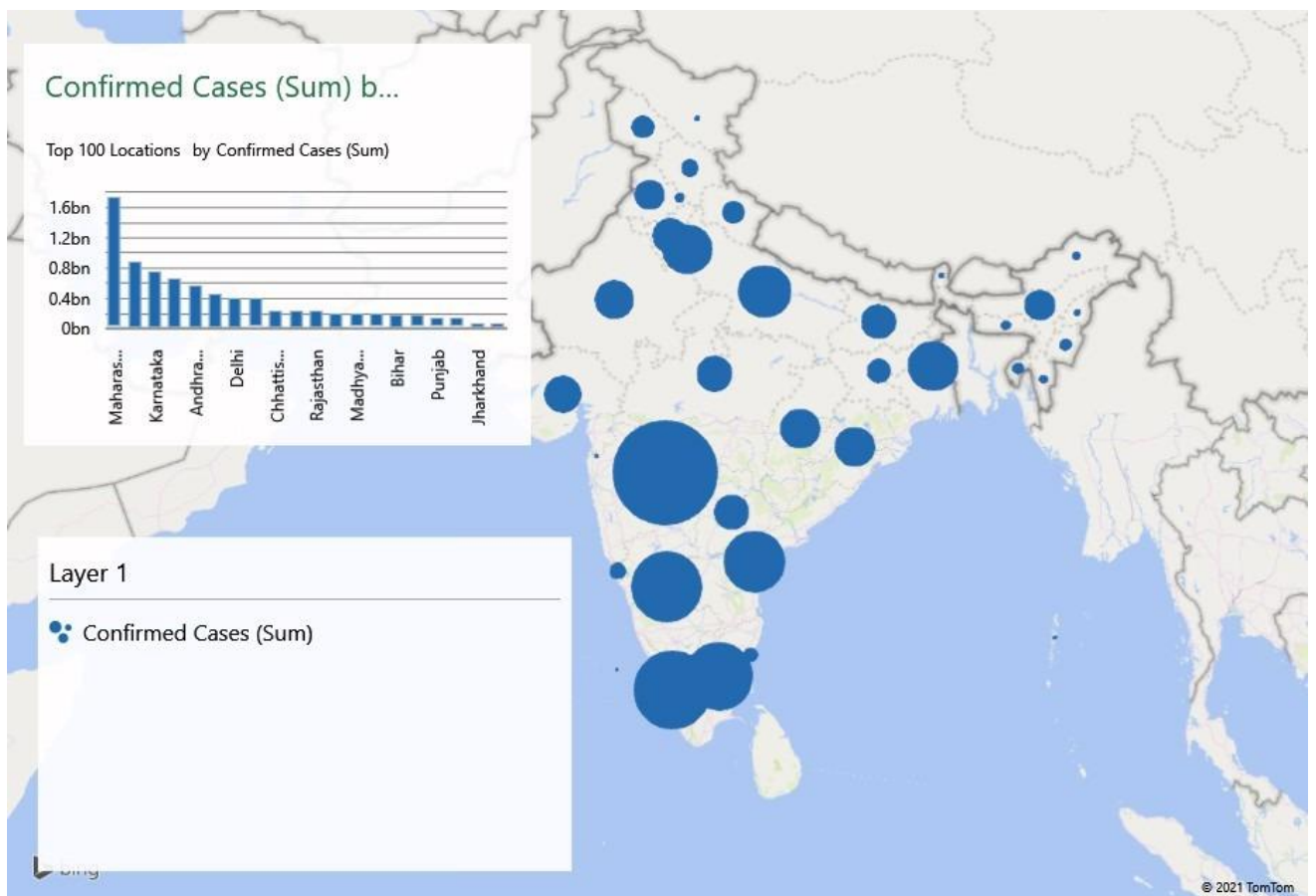
3. Total number of confirmed cases

Desc:

This will be showing us the total number of confirmed cases in India during the past years.

Results:

Here is a little chuck of data



4. Total number of death by region:

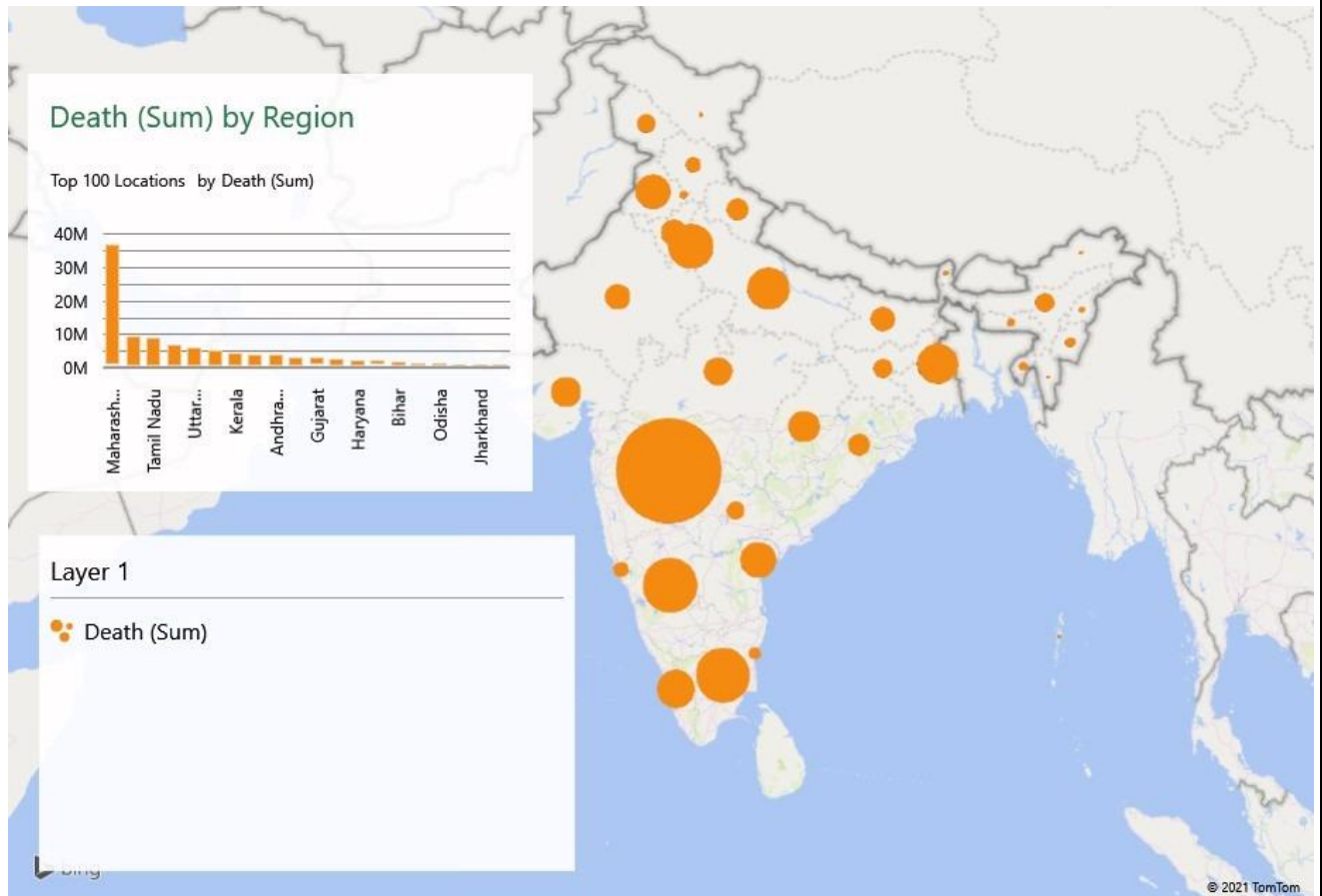
Description:

This will be showing us the death toll that has been reached during the past few years.

Specific function and requirements:

We have to create a pivot table.

Results:

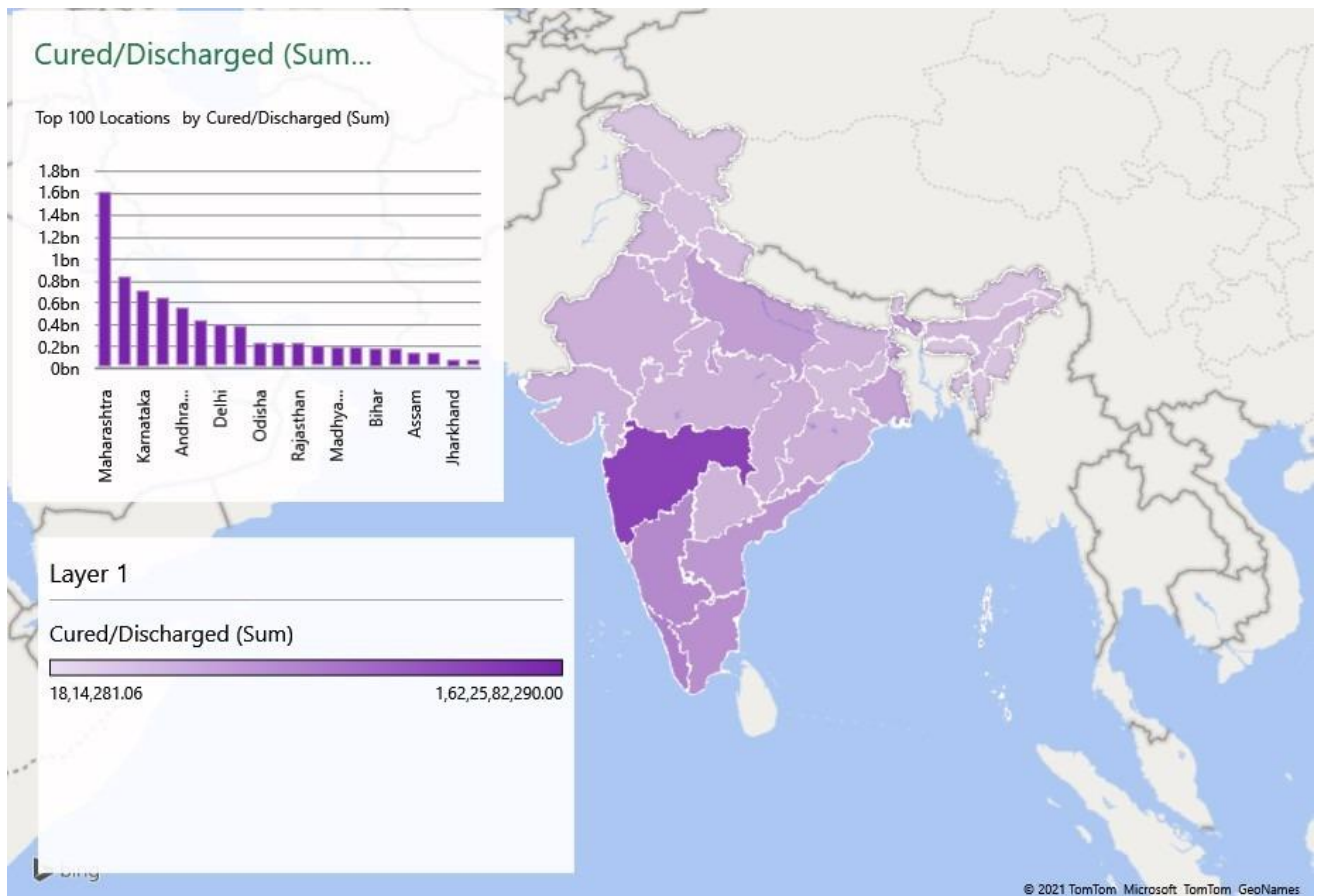


5. Total number of discharged:

Desc:

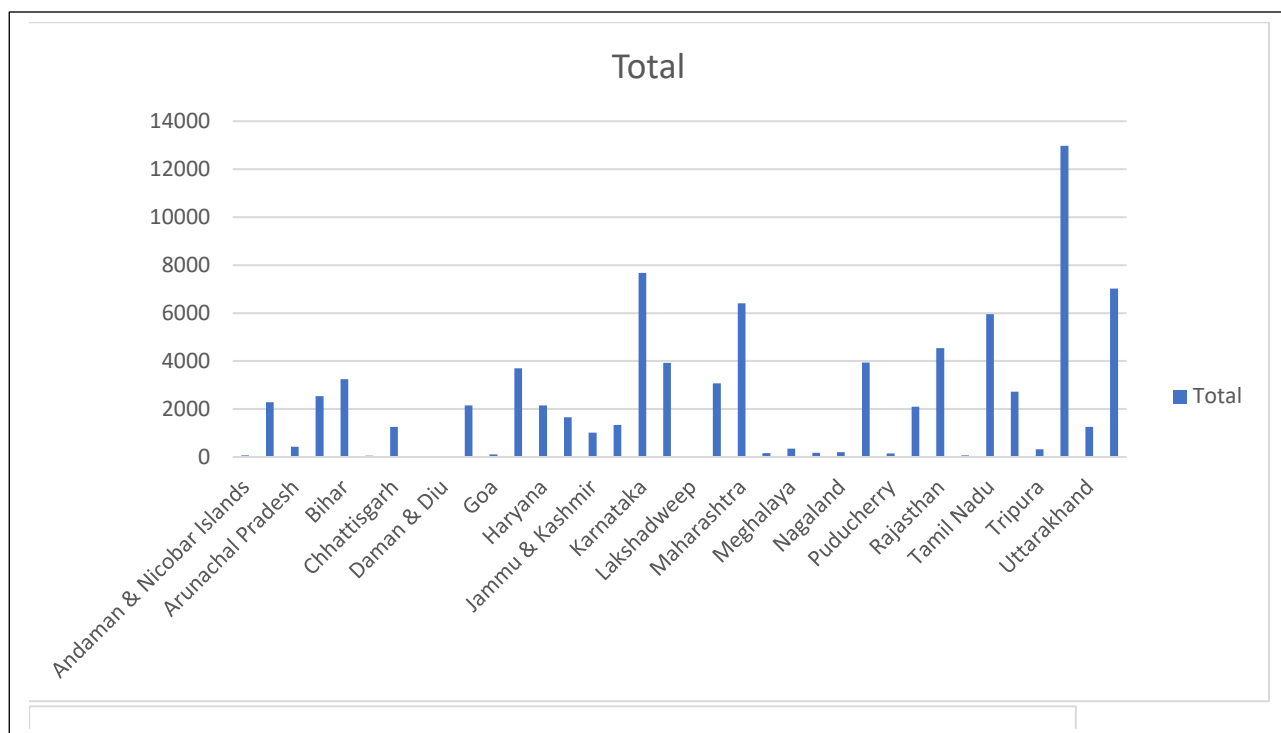
This will be showing the number of discharged patients as per region during the past years.

Result:

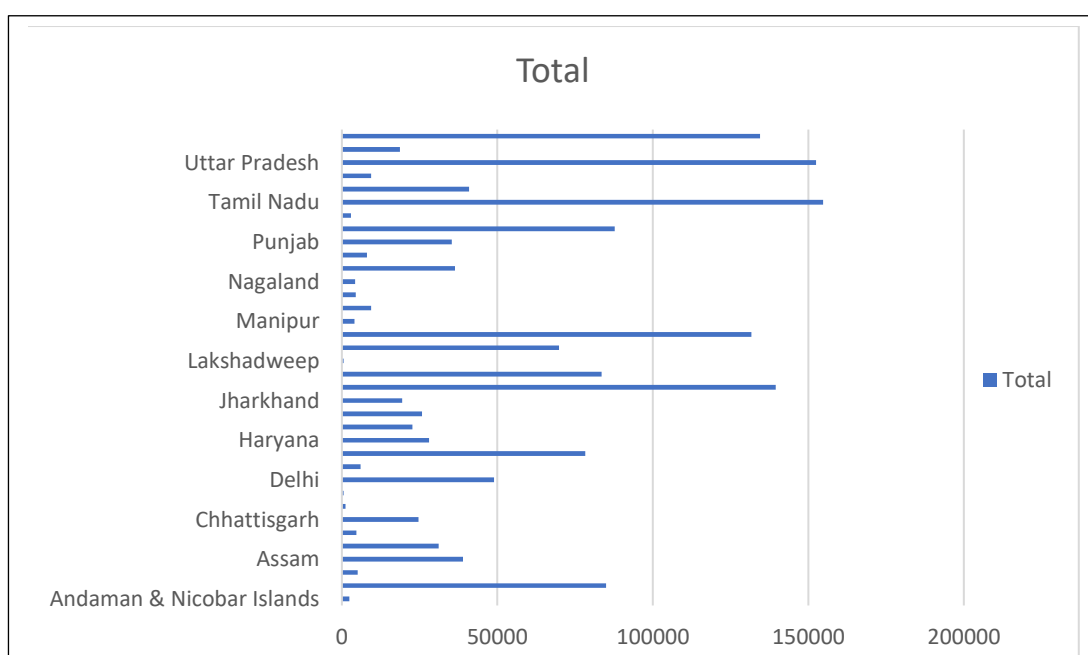


ANALYSIS RESULTS

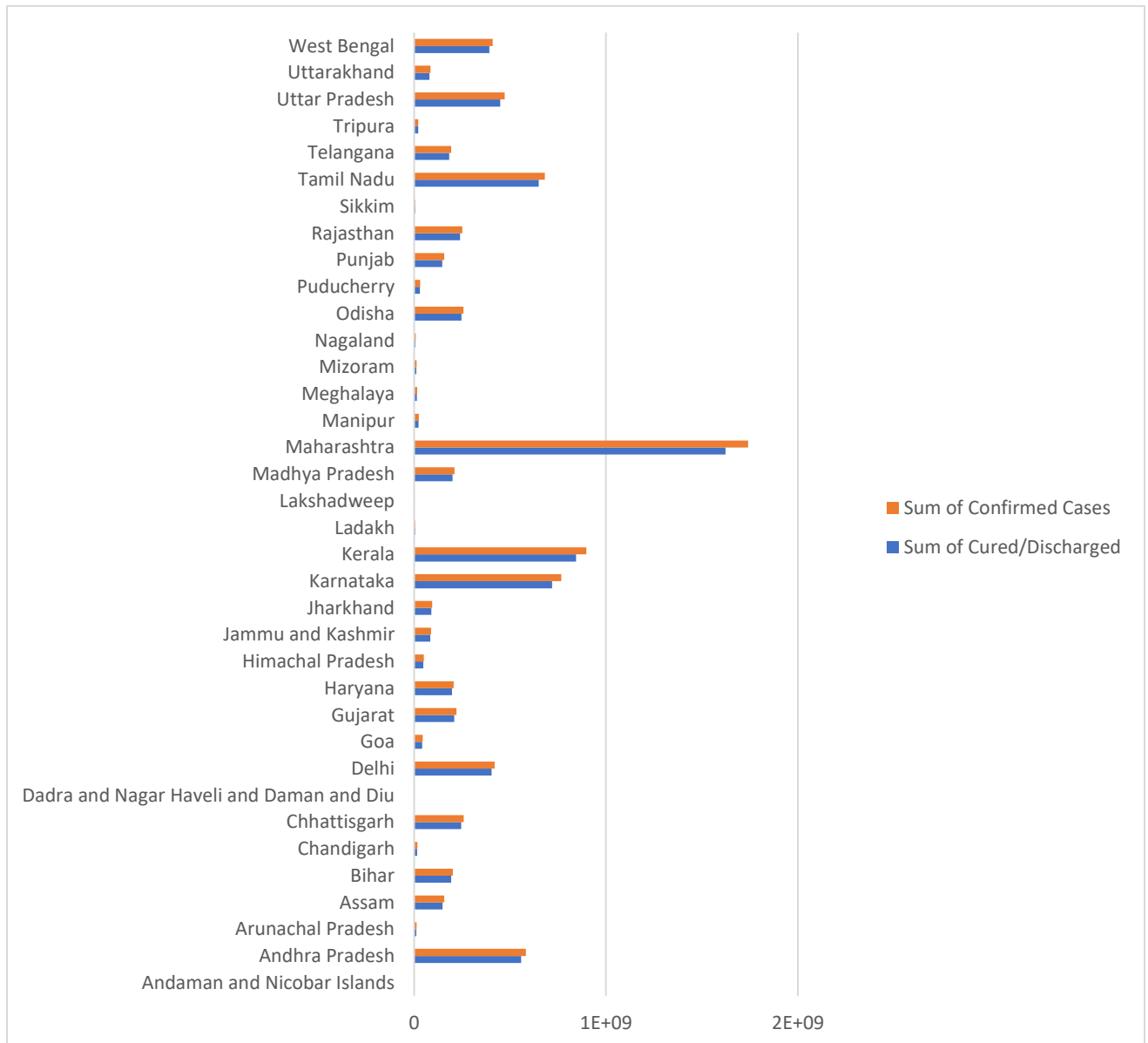
1. Total Hospitals



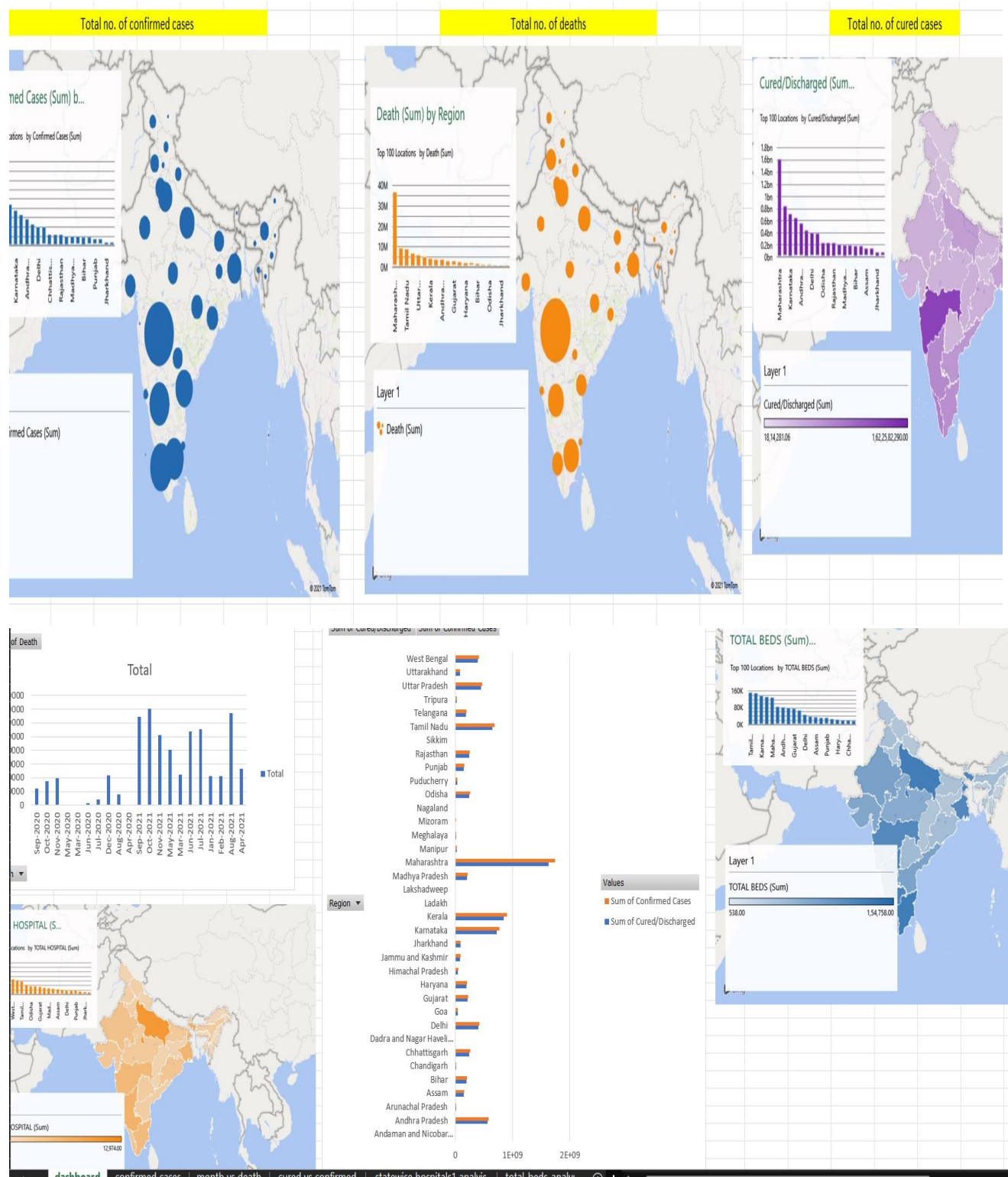
2. Total beds



3. Cured Cases



Final Dashboard



Dashboard Will be giving all the brief details of the of work done on the project.

Future Scope

1.Prediction Model:

India has not reached the peak yet , once it reaches the peak the prediction model can be built to show that how much time it will take to get things back to the normal

2.Sentiment Analysis:

India has never experience such pandemic in last 100 years so what do people think about this pandemic, lockdown, government approach/policies etc can be studied to have sentiment insight of this pandemic.

Reference and Bibliography

- Analytics Vidhya
- Kaggle
- Wikipedia.com
- Google.com

