COVID - 19 VACCINATION: DATA ANALYSIS AND VISUALIZATION

RESEARCH INTERNSHIP FOR STUDENTS OF OTHER INSITUTIONS

DEPARTMENT OF INFORMATION TECHNOLOGY
SRI SIVASUBRAMANIYA NADAR COLLEGE OF ENGINEERING
KALAVAKKAM, CHENNAI 603 110

Submitted by

PREMI JAWAHAR VASAGAM (2017504064)

Department of Electronics Engineering
Madras Institute of Technology, Anna University
Chennai - 600044
4th Year

JULY 2021

SSN COLLEGE OF ENGINEERING : CHENNAI 603110

BONAFIDE CERTIFICATE

Certified that project report titled **"COVID-19 VACCINATION: DATA ANALYSIS AND VISUALIZATION"** is the bonafide work of Premi Jawahar Vasagam (2017504064 – 4th Year), Madras Institute of Technology, Anna University, Chennai - 600044 who carried out the project work under our supervision as a research internship project in the Department of Information Technology, SSN College of Engineering during June – July 2021.

SIGNATURE	**SIGNATURE**
Dr. C. Aravindan	Dr. K. Kabilan
HEAD OF DEPARTMENT	**SUPERVISOR**
Professor	Assistant Professor
Department of Information Technology	Department of Information Technology
SSN College of Engineering	SSN College of Engineering
Kalavakkam – 603110.	Kalavakkam – 603110.

Place: Chennai **Date:** 02/08/2021

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Sri Sivasubramaniya Nadar College of Engineering, for giving me this opportunity to take part in this Research Internship for Students of Other Institutions.

I also wish to express my sincere gratitude to Dr. C. ARAVINDAN, Professor and Head of Department of Information Technology, SSN College of Engineering.

With utmost respect, I would like to express my sincere gratitude to my mentor Dr. K. KABILAN, Assistant Professor, Department of Information Technology, for his constant support, technical inputs and guidance which has motivated me through the course of this research internship project.

I would also like to thank Dr. K. R. UTHAYAN, Associate Professor, Department of Information Technology for his valuable insights and motivation that helped me immensely in completing this research internship project.

I wish to express my sincere thanks to all the panel members for their suggestions and encouragement during the review sessions.

Premi Jawahar Vasagam

ABSTRACT

Coronavirus disease (COVID-19) is a communicable disease that is caused by a newly discovered coronavirus. It was first identified in Wuhan, China and since then, it has spread all over the world and has caused the current pandemic.

The purpose is to create a visual representation dashboard on Tableau regarding the significance of vaccines on the infection rate and see the trend of cases over various periods of time during the vaccination drive. The focus is on how the vaccination drive has impacted the time period before, during and after the 2nd wave and vice versa in India with added information on the situation in Tamil Nadu as well. Additionally, time series analysis using ARIMA model has been carried out to forecast the future scenario in India given that the vaccination drive in India is carried out at the same pace. The significance of the vaccine on the number of cases can be understood through this analysis.

The data was collected from COVID19-India's GitHub repository. It has all the data regarding total cases, daily cases, recovered and deceased cases for whole of India and state wise as well. Data regarding vaccinations was also collected from the same site for both India and districts of Tamil Nadu.

Through this project, the actual impact of the pandemic and the significance of the vaccine in controlling it can be realized. Since the dashboards are interactive, it engages the user and relevant information can be accessed easily.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation Function
AR	Auto Regressive
ARIMA	Autoregressive Integrated Moving Average
COVID-19	Coronavirus Disease 2019
MA	Moving Average
PACF	Partial Autocorrelation Function
TSA	Time Series Analysis

CHAPTER 1
INTRODUCTION

COVID-19 has completely changed the way the world works and has caused unprecedented scenarios, and everyone has had to cope with it and deal with personal losses of many types. Information on this disease can be found everywhere and it might even be overwhelming at times to see the information all at once, but it is necessary that the public know the impact of the virus. This will help them realize the extent of danger this virus poses and make them more aware of the happenings around them. Visualizations are easy to understand for all types of people and it can help a large audience keep themselves informed. Seeing a visual representation is always more effective than just reading about it or listening to the numbers on the news.

There is a need for visualizations since it also helps us clearly understand the relation between the cases and the vaccinations that are being carried out. We can see how both these factors affect one another and the impact of vaccination can be seen easily as the days go by and the number of vaccinated people increase. Getting an overall view of the status of the nation will also help us understand which states are still lagging behind and which states are doing a good job. We can analyze further after we are given all the facts needed to arrive at a conclusion.

Time Series Analysis (TSA) is a suitable forecasting method for this project since the outcome variable in our model is dependent on a single variable that is time. We are using the ARIMA model for this project since the assumptions required for it to work are met. The data has been made stationary, it is univariate, and it is in time series format. TSA is used to forecast the future scenario with the ongoing vaccination drive. This will give us a good idea of what to expect and how the situation will look in the best- and worst-case scenarios.

The vaccination has been proven to bring down the severity of the infection if the virus has been contracted and greatly reduces the burden on the healthcare system in India which saw a huge number of cases during the 2nd wave. No one was prepared for this extreme number of casualties and it resulted in lots of unfortunate deaths. The data that is used in this project is from the day COVID-19 vaccinations were started in India, i.e., 16th January 2021. The 2nd wave started just 2 months after the commencement of vaccinations and not many people had been vaccinated at that point. With this project, we will be able to analyze the relationships between the two factors.

The organization of the report is as follows: Chapter 2 elaborates the literature survey conducted for the project. Chapter 3 describes the datasets and techniques used for visualization. Chapter 4 deals with the time series analysis and model used. Chapter 5 discusses the results and its analysis. Chapter 6 concludes with the findings.

CHAPTER 2
LITERATURE SRUVEY

Understanding COVID-19 using Data Visualization (Chauhan R., Goel P., et al (2021))

Purpose of the dashboard is to provide researchers and enthusiasts a place where they can analyze and visualize different aspects, trends, and patterns of COVID-19. It has been divided into 5 tabs that each serve a different purpose: Data Summary, World Data, Visualization, Information and News. The data is collected from trusted sources and made sure that it is updated, correct and easily available. Dashboard contains information like total confirmed cases, active cases, deceased cases, most affected countries and continent, recovery rate, fatality rate, etc. World Data contains special plots for showing each country's situation quickly. It also shows trend of virus in various countries. Visualization gives in-depth analysis and visualization of concerned country. Any country can be chosen and information regarding that country can be viewed. Information page is aimed at removing misconceptions regarding the virus and safety precautions that need to be taken. News page gives live and updated news about the virus from trusted sources. Data is directly collected from John Hopkins University's GitHub Repository. All the plots are interactive so as to increase user experience and engagement.

Visual Exploratory Data Analysis of Covid-19 Pandemic (Saini S. K., Dhull V., et al (2020))

It is important to analyze the worldwide pandemic spread so that certain guide strategies can be set for complete situational awareness and application of conventional methodologies to control the impacts caused by it globally. This paper is composed of the visual exploratory data analysis of the countries based on the number of confirmed, recovered and death cases along with the comparative analysis of the mortality and recovery rate for nearly 222 nations worldwide. This study can be used to evaluate the rise of risks in a given area by comparing the count of the cases via visual analysis and work on the set up of some strategies to control its spread globally.

Big Data Visualization and Visual Analytics of COVID-19 Data (Leung C. K., Chen Y., et al ((2020))

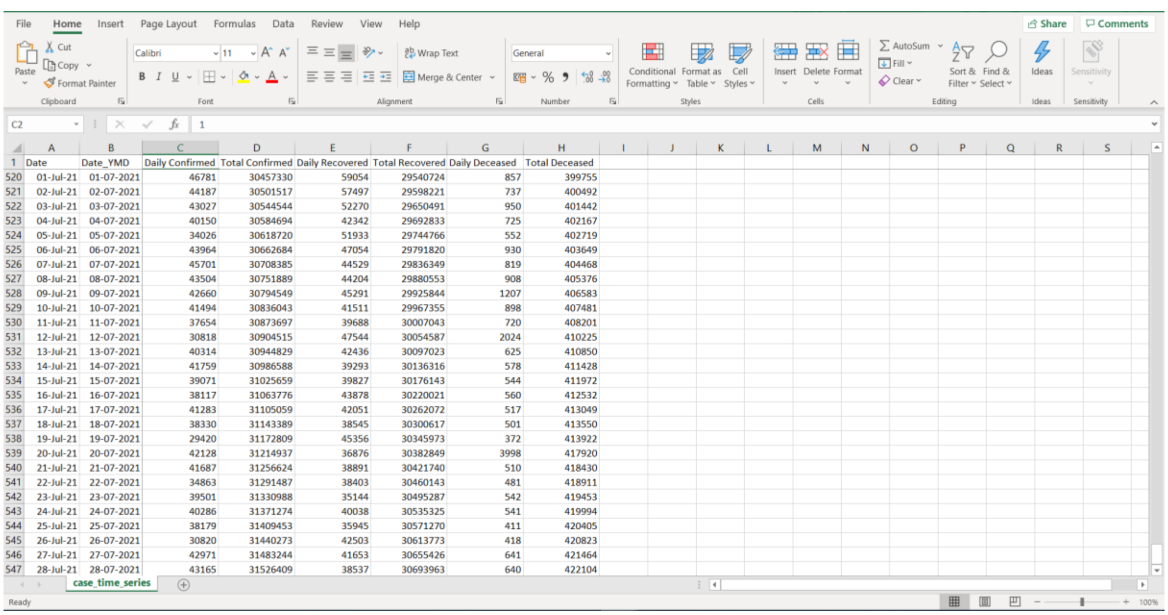
A huge amount of data has been generated and collected from a wide variety of rich data sources. Embedded in these big data are useful information and valuable knowledge. An example is healthcare and epidemiological data such as data related to patients who suffered from epidemic diseases like the coronavirus disease 2019 (COVID-19). Knowledge discovered from these epidemiological data helps researchers, epidemiologists, and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. As “a picture is worth a thousand words”, having methods to visualize and visually analyze these big data makes it easily to comprehend the data and the discovered knowledge. A big data visualization and visual analytics tool for visualizing and analyzing COVID-19 epidemiological data has been done. The tool helps users to get a better understanding of information about the confirmed cases of COVID-19. Although this tool is designed for visualization and visual analytics of epidemiological data, it is applicable to visualization and visual analytics of big data from many other real-life applications and services.

CHAPTER 3
DATASETS AND VISUALIZATION TECHNIQUES

This section gives detailed view about the datasets and visualization techniques that were used to create the dashboards for this project in the form of modules.

3.1 DATASETS USED

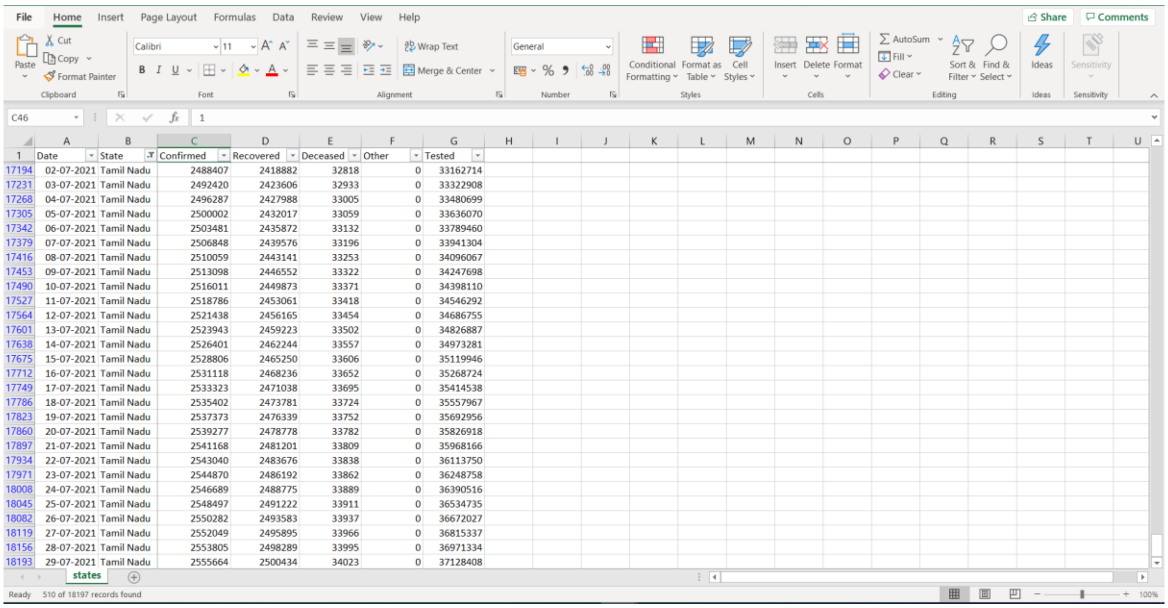
For this project, the datasets were collected from COVID19-India's GitHub repository. There were various datasets to choose from but for this project, three datasets were chosen. The data is updated on a daily basis and its credibility has been confirmed. Time series data for India and its states, vaccination data: state wise and district wise datasets were used. First dataset contains India level timeseries for confirmed, recovered, and deceased cases. The datapoints for this project start from when the vaccination drive started in India, i.e., 16th January 2021. The 2nd wave started in India around 16th March 2021.



Date	Date_YMD	Daily Confirmed	Total Confirmed	Daily Recovered	Total Recovered	Daily Deceased	Total Deceased
01-Jul-21	01-07-2021	46781	30457330	59054	29540724	857	399755
02-Jul-21	02-07-2021	44187	30501517	57497	29598221	737	400492
03-Jul-21	03-07-2021	43027	30544544	52270	29650491	950	401442
04-Jul-21	04-07-2021	40150	30584694	42342	29692833	725	402167
05-Jul-21	05-07-2021	34026	30618720	51933	29744766	552	402719
06-Jul-21	06-07-2021	43964	30662684	47054	29791820	930	403649
07-Jul-21	07-07-2021	45701	30708385	44529	29836349	819	404468
08-Jul-21	08-07-2021	43504	30751889	44204	29880553	908	405376
09-Jul-21	09-07-2021	42660	30794549	45291	29925844	1207	406583
10-Jul-21	10-07-2021	41494	30836043	41511	29967355	898	407481
11-Jul-21	11-07-2021	37654	30873697	39688	30007043	720	408201
12-Jul-21	12-07-2021	30818	30904515	47544	30054587	2024	410225
13-Jul-21	13-07-2021	40314	30944829	42436	30097023	625	410850
14-Jul-21	14-07-2021	41759	30986588	39293	30136316	578	411428
15-Jul-21	15-07-2021	39071	31025659	39827	30176143	544	411972
16-Jul-21	16-07-2021	38117	31063776	43878	30220021	560	412532
17-Jul-21	17-07-2021	41283	31105059	42051	30262072	517	413049
18-Jul-21	18-07-2021	38330	31143389	38545	30300617	501	413550
19-Jul-21	19-07-2021	29420	31172809	45356	30345973	372	413922
20-Jul-21	20-07-2021	42128	31214937	36876	30382849	3998	417920
21-Jul-21	21-07-2021	41687	31256624	38891	30421740	510	418430
22-Jul-21	22-07-2021	34863	31291487	38403	30460143	481	418911
23-Jul-21	23-07-2021	39501	31330988	35144	30495287	542	419453
24-Jul-21	24-07-2021	40286	31371274	40038	30535325	541	419994
25-Jul-21	25-07-2021	38179	31409453	35945	30571270	411	420405
26-Jul-21	26-07-2021	30820	31440273	42503	30613773	418	420823
27-Jul-21	27-07-2021	42971	31483244	41653	30655426	641	421464
28-Jul-21	28-07-2021	43165	31526409	38537	30693963	640	422104

Figure 3.1.1 Time Series Data for India

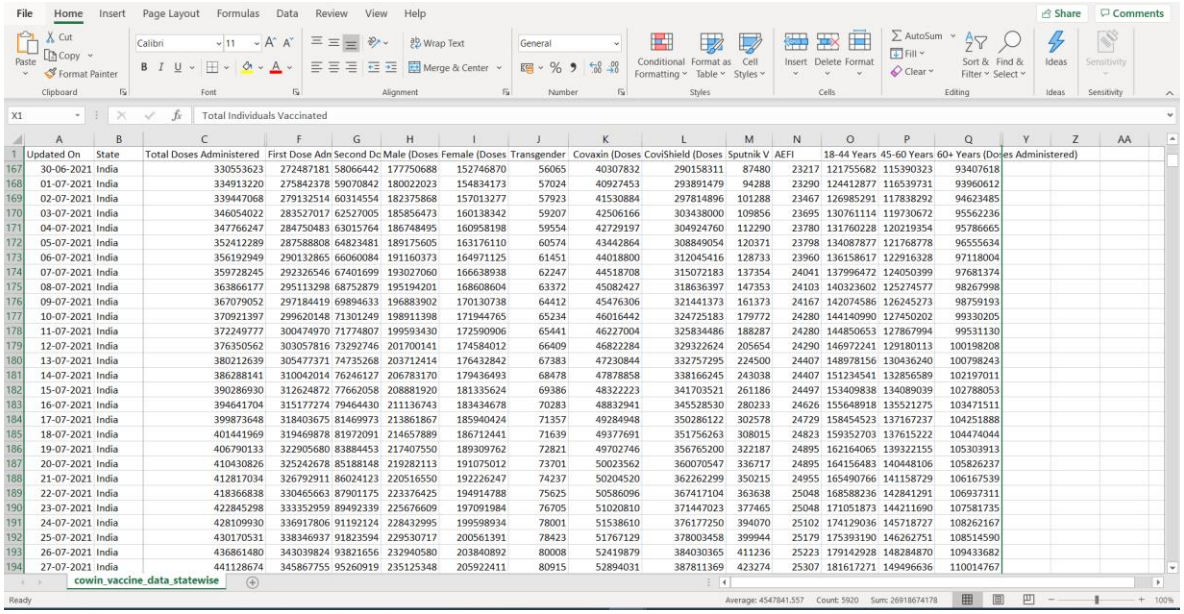
Second dataset contains time series data for each state in India and has the same type of data as time series data for India. Data regarding Tamil Nadu was collected from this dataset.



	Date	State	Confirmed	Recovered	Deceased	Other	Tested
17194	02-07-2021	Tamil Nadu	2488407	2418882	32818	0	33162714
17231	03-07-2021	Tamil Nadu	2492420	2423606	32933	0	33322908
17268	04-07-2021	Tamil Nadu	2496287	2427988	33005	0	33480699
17305	05-07-2021	Tamil Nadu	2500002	2432017	33059	0	33636070
17342	06-07-2021	Tamil Nadu	2503481	2435872	33132	0	33789460
17379	07-07-2021	Tamil Nadu	2506848	2439576	33196	0	33941304
17416	08-07-2021	Tamil Nadu	2510059	2443141	33253	0	34096067
17453	09-07-2021	Tamil Nadu	2513098	2446552	33322	0	34247698
17490	10-07-2021	Tamil Nadu	2516011	2449873	33371	0	34398110
17527	11-07-2021	Tamil Nadu	2518786	2453061	33418	0	34546292
17564	12-07-2021	Tamil Nadu	2521438	2456165	33454	0	34686755
17601	13-07-2021	Tamil Nadu	2523943	2459223	33502	0	34826887
17638	14-07-2021	Tamil Nadu	2526401	2462244	33557	0	34973281
17675	15-07-2021	Tamil Nadu	2528806	2465250	33606	0	35119946
17712	16-07-2021	Tamil Nadu	2531118	2468236	33652	0	35268724
17749	17-07-2021	Tamil Nadu	2533323	2471038	33695	0	35414538
17786	18-07-2021	Tamil Nadu	2535402	2473781	33724	0	35557967
17823	19-07-2021	Tamil Nadu	2537373	2476339	33752	0	35692956
17860	20-07-2021	Tamil Nadu	2539277	2478778	33782	0	35826918
17897	21-07-2021	Tamil Nadu	2541168	2481201	33809	0	35968166
17934	22-07-2021	Tamil Nadu	2543040	2483676	33838	0	36113750
17971	23-07-2021	Tamil Nadu	2544870	2486192	33862	0	36248758
18008	24-07-2021	Tamil Nadu	2546689	2488775	33889	0	36390516
18045	25-07-2021	Tamil Nadu	2548497	2491222	33911	0	36534735
18082	26-07-2021	Tamil Nadu	2550282	2493583	33937	0	36672027
18119	27-07-2021	Tamil Nadu	2552049	2495895	33966	0	36815337
18156	28-07-2021	Tamil Nadu	2553805	2498289	33995	0	36971334
18193	29-07-2021	Tamil Nadu	2555664	2500434	34023	0	37128408

Figure 3.1.2 Time Series Data for Tamil Nadu

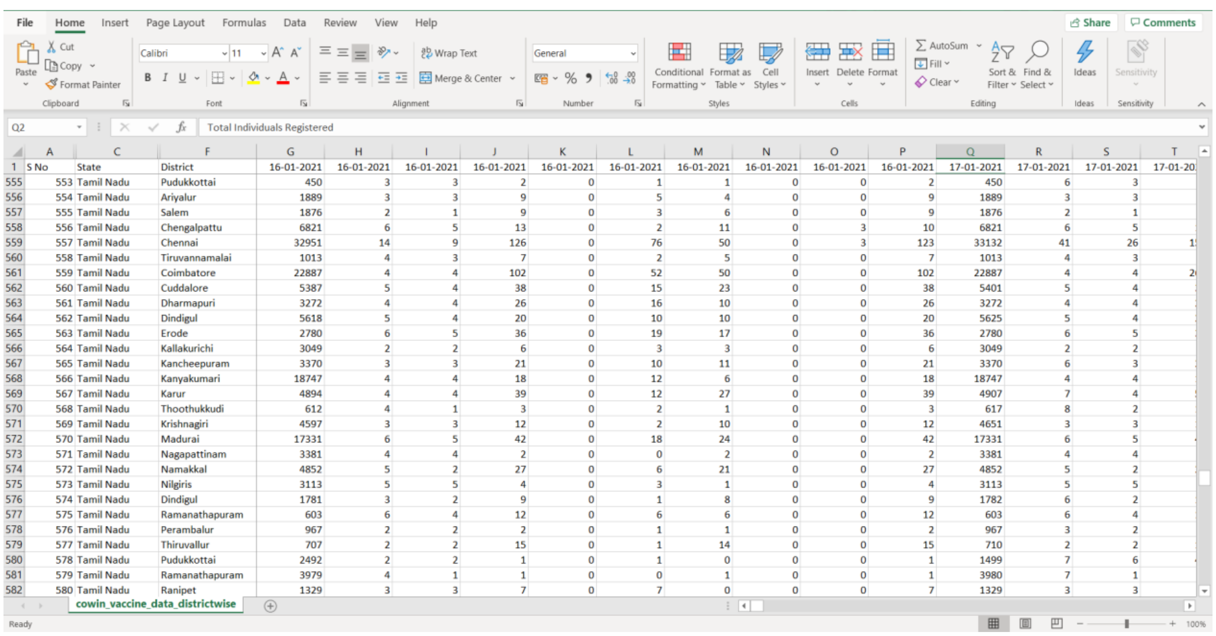
Third dataset contains key data points from CoWin database at a state level. Vaccination information on individual states is present. There are many data points available in this dataset such as total doses, type of vaccine (CoviShield, Covaxin or Sputnik V), age group and gender of vaccinated individual, and dosage type (1st or 2nd dose) but we are concentrating only on dosage type and total vaccinated individuals. Vaccination data for whole of India and Tamil Nadu separately was collected from this dataset



	Updated On	State	Total Doses Administered	First Dose Adn	Second Dc	Male (Doses	Female (Doses	Transgender	Covaxin (Doses	CoviShield (Doses	Sputnik V	AEFI	18-44 Years	45-60 Years	60+ Years (Do
167	30-06-2021	India	330553623	272487181	58066442	177750688	152746870	56065	40307832	290158311	87480	23217	121755682	115390323	93407618
168	01-07-2021	India	334913220	275842378	59070842	180022023	154834173	57024	40927453	293891479	94288	23290	124412877	116539731	93960612
169	02-07-2021	India	339447068	279132514	60314554	182375868	157013277	57923	41530884	297814896	101288	23467	126985291	117838292	94623485
170	03-07-2021	India	346054022	283527017	62527005	185856473	160138342	59207	42506166	303438000	109856	23695	130761114	119730672	95562236
171	04-07-2021	India	347766247	284750483	63015764	186748495	160958198	59554	42729197	304924760	112290	23780	131760228	120219354	95786665
172	05-07-2021	India	352412289	287588808	64823481	189175605	163176110	60574	43442864	308849054	120371	23798	134087877	121768778	96555634
173	06-07-2021	India	356192949	290132865	66060084	191160373	164971125	61451	44018800	312045416	128733	23960	136158617	122916328	97118004
174	07-07-2021	India	359728245	292326546	67401699	193027060	166638938	62247	44518708	315072183	137354	24041	137996472	124050399	97681374
175	08-07-2021	India	363866177	295113298	68752879	195194201	168608604	63372	45082427	318636397	147353	24103	140323602	125274577	98267998
176	09-07-2021	India	367079052	297184419	69894633	196883902	170130738	64412	45476306	321441373	161373	24167	142074586	126245273	98759193
177	10-07-2021	India	370921397	299620148	71301249	198911398	171944765	65234	46016442	324725183	179772	24280	144140990	127450202	99330205
178	11-07-2021	India	372249777	300474970	71774807	199593430	172590906	65441	46227004	325834486	188287	24280	144850653	127867994	99531130
179	12-07-2021	India	376350562	303057816	73292746	201700141	174584012	66409	46822284	329322624	205654	24290	146972241	129180113	100198208
180	13-07-2021	India	380212639	305477371	74735268	203712414	176432842	67383	47230844	332757295	224500	24407	148978156	130436240	100798243
181	14-07-2021	India	386288141	310042014	76246127	206783170	179436493	68478	47878858	338166245	243038	24407	151234541	132856589	102197011
182	15-07-2021	India	390286930	312624872	77662058	208881920	181335624	69386	48322223	341703521	261186	24497	153409838	134089039	102788053
183	16-07-2021	India	394641704	315177274	79464430	211136743	183434678	70283	48832941	345528530	280233	24626	155648918	135521275	103471511
184	17-07-2021	India	399873648	318403675	81469973	213861867	185940424	71357	49284948	350286122	302578	24729	158454523	137167237	104251888
185	18-07-2021	India	401441969	319469878	81972091	214657889	186712441	71639	49377691	351756263	308015	24823	159352703	137615222	104474044
186	19-07-2021	India	406790133	322905680	83884453	217407550	189309762	72821	49702746	356765200	322187	24895	162164065	139322155	105303913
187	20-07-2021	India	410430826	325242678	85188148	219282113	191075012	73701	50023562	360070547	336717	24895	164156483	140448106	105826237
188	21-07-2021	India	412817034	326792911	86024123	220516550	192226247	74237	50204520	362262299	350215	24955	165490766	141158729	106167539
189	22-07-2021	India	418366838	330465663	87901175	223376425	194914788	75625	50586096	367417104	363638	25048	168588236	142841291	106937311
190	23-07-2021	India	422845298	333352959	89492339	225676609	197091984	76705	51020810	371447023	377465	25048	171051873	144211690	107581735
191	24-07-2021	India	428109930	336917806	91192124	228432995	199598934	78001	51538610	376177250	394070	25102	174129036	145718727	108262167
192	25-07-2021	India	430170531	338346937	91823594	229530717	200561391	78423	51767129	378003458	399944	25179	175393190	146262751	108514590
193	26-07-2021	India	436861480	343039824	93821656	232940580	203840892	80008	52419879	384030365	411236	25223	179142928	148284870	109433682
194	27-07-2021	India	441128674	345867755	95260919	235125348	205922411	80915	52894031	387811369	423274	25307	181617271	149496636	110014767

Figure 3.1.3 Vaccination Data for India

Fourth dataset contains key data points from CoWin database at a district level. Here, information on each district within a state is present and contains the same type of information as the third dataset. Vaccination data regarding the different districts in Tamil Nadu were collected from this dataset.



	S No	State	District	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	16-01-2021	17-01-2021	17-01-2021	17-01-2021
555	553	Tamil Nadu	Pudukkottai	450	3	3	2	0	1	1	0	0	2	450	6	3
556	554	Tamil Nadu	Ariyalur	1889	3	3	9	0	5	4	0	0	9	1889	3	3
557	555	Tamil Nadu	Salem	1876	2	1	9	0	3	6	0	0	9	1876	2	1
558	556	Tamil Nadu	Chengalpattu	6821	6	5	13	0	2	11	0	3	10	6821	6	5
559	557	Tamil Nadu	Chennai	32951	14	9	126	0	76	50	0	3	123	33132	41	26
560	558	Tamil Nadu	Tiruvannamalai	1013	4	3	7	0	2	5	0	0	7	1013	4	3
561	559	Tamil Nadu	Coimbatore	22887	4	4	102	0	52	50	0	0	102	22887	4	4
562	560	Tamil Nadu	Cuddalore	5387	5	4	38	0	15	23	0	0	38	5401	5	4
563	561	Tamil Nadu	Dharmapuri	3272	4	4	26	0	16	10	0	0	26	3272	4	4
564	562	Tamil Nadu	Dindigul	5618	5	4	20	0	10	10	0	0	20	5625	5	4
565	563	Tamil Nadu	Erode	2780	6	5	36	0	19	17	0	0	36	2780	6	5
566	564	Tamil Nadu	Kallakurichi	3049	2	2	6	0	3	3	0	0	6	3049	2	2
567	565	Tamil Nadu	Kancheepuram	3370	3	3	21	0	10	11	0	0	21	3370	6	3
568	566	Tamil Nadu	Kanyakumari	18747	4	4	18	0	12	6	0	0	18	18747	4	4
569	567	Tamil Nadu	Karur	4894	4	4	39	0	12	27	0	0	39	4907	7	4
570	568	Tamil Nadu	Thoothukkudi	612	4	1	3	0	2	1	0	0	3	617	8	2
571	569	Tamil Nadu	Krishnagiri	4597	3	3	12	0	2	10	0	0	12	4651	3	3
572	570	Tamil Nadu	Madurai	17331	6	5	42	0	18	24	0	0	42	17331	6	5
573	571	Tamil Nadu	Nagapattinam	3381	4	4	2	0	0	2	0	0	2	3381	4	4
574	572	Tamil Nadu	Namakkal	4852	5	2	27	0	6	21	0	0	27	4852	5	2
575	573	Tamil Nadu	Nilgiris	3113	5	5	4	0	3	1	0	0	4	3113	5	5
576	574	Tamil Nadu	Dindigul	1781	3	2	9	0	1	8	0	0	9	1782	6	2
577	575	Tamil Nadu	Ramanathapuram	603	6	4	12	0	6	6	0	0	12	603	6	4
578	576	Tamil Nadu	Perambalur	967	2	2	2	0	1	1	0	0	2	967	3	2
579	577	Tamil Nadu	Thiruvallur	707	2	2	15	0	1	14	0	0	15	710	2	2
580	578	Tamil Nadu	Pudukkottai	2492	2	2	1	0	1	0	0	0	1	1499	7	6
581	579	Tamil Nadu	Ramanathapuram	3979	4	1	1	0	0	1	0	0	1	3980	7	1
582	580	Tamil Nadu	Ranipet	1329	3	3	7	0	7	0	0	0	7	1329	3	3

Figure 3.1.4 Vaccination Data for Tamil Nadu Districts

3.2 VISUALIZATION TECHNIQUES

The visualizations in this project have been done in Tableau. It is a data visualization platform that helps to simplify raw data into the form of dashboards and worksheets. This is very useful in trying to understand the data easily and accurately. It combines Python's packages and uses Tableau's SQL database connection, so it becomes efficient to describe and visualize the data. Tableau was preferred over Python for the visualization part since it is specifically made for creating visualizations and the dashboards are interactive as well. This creates a better experience for the user and no prior knowledge is needed to navigate through various areas of the dashboards.

Various workbooks had been created that showcases the situation in India and Tamil before, during and after the 2nd wave. Two maps were also plotted which shows vaccination status of various states in India and then the districts in Tamil Nadu

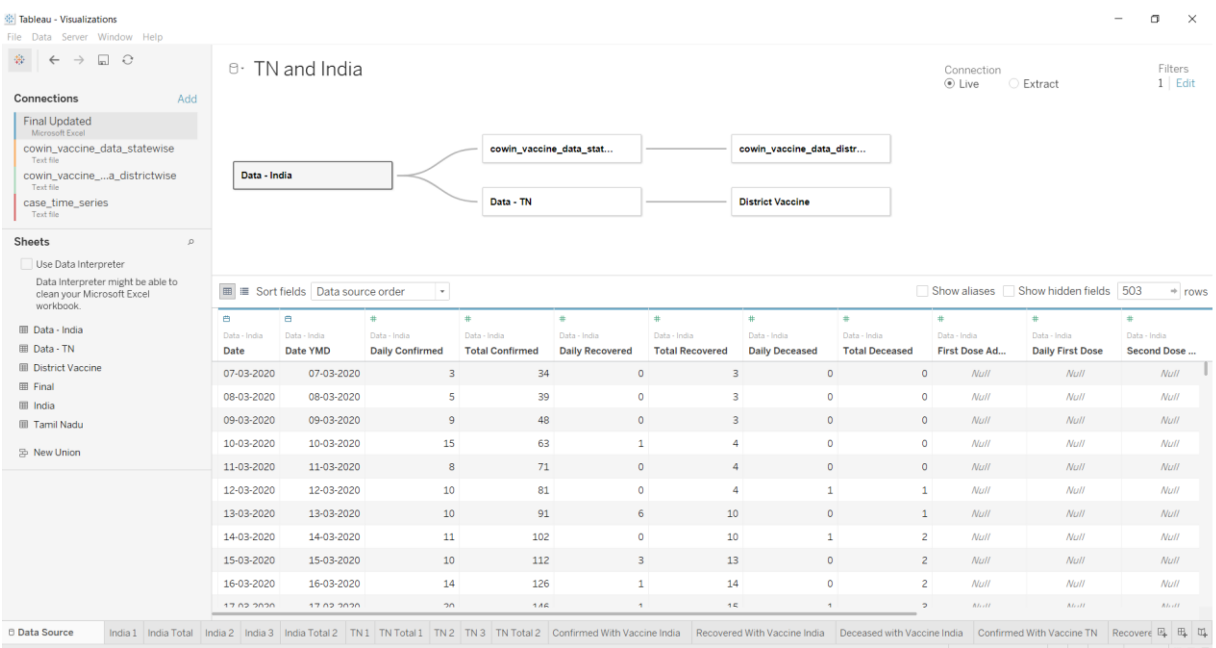


Figure 3.2.1 Data Source in Tableau

CHAPTER 4
TIME SERIES ANALYSIS

4.1 COMPONENTS OF TIME SERIES ANALYSIS

As mentioned previously, time series analysis is used in this project since the outcome in our model is dependent on a single variable: time. There are 3 components in time series analysis:

1. General Trend
2. Seasonality
3. Irregular Fluctuations

General trend is the trend line of our data. They are used to predict the continuation of a certain trend available. It is also used to identify the correlation between two variables (timeline and number of confirmed cases in this project) by observing the trend in both of them simultaneously.

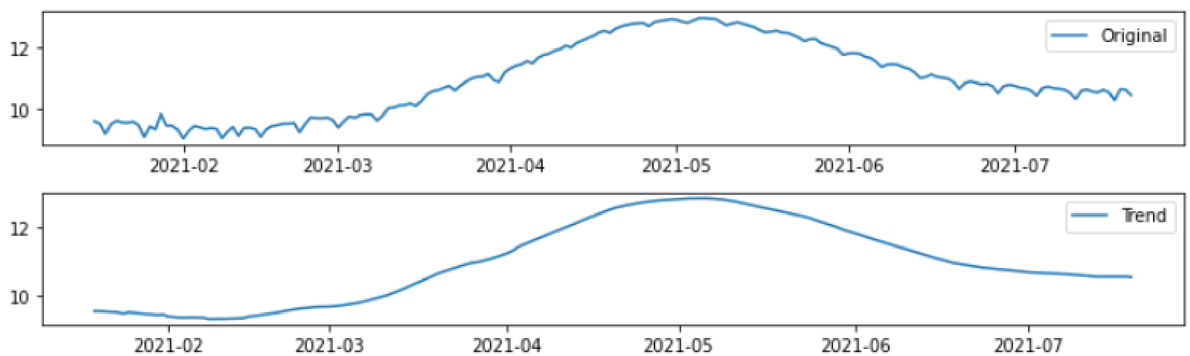


Figure 4.1.1 Trend Line Comparison with Original Data

Seasonality is major factor in TSA. In this project, seasonality can be taken as a spike in data points. This would be due to the sudden outbreak of a new variant as was the case during the 2nd wave. The peak was sudden and then decreased after a while.

Irregular fluctuations are the random components of TSA. They are uncontrolled situations where the number of confirmed cases would change. This could be because of backdated data being added later. This does not mean there was a sudden spike in cases. This is known as random effect.

4.2 IMPLEMENTATION OF TIME SERIES ANALYSIS

We are using TSA to forecast the number of cases in the upcoming months in India. For this, we are using ARIMA model. The relevant datapoints had to extracted from the dataset and the data had to be cleaned to remove a few null values.


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.pylab import rcParams
rcParams['figure.figsize']=10,6
```

```
url = ('https://raw.githubusercontent.com/Premi04/Covid-19-Visualization/main/case_time_series1.csv')
case=pd.read_csv(url)
final = case.drop(['Date'],axis=1).iloc[352:,:2]


final['Date_YMD'] = pd.to_datetime(final['Date_YMD'], infer_datetime_format=True)
indexedDataset = final.set_index(['Date_YMD'])
indexedDataset.head()
```

	Daily Confirmed
Date_YMD	
2021-01-16	15050
2021-01-17	13962
2021-01-18	9987
2021-01-19	13787
2021-01-20	15279

Figure 4.2.1 Initialization of Data

Next, the data has to be verified to be stationary, i.e., have equal mean, variance, and covariance across different time intervals.

```
from statsmodels.tsa.stattools import adfuller
def test_stationarity(timeseries):
  movingAverage = timeseries.rolling(window=14).mean()
  movingSTD = timeseries.rolling(window=14).std()

  orig=plt.plot(timeseries, color='blue', label='Original')
  mean=plt.plot(movingAverage, color='red', label='Rolling Mean')
  std=plt.plot(movingSTD,color='black',label='Rolling Std')
  plt.legend(loc='best')
  plt.title('Rolling Mean and Standard Deviation')
  plt.show(block=False)

  print('Results of Dickey-Fuller Test: ')
  dftest = adfuller(timeseries['Daily Confirmed'], autolag='AIC')
  dfoutput = pd.Series(dftest[0:4],index = ['Test Statistic', 'p-value', '#Lags Used', 'No. of Observations Used'])
  for key,value in dftest[4].items():
    dfoutput['Critical Value (%s)' %key] = value
  print(dfoutput)
```

Figure 4.2.2 Testing for Stationarity

```
test_stationarity(datasetLogScaleMinusMovingAverage)
```

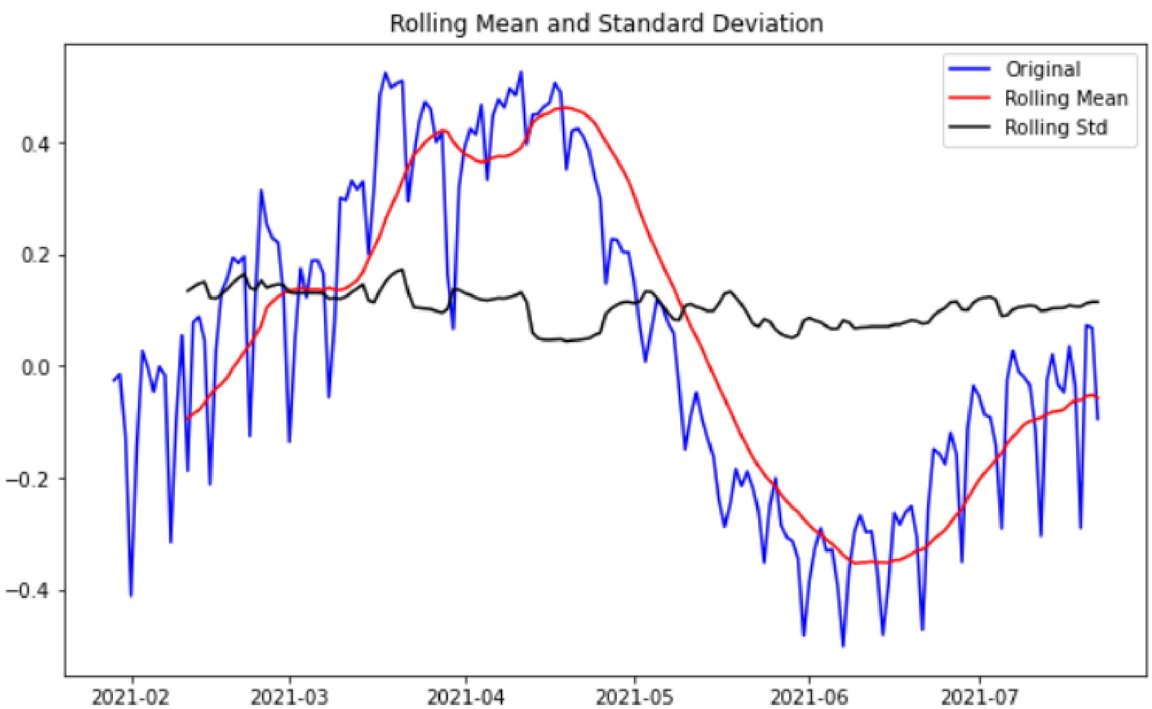


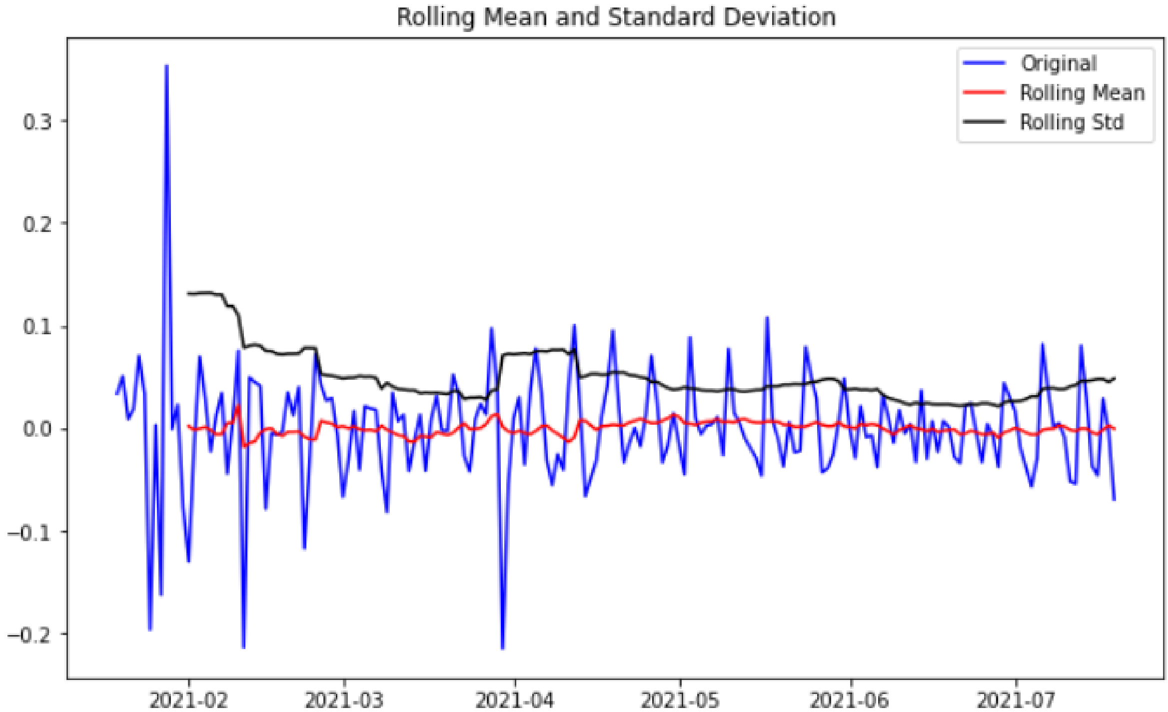
Figure 4.2.3 Rolling Mean and Standard Deviation

From the above figure, we can see that the data is not stationary, the mean line is varying, and the variance is differing as well. The data has to be transformed to be made stationary. There are different techniques to make the data stationary. Data transformations like logarithms can help to stabilize the variance. Differencing can also be used to stabilize the mean. Differencing means calculating the difference between consecutive observations. Both logarithms and differencing has been applied to make the data stationary.

```
from statsmodels.tsa.seasonal import seasonal_decompose
decomposition = seasonal_decompose(indexedDataset_logScale)

trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

plt.subplot(411)
plt.plot(indexedDataset_logScale, label = 'Original')
plt.legend(loc='best')
plt.subplot(412)
plt.plot(trend, label = 'Trend')
plt.legend(loc='best')
plt.subplot(413)
plt.plot(seasonal, label = 'Seasonality')
plt.legend(loc='best')
plt.subplot(414)
plt.plot(residual, label = 'Residuals')
plt.legend(loc='best')
plt.tight_layout()

decomposedLogData = residual
decomposedLogData.dropna(inplace=True)
test_stationarity(decomposedLogData)
```

Figure 4.2.4 Making the Data Stationary

```python
decomposedLogData = residual
decomposedLogData.dropna(inplace=True)
test_stationarity(decomposedLogData)
```




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.318799
p-value                          0.014059
#Lags Used                      14.000000
No. of Observations Used       167.000000
Critical Value (1%)             -3.470126
Critical Value (5%)             -2.879008
Critical Value (10%)            -2.576083
dtype: float64
```

Figure 4.2.5 Stationary Data

The data transformation has been done and it is visible that the mean is constant now compared to previous data. It has also been verified using the Augmented Dickey-Fuller Test. This tells us if the data is stationary or not. The p-value of the test is 0.014 on a significance of level of 0.05, i.e., the data is stationary, and the analysis can proceed.

4.3 MODELING FOR TIME SERIES ANALYSIS

As mentioned previously, ARIMA model will be used for the Time Series Analysis. ARIMA model forecasts results based on its previous values and there are three parameters in this model.

1. p: This stands for Auto Regressive (AR). It refers to when past values are used to predict the future values
2. d: This stands for Integration. It takes in the amount of differencing that is used for the TSA
3. q: This stands for Moving Average (MA). It is the average that is calculated when different intervals are taken

The p, d, and q values are chosen depending on the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) graphs. For stationary data, the values must be between the upper and lower dotted lines.

```
from statsmodels.tsa.stattools import acf,pacf
lag_acf = acf(datasetLogDiffShifting, nlags=20)
lag_pacf = pacf(datasetLogDiffShifting, nlags=20,method='ols')

plt.subplot(121)
plt.plot(lag_acf)
plt.axhline(y=0,linestyle='--',color='gray')
plt.axhline(y=-1.96/np.sqrt(len(datasetLogDiffShifting)),linestyle='--',color='gray')
plt.axhline(y=1.96/np.sqrt(len(datasetLogDiffShifting)),linestyle='--',color='gray')
plt.title('Autocorrelation function')

plt.subplot(122)
plt.plot(lag_pacf)
plt.axhline(y=0,linestyle='--',color='gray')
plt.axhline(y=-1.96/np.sqrt(len(datasetLogDiffShifting)),linestyle='--',color='gray')
plt.axhline(y=1.96/np.sqrt(len(datasetLogDiffShifting)),linestyle='--',color='gray')
plt.title('Partial Autocorrelation function')
plt.tight_layout()
```

Figure 4.3.1 Code for ACF and PACF Graphs

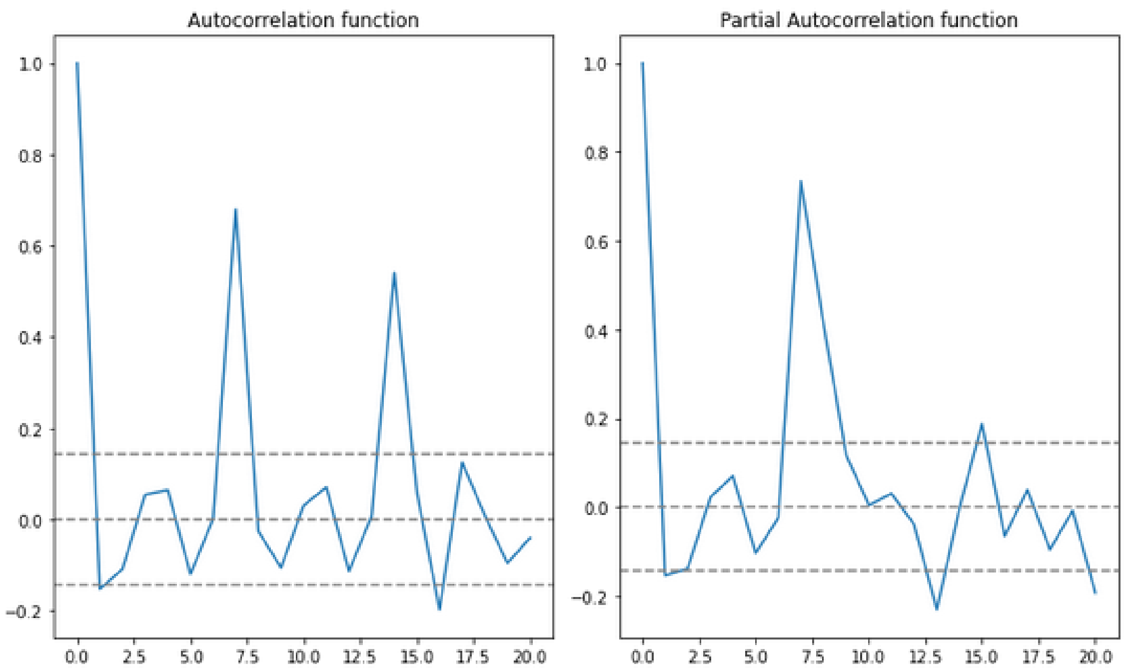


Figure 4.3.2 ACF and PACF Graphs

As we can see, most of the points are between the upper and lower dotted lines. The p value is selected from the PACF graph and the q value is selected from the ACF graph. The point where the line cuts the 0 point on the X-axis is the value we will chose. In both graphs, that point is 2. So, the values for this model are p = 2, d = 1, q = 2. This is the order of the model (2, 1, 2).

```
from statsmodels.tsa.arima_model import ARIMA

model = ARIMA(indexedDataset_logScale, order=(2,1,2))
results_AR = model.fit(disp=-1)
plt.plot(datasetLogDiffShifting)
plt.plot(results_AR.fittedvalues, color='red')
plt.title('RSS: %.4f'%sum((results_AR.fittedvalues-datasetLogDiffShifting['Daily Confirmed'])**2))
print('Plotting AR model')
```

Figure 4.3.3 Plotting ARIMA Model

CHAPTER 5
RESULTS AND ANALYSIS

In this chapter, the visualizations and time series analysis results that were obtained have been attached and their analysis has been discussed.

As mentioned previously, all visualizations have been done in Tableau and the time series analysis was done using Python.

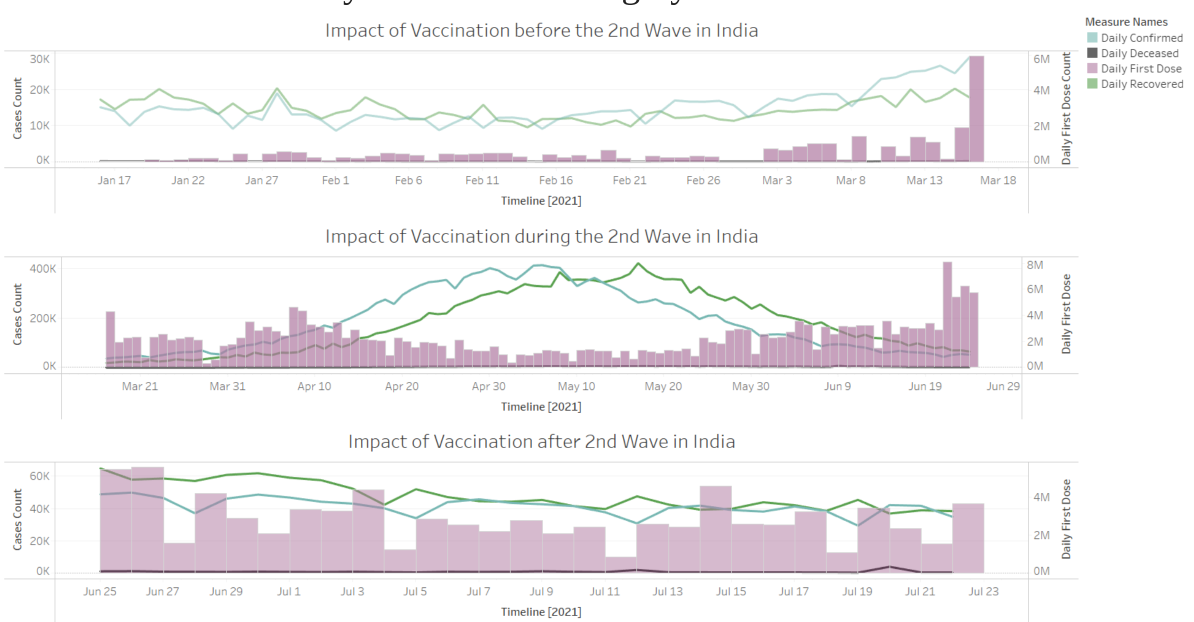


Figure 5.1 Impact on India

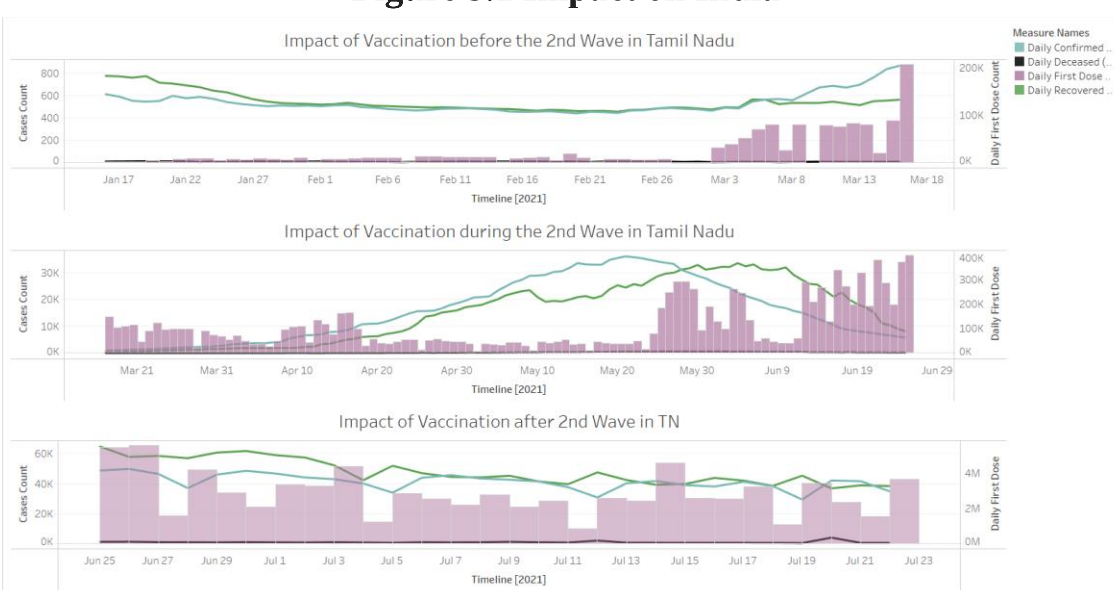


Figure 5.2 Impact on Tamil Nadu

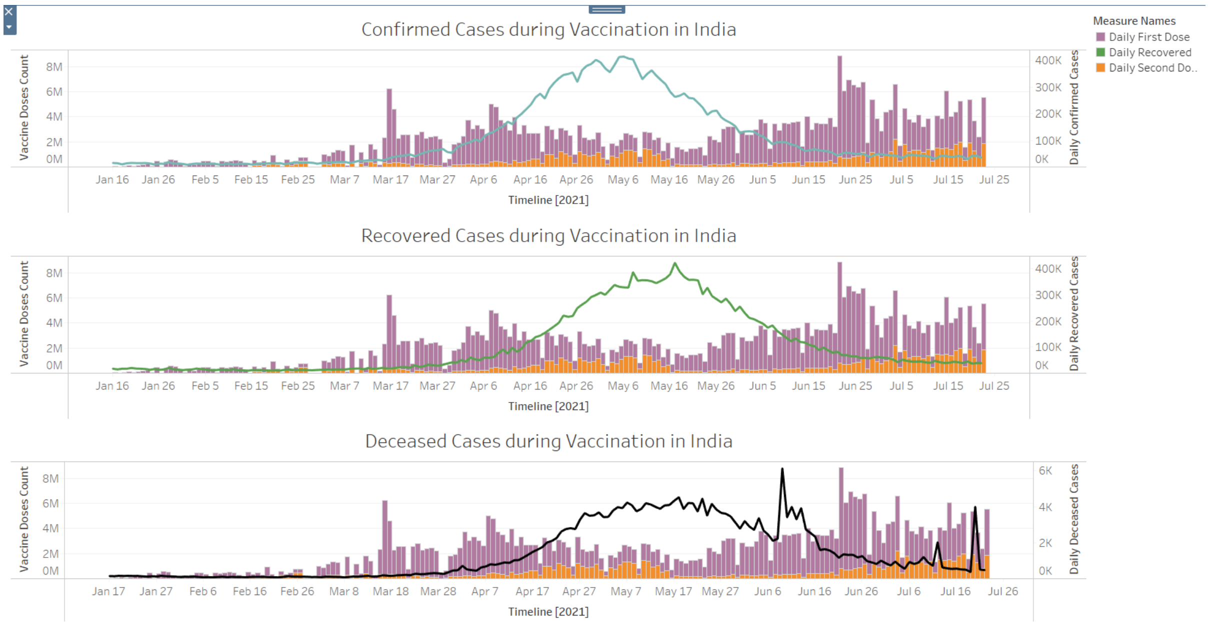


Figure 5.3 Impact of Cases on Vaccination Drive in India

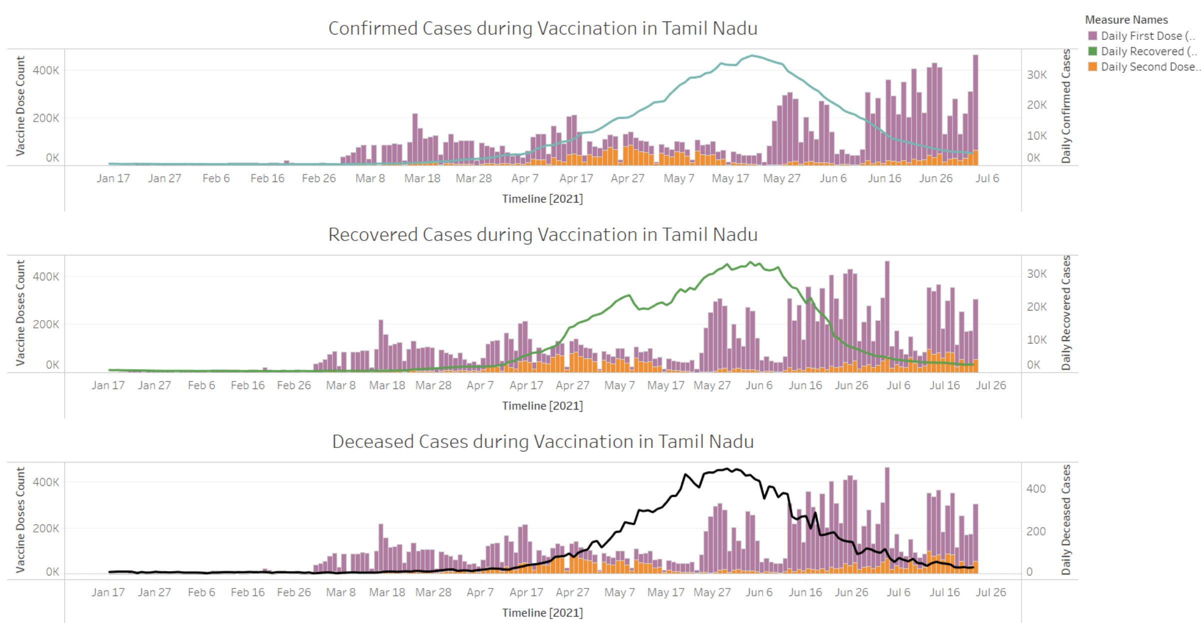


Figure 5.4 Impact of Cases on Vaccination Drive in Tamil Nadu

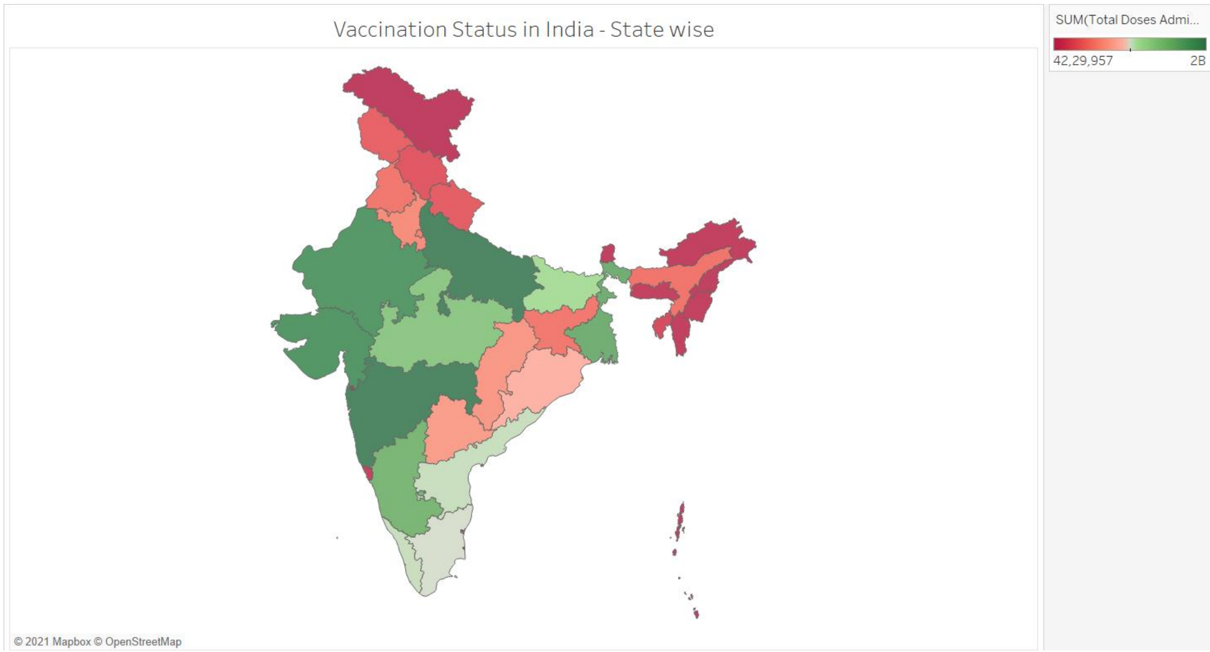


Figure 5.5 Status of Vaccination Drive in India-State Wise

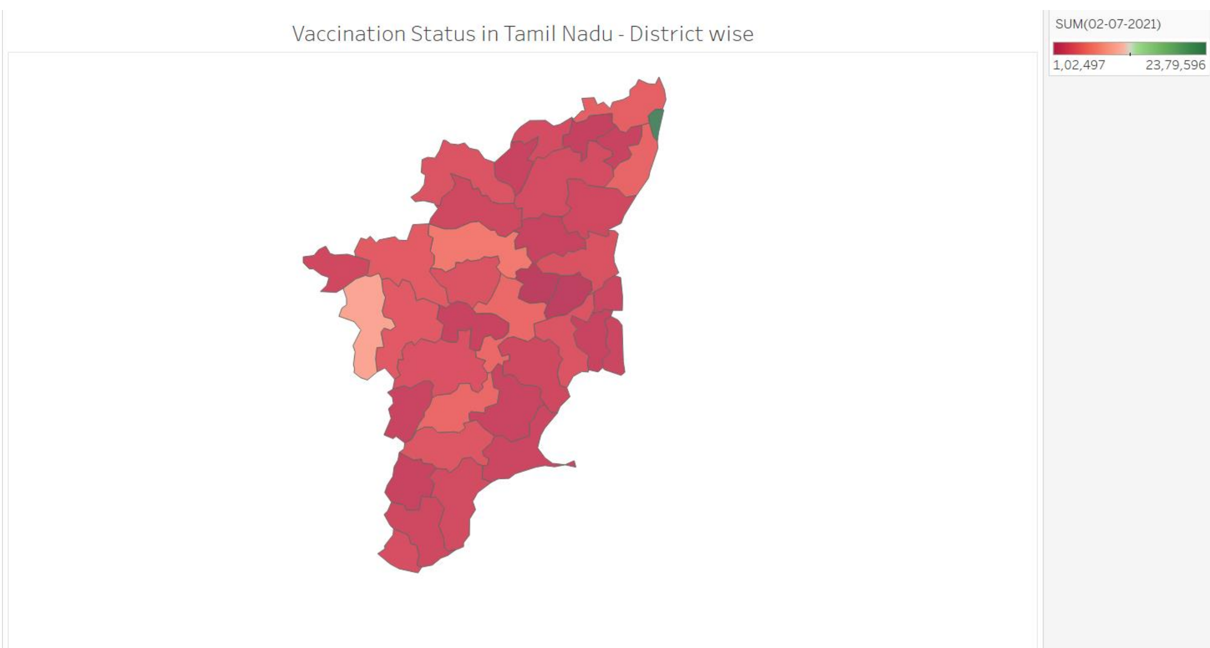


Figure 5.6 Status of Vaccination Drive in India-District Wise

Figures 5.1 and 5.2 show the total number of vaccinations that were given along with the curves for confirmed, recovered, and deceased cases. This was made possible by using a dual axis graph. Comparing both the graphs, we can see that Tamil Nadu peaked almost 2 weeks later than the rest of India. The number of vaccinations also decreased significantly as people were hesitant to go out during the rising cases and the lockdowns were also strict in some places. We can also see that the total lockdowns on Sundays in Tamil Nadu also majorly affected the drive. These days are visible by extremely low values in the graph.

Figures 5.3 and 5.4 show the impact of the increasing cases on the vaccination drive. There are 3 separate graphs for confirmed, recovered, and deceased cases respectively. It also shows the type of dosage. First dose is in purple and second dose is in orange. We can see a sudden spike in the number of deceased cases in India, but this is just backdated data being added at a later date. The cases start dropping and the vaccinations pick up pace as people realize the importance of the vaccine and since many people had received one dose, rate of future infections started decreasing.

Figures 5.5 and 5.6 compare state wise vaccination status, the most affected and populous states have the highest number of vaccinated people, i.e., Maharashtra, followed by Uttar Pradesh. This had been done to combat the exponentially increasing cases and bring them under control. Same trend can be found in districts of Tamil Nadu, where most affected districts had more vaccinated people i.e., Chennai and the Coimbatore

```
1 results_ARIMA.plot_predict(1,248)
2 #results_ARIMA.forecast(steps=120)
```

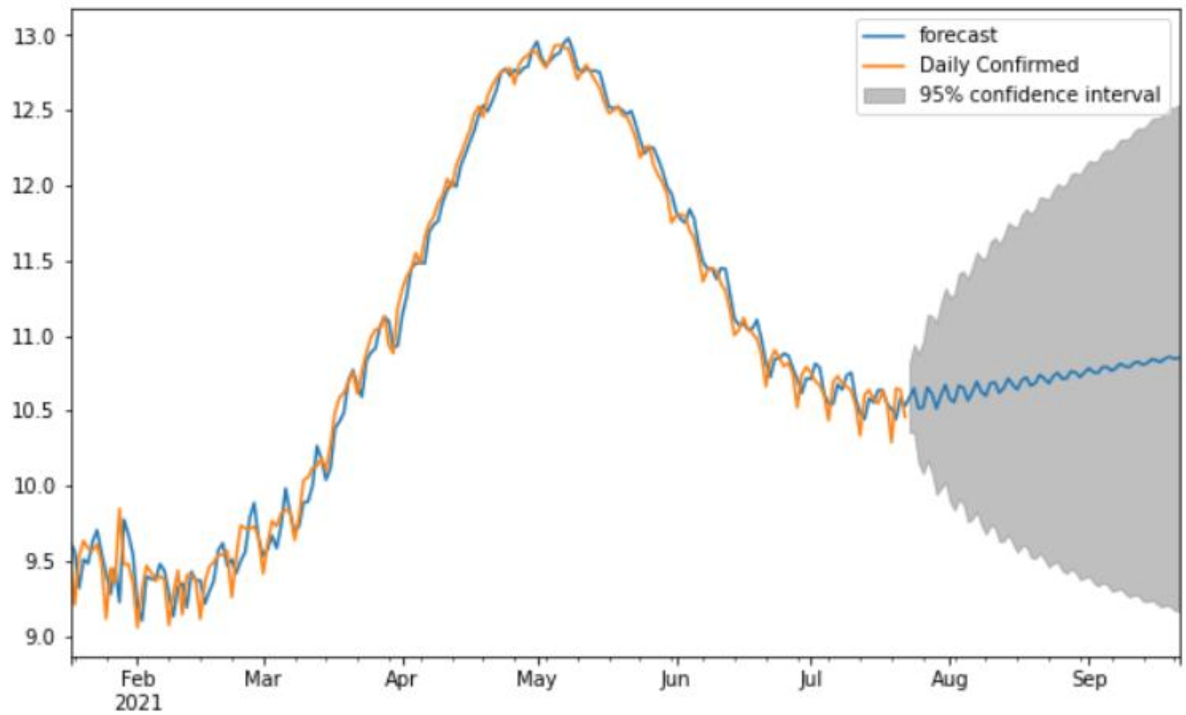


Figure 5.7 Forecasted Number of Cases

This figure shows the forecasted number of cases for the upcoming 2 months. With a confidence interval of 95%, we can safely account for both best- and worst-case scenarios in India. It has been predicted that the number of daily cases can continue decreasing in the best-case scenario. The orange line represents the data from the dataset and the blue line is the forecasted number of cases through ARIMA model. This will be made possible by maintaining the current rate of vaccination and it can be bettered by constantly increasing the rate to give immunity to the majority of the population.

CHAPTER 6
CONCLUSION

The vaccination drive was severely affected during the 2nd wave in India since many states had imposed lockdowns of varying strictness and the people were not able to get vaccinated like before. From the visualizations, it can be verified that the vaccination drive helped in bringing down the number of cases during the middle of the 2nd wave since the previously vaccinated people had more protection against the virus and had less chances of contracting the virus. Even if they did contract the virus, vaccines help in reducing the severity of the disease and greatly reduce the number of fatalities.

With time series analysis being carried out, it has been forecasted that number of daily cases will not undergo any drastic increase. We can also predict that future waves can be effectively handled, and with the worst-case scenario where there is no increase in rate of vaccination, possible future peaks will be lower than that of the 2nd wave in India. This is only possible because of vaccinations and if the current trend continues, there are good chances of the cases remaining low provided that new COVID-19 variants do not cause a surge in cases. We can expect the best-case scenario when more vaccinations are carried out. This will lead to higher chances of cases continuing to decrease and even sudden waves can be handled well.

CHAPTER 7
REFERENCES

[1] Dey, Samrat K., et al. "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach." Journal of medical virology 92.6 (2020): 632-638

[2] Comba, Joao LD. "Data visualization for the understanding of COVID-19." Computing in Science & Engineering 22.6 (2020): 81-86.

[3] Jentner, Wolfgang, and Daniel A. Keim. "Visualization and visual analytic techniques for patterns." High-Utility Pattern Mining (2019): 303-337

[4] Le Bras, Pierre, et al. "Visualising covid-19 research." arXiv preprint arXiv:2005.06380 1 (2020)

[5] Maurya, Sujeet, and Shikha Singh. "Time Series Analysis of the Covid-19 Datasets." 2020 IEEE International Conference for Innovation in Technology (INOCON). IEEE, 2020

[6] Gecili, Emrah, Assem Ziady, and Rhonda D. Szczesniak. "Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy." Plos one 16.1 (2021): e0244173

[7] Bhangu, Kamalpreet Singh, Jasminder Sandhu, and Luxmi Sapra. "Time series analysis of COVID-19 cases." World Journal of Engineering (2021)