Statistical Analysis and Visualization of Covid-19

Professor, M Shanmugham Shoba [1], Raahul Narayana Reddy K [2], Prasanna Bhat [3], Apurba Bhattacharjee [4], Srinivas M [5]

[1]*Sr. Asst. Professor, Dept. of Information Science and Engineering, New Horizon College of Engineering, Karnataka, India*
[2-5]*Student, Dept. of Information Science and Engineering, New Horizon College of Engineering, Karnataka, India*

---***---

Abstract - The current technology which is available today, we are proposing a system which is used to implement the statistical and visualization of the COVID-19 cases dataset using data analysis techniques where it will help to give the graphical representation of the COVID-19 situation as the graphical representation gives a serious outlook so that people can now be serious and take precaution to keep themselves safe until the graph shows a decline. We are developing a project that people can access the data related to the increased amount of covid cases using the statistical analysis and visualization data of the covid. This project will help the people to understand the growth and information of the amount of covid cases. The existing system as we have observed that they have only displayed the detail of covid in statistical and visualization of the data set which was not up to date and they have only displayed few data analysis where people won't be updated with cases in the world. We are going to design a statistical and visualization of covid with the latest update of the data set with various graph methods.

This project will also help everyone who is connected to the internet which would reduce the chance of being affected and will reduce the death rate. It will inform the user about the current cases in the way of statistical and visualization of data. This will help the end-users to understand and take the proper precaution of the covid disease. So we chose this topic to give the graphical representation of the COVID-19 situation as the graphical representation gives a serious outlook so that people can now be serious and take precautions to keep themselves safe until the graph shows a decline.

Keywords: Coronavirus, COVID-19, Data analysis, Visualization

1. INTRODUCTION

The novel coronavirus (COVID-19) was widely reported to have first been detected in Wuhan (Hebei area of land, China) in December 2019. After the first sudden start of something bad like disease, COVID-19 continued to spread to all areas of a country in China and very quickly spread to other countries within and outside of Asia. Now, over 45 million cases of infected people have been confirmed in over 180 countries with over 1 million deaths. Although the foundations of this disease are almost the same as the extreme sudden and serious lung-related disease (SARS) virus that took hold of Asia in 2003, it is shown to spread much more easily and there now exists no (disease-preventing treatment)[1].

Statistics is the control/field of study that concerns the collection, organization, analysis, understanding, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is ordinary, to begin with, a (related to studying numbers) population or a (related to studying numbers) model to be studied. Groups of people can be many different kinds of people or things groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments. When (official count of everyone who lives in a country, city, etc.) data cannot be collected, (people who work with numbers for a living) collect data by developing clearly stated/particular experiment designs and survey samples. Representative sampling promises that guesses (based on what you've been told) and ends/ results can (well enough/good enough/in a smart way) extend from the sample to the population as a whole.

An experimental study involves taking measurements of the system under study, controlling/moving around/misleading the system, and then taking added/more measurements using the same procedure to decide/figure out if the moving around/misleading and tricking has changed the values of the measurements. In contrast, a (related to watching or recording something) study does not involve experimental moving around/misleading and tricking.

1.1 Purpose of study

To implement statistical and visualization of covid using data analysis. With the current technology available today, we are proposing a system that is used to implement the statistical and visualization of the COVID-19 cases dataset using data analysis techniques where it will help to give the graphical representation of the COVID-19 situation as the graphical representation gives a serious outlook so that people can now be serious and take precaution to keep themselves safe until the graph shows a decline an existing system as we have observed that they have only displayed the detail of covid in statistical and visualization of the data set which was not up to date and they have only displayed few data analysis where people won't be updated with cases in the world. We are going to design a statistical and visualization of covid with the latest update of the data set with various graph methods.

2. Scope of the Project

As explained earlier in the relevance of the project from the beginning of corona which started in December 2019 the number of growing cases was not clear. There was no proper information on the internet or TV. There was no information on the growth of corona. Without the proper information on the growth of corona in the end the cases increased that it was out of control. Even during the rapidly increased cases of the corona, there was no proper information about the exact number of cases. This had led to misleading the information which was broadcasting on television [2].

In 1919 when the Spanish flu emerged it led to 100million deaths of people because of low health infrastructure and unavailability of the proper information. Since then the health infrastructure has been improved and people are being informed. In 2020 we have experienced the severe growth of the global pandemic which led to a huge number of death cases and without any proper medications, people are dying. Comparing to Spanish flu the covid deaths are much lesser due to the improved infrastructure and digitalization.

This project will help everyone who is connected to the internet which would reduce the chance of being affected and will reduce the death rate. It will inform the user about the current cases in the way of statistical and visualization of data. This will help the end-users to understand and take the proper precaution of the covid disease.

3. Problem Definition

The problem definition is to build an application that aims at creating an improved statistical and visualization of covid data which overcomes the drawbacks of the existing system.

These are the problems that are faced by the existing system are:

- Irregularity in data.
- No proper updation of data.
- No proper visualization of data.
- No auto updation of data.
- Much misleading information.

4. Problem Explanation

Since the very beginning of corona which started in December 2019 the growth of the cases was not clear. There was no proper information on the internet or TV. There was no information on the growth of corona. Without the proper information on the growth of corona in the end the cases increased that it was out of control [3]. Even during the rapidly increased cases of the corona, there was no proper information about the exact number of cases. This had led to misleading the information which was broadcasting on television. This project which is statistical analysis and visualization of data will help everyone who is connected to the internet which would reduce the chance of being affected and will reduce the death rate. It will also inform the user about the current cases in the way of statistical and visualization of data. This will help the end-users to understand and take the proper precaution of the covid disease. This project gives the exact information of the covid cases which will also help people being updated on the growth of the cases.

5. General Description of the System

To implement statistical and visualization of covid using data analysis. With the current technology available today, we are proposing a system that is used to implement the statistical and visualization of the COVID-19 cases dataset using data analysis techniques where it will help to give the graphical representation of the COVID-19 situation as the graphical representation gives a serious outlook so that people can now be serious and take precaution to keep themselves safe until the graph shows a decline an existing system as we have observed that they have only displayed the detail of covid in statistical and visualization of the data set which was not up to date and they have only displayed few data analyses where people won't be updated with cases in the world. We are going to design a statistical and visualization of covid with the latest update of the data set with various graph methods [4].

6. Technical Requirement of the System

A technical requirement relates to the technical aspects that an application must fulfill, such as performance-related problems, reliability matters, and availability concerns. These types of requirements are often called service-level requirements or non-functional requirements.

System requirements define the quality of service a released application must provide to meet the business necessities arrived at. We use the convention analysis and use cases organized with the business requirements to develop application requirements.

1. Hardware Requirements

Hardware requirements are the most common set of requirements defined by any operating system or software application. It is the physical computer resources, also known as hardware.
A hardware requirements list is often complemented by a hardware compatibility list, especially for Operating Systems. The following are needed to efficiently use the application.

Processor - Intel Core i3 and above
Speed - 2.5 GHz
RAM - 8 GB (min)
Hard Disk - 50 GB

A good processor and CPU speed are required to run Jupyter efficiently. The RAM helps in the faster execution of the application. The large storage is used for holding the IDE and the various components of the application. The system should have an internet connection.

2. Software Requirements

Software requirements define software resource fundamentals that need to be installed on a workstation to provide optimum working of the software. The following are required for optimal development and usage of the application.

Operating System - Windows 7 and above
Programming Language - Python 3.7
Compiler - Anaconda, Jupyter

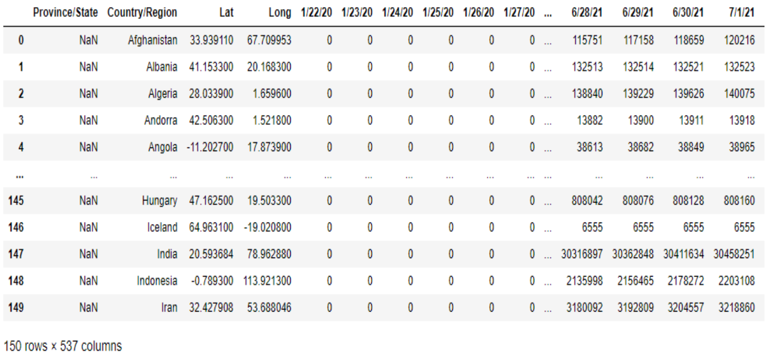
We require the operating system to be Windows 7 or above so Anaconda can run efficiently. This project has been written in Python 3.7. The code is executed on Anaconda.

7. Different Modules of the Project

The various modules used in this project are:

1. Exploring Global Covid Cases

Displaying a table of all country covid cases from the first report to the latest report.



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	6/28/21	6/29/21	6/30/21	7/1/21
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	115751	117158	118659	120216
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	132513	132514	132521	132523
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	138840	139229	139626	140075
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	13882	13900	13911	13918
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	38613	38682	38849	38965
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
145	NaN	Hungary	47.162500	19.503300	0	0	0	0	0	0	...	808042	808076	808128	808160
146	NaN	Iceland	64.963100	-19.020800	0	0	0	0	0	0	...	6555	6555	6555	6555
147	NaN	India	20.593684	78.962880	0	0	0	0	0	0	...	30316897	30362848	30411634	30458251
148	NaN	Indonesia	-0.789300	113.921300	0	0	0	0	0	0	...	2135998	2156465	2178272	2203108
149	NaN	Iran	32.427908	53.688046	0	0	0	0	0	0	...	3180092	3192809	3204557	3218860

150 rows × 537 columns

Fig 1: - Exploring Global covid cases from 22/1/20 till 7/7/21

2. Country specific graphs

Displaying the countries' covid confirmed cases and confirmed deaths using a line graph.

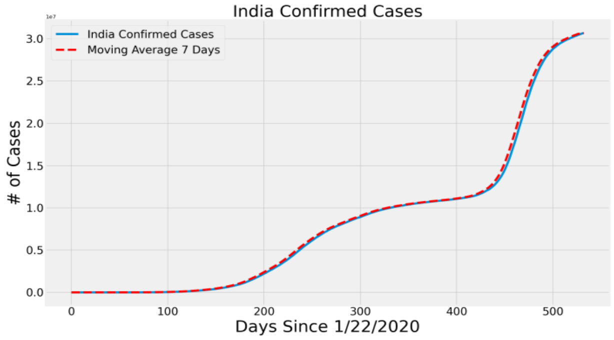


Fig 2: - India confirmed cases

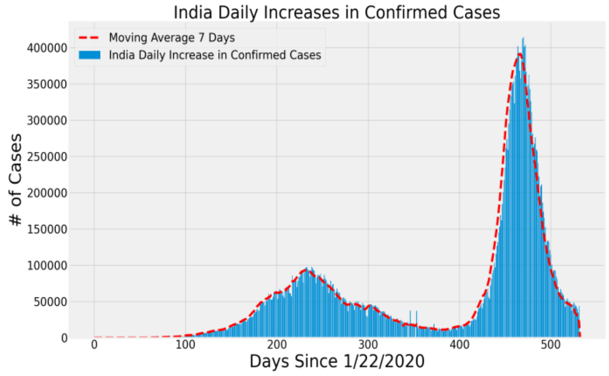


Fig 3: - India daily increase in confirmed cases

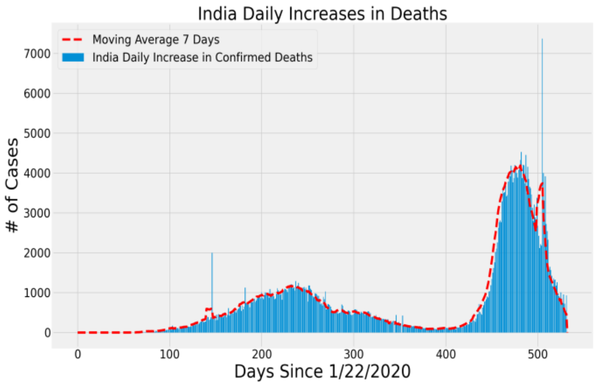


Fig 4: - India daily increase in confirmed deaths

3. Prediction for the conformed covid case using the algorithms

These three models predict future covid cases on a global level.

The prediction models include

- Support Vector Machine
- Polynomial Regression
- Bayesian Ridge Regression

Support Vector Machine: - It is also known as SVM which is one of the most popular supervised learning algorithms which is used for both classification and regression problems, primarily it is basically used for Classification problems in ML.

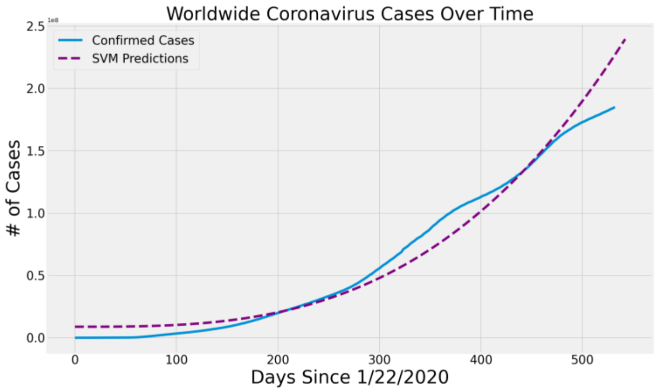


Fig 5: - Predication of conformed covid cases using SVM (**Support Vector Machine)**

Polynomial Regression: - It is one of the regression algorithms which models the relationship between a Dependent (y) and Independent variable (x) as the nth degree polynomial. The Polynomial regression equation is y= b0+b1x1+ b2x12+ b2x13+...... bnx1n. It is also known as a special case of multiple linear regression In Machine Learning.

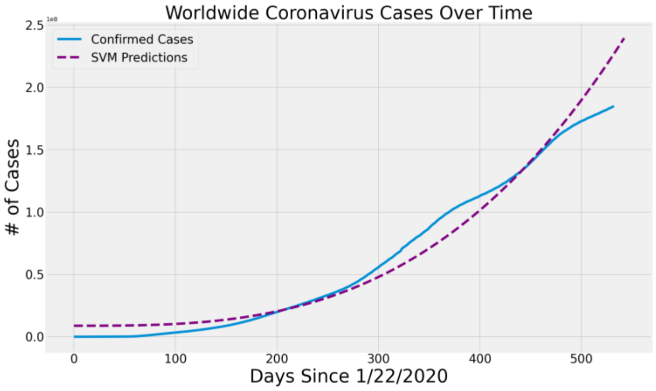


Fig 6: - Predication of conformed covid cases using PR (**Polynomial Regression)**

Bayesian Ridge Regression: - It allows a natural mechanism to survive the poorly distributed data by formatting the linear regression using probability distributors rather than point estimates and the output 'y' is assumed to draw from a probability distribution rather than estimated in a single value

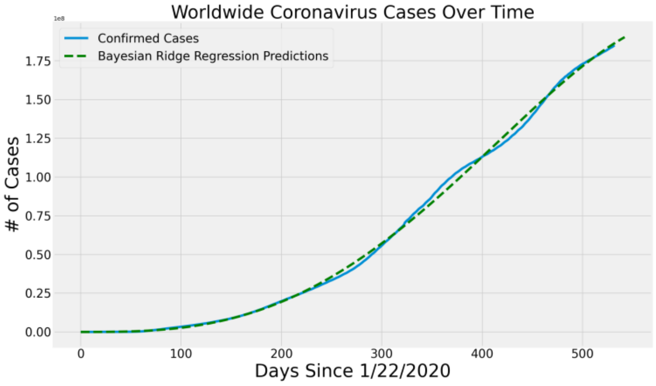
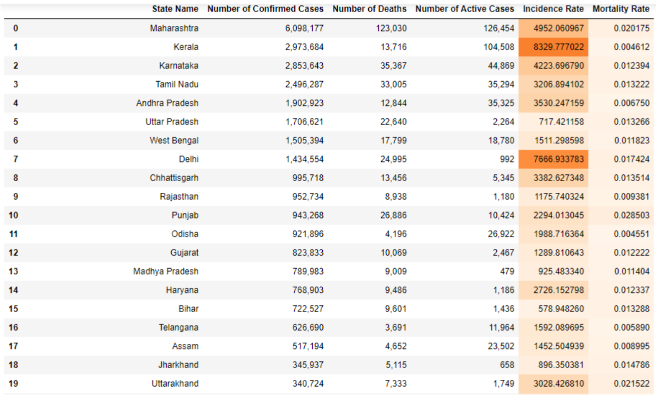


Fig 7: - Predication of conformed covid cases using BRR (**Bayesian Ridge Regression)**

4. Data table

Displaying a table with the latest data of each country with all its states.



	State Name	Number of Confirmed Cases	Number of Deaths	Number of Active Cases	Incidence Rate	Mortality Rate
0	Maharashtra	6,098,177	123,030	126,454	4952.060967	0.020175
1	Kerala	2,973,684	13,716	104,508	8329.777022	0.004612
2	Karnataka	2,853,643	35,367	44,869	4223.696790	0.012394
3	Tamil Nadu	2,496,287	33,005	35,294	3206.894102	0.013222
4	Andhra Pradesh	1,902,923	12,844	35,325	3530.247159	0.006750
5	Uttar Pradesh	1,706,621	22,640	2,264	717.421158	0.013266
6	West Bengal	1,505,394	17,799	18,780	1511.298598	0.011823
7	Delhi	1,434,554	24,995	992	7666.933783	0.017424
8	Chhattisgarh	995,718	13,456	5,345	3382.627348	0.013514
9	Rajasthan	952,734	8,938	1,180	1175.740324	0.009381
10	Punjab	943,268	26,886	10,424	2294.013045	0.028503
11	Odisha	921,896	4,196	26,922	1988.716364	0.004551
12	Gujarat	823,833	10,069	2,467	1289.810643	0.012222
13	Madhya Pradesh	789,983	9,009	479	925.483340	0.011404
14	Haryana	768,903	9,486	1,186	2726.152798	0.012337
15	Bihar	722,527	9,601	1,436	578.948260	0.013288
16	Telangana	626,690	3,691	11,964	1592.089695	0.005890
17	Assam	517,194	4,652	23,502	1452.504939	0.008995
18	Jharkhand	345,937	5,115	658	896.350381	0.014786
19	Uttarakhand	340,724	7,333	1,749	3028.426810	0.021522

Fig 8: - Data table of states in India with its covid cases

5. Bubble Plot

Displaying the covid cases of the top 10 countries in the form of a bubble plot.

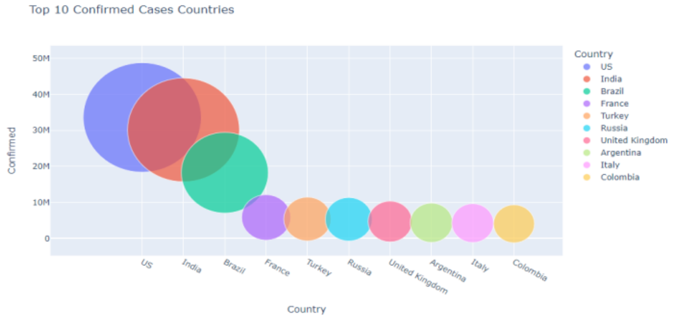


Fig 9: - Top 10 countries confirmed cases using bubble plot

6. Bar chart visualization for covid-19

Displaying the covid cases of top 10 countries with its latest data in the form of a bar chart

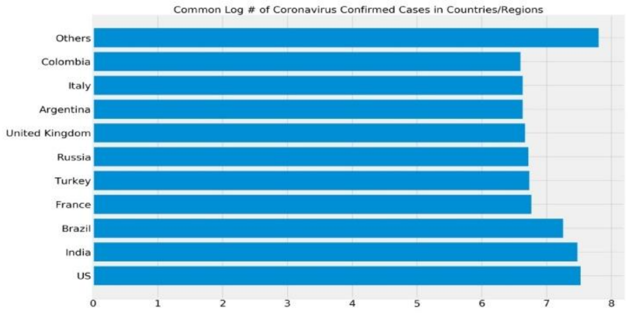


Fig 10: - common log of covid confirmed cases in countries/region

7. Pie chart visualization for covid-19

Displaying the covid cases of each country with its states in the form of a Pie chart.

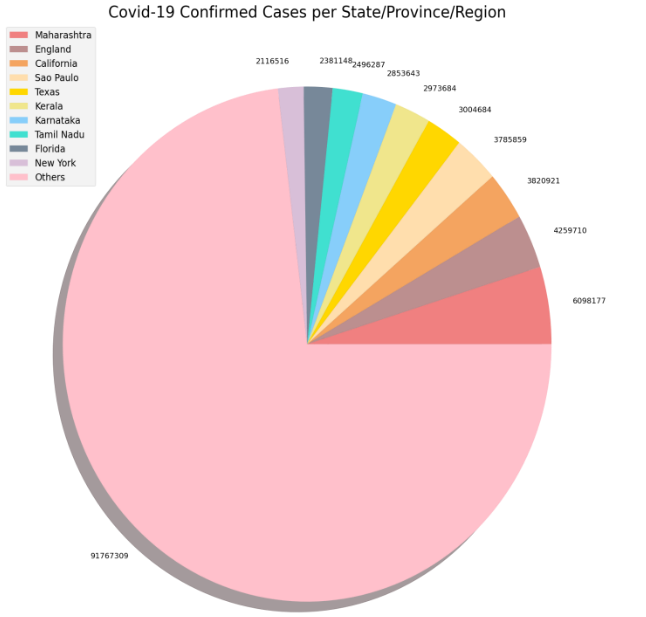


Fig 11: - covid confirmed cases in India using the pie chart

8 . Conclusions

Misleading information is not good for people. People should be provided with the proper information which is very easy to understand. We are developing a project that people can access the data related to the increased amount of covid cases using the statistical analysis and visualization data of the covid. This project will help people to understand the growth and information of the amount of covid cases [5].

To implement statistical and visualization of covid using data analysis. With the current technology available today, we are proposing a system that is used to implement the statistical and visualization of the COVID-19 cases dataset using data analysis techniques where it will help to give the graphical representation of the COVID-19 situation as the graphical representation gives a serious outlook so that people can now be serious and take precaution

to keep themselves safe until the graph shows a decline.

This project will help everyone who is connected to the internet which would reduce the chance of being affected and will reduce the death rate. It will inform the user about the current cases in the way of statistical and visualization of data. This will help the end-users to understand and take the proper precaution of the covid disease.

9. References

[1] Araujo MB, Naimi B. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. Medrxiv; 2020.

[2] An Exploratory Data Analysis of COVID-19 in India, Vol. 9 Issue 04, April-2020, author: Sarvam Mittal Data Science IIIT-B Banglore, India

[3] Analysis of COVID-19 in India using Exploratory Method, Vol. 9 Issue 09, September-2020, author: Adarsh Satsangi Computer Science & Engineering SRM-IST Delhi-NCR, India

[4] Data Analysis for COVID-19, Volume XII, Issue V, May/2020, authors: Noopur Khare, Meenakshi Jha, Runjhun Mathur, Abhimanyu Kumar Jha.

[5] An Investigatory Data Analysis of COVID 19 India Data: About First Three Phases of Lockdown in India, Volume: 07 Issue: 05 | May 2020, authors: E.Prabhakar, T.M.Jacob, Aman Khurana, N.Santhiya