

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology in Information Technology

By

VASANTH P
REG. NO. 39120132



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU

APRIL 2022



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
(Established under Section 3 of UGC Act, 1956)
JEPPIAAR NAGAR, RAJIV GANDHI SALAI
CHENNAI- 600119
www.sathyabama.ac.in



DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of VASANTH P
(Reg. No: 39120132) who carried out the project entitled "**ANALYSIS OF
EPIDEMIC USING DATA SCIENCE**" under my supervision from June 2022 to
APRIL 2022.

Internal Guide

Name of the Guide

DR.V.NIRMAL RANI ,M.E,Ph.D

Head of the Department

DR.R.SUBHASHINI, M.E,Ph.D

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **VASANTH P** hereby declare that the project report entitled “**ANALYSIS OF EPIDEMIC USING DATA SCIENCE**” done by me under the guidance of **Dr. V.NIRMAL RANI, M.E,Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in **INFORMATION TECHNOLOGY**.

DATE:25.04.2022

PLACE:CHENNAI

A handwritten signature in black ink, appearing to read 'P. Vasanth', with a stylized flourish at the end.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean**, School of Computing, **Dr. R. Subhashini, M.E., Ph.D. Head of the Department** of **Information Technology** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.V.NIRMAL RANI M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE



icmr **NIE**
INDIAN COUNCIL OF
MEDICAL RESEARCH | NATIONAL INSTITUTE OF
EPIDEMIOLOGY

ICMR - NATIONAL INSTITUTE OF EPIDEMIOLOGY
Chennai, Tamil Nadu, India



TO WHOMEVER IT MAY CONCERN

THIS IS TO CERTIFY THAT **VASANTH P** IS CURRENTLY PURSUING INTERNSHIP AT
ICMR-NIE AS ON 25.04.2022.

A handwritten signature in blue ink, appearing to be "V. P.", with a long horizontal stroke extending to the right.

HEAD

Computing and Information Science

ABSTRACT

FAST Application is Facial Attendance System Technology Application which is an automated attendance system. This system is a prototype model for AI based Attendance System. This system provides a solution using IoT, Microsoft Cognitive Services and JavaScript Frameworks like Angular, Nodejs, Johnny-Five.

Automation is a necessity in the current times as it makes processes more economical and affordable in the long run. It also frees humans from performing banal tasks day in and day out. Once a process is automated the only check that is to be performed is whether it is turned on or not. Automated processes are not prone to errors and even if an error is identified rectification is easy and can be applied system wide without any delay.

Taking attendance is a mundane but necessary task in many academic institutions. In this paper, an alternate and more efficient method of taking attendance is proposed that uses facial recognition and cloud based IoT

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	i
	LIST OF FIGURES	ii
	LIST OF TABLES	iii
	LIST OF ABBREVIATIONS	iv
1	INTRODUCTION	10

	1.1 ICMR-NIE PROJECT	10
	1.2 MODERN TECHNOLOGY & EPIDEMIC	12
2.	AIM AND SCOPE OF THE PROJECT	16
	2.1 ANALYSE THE EPIDEMIC	16
	2.2 DEVELOP MEASURES	17
	2.3 EVALUATE	17
3.	EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED	18
	3.1 EPIDEMIC DATA	18
	3.2 DATA MODIFICATION	20
	3.3 BIG DATA TOOLS-TABLEAU	27
4.	RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS	9
	4.1 ANALYTICAL RESULTS	9
	4.2 RECOMMENDATIONS	9
5	SUMMARY AND CONCLUSIONS	28
	REFERENCES	35
	APPENDIX	37
	A.SCREENSHOTS	14

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
1.1	data set I	21
1.2	data set II	22
2.1	analysis chart I	31
3.1	analysis chart II	32

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO.
-----------	------------	----------

1. INTRODUCTION

1.1 ICMR-NIE PROJECT-

This is ICMR-ICMR-NATIONAL INSTITUTE OF EPIDEMIOLOGY project . the aim of the project is to use modern technology to analyze the spread of the epidemic /pandemic.

IN this project data science, big data tools are used to effectively analyse the mode of spread of the pandemic. Modern data science offers epidemiologists a marked advantage in the single most important factor in tracking the spread of an outbreak: speed. [Every passing day that data sits in storage without being processed leads to the loss of life](#), as was described in the Harvard School of Public Health Magazine.

However, it's not just the speed with which multiple data sources can be integrated and analyzed with data science techniques. It's also the accessibility of the resulting analyses. Only rarely do public health officials perform their most critical work in the safe and cloistered confines of offices, with powerful computers and high-speed Internet access at their fingertips.

Traditionally, public health workers have been able to work only retrospectively, acting on information that comes in after crises are already underway. It takes time to draw blood, get lab results, and compile and interpret the information. Epidemiologists are sometimes days behind the curve when trying to assess the spread of disease out in the field.

But data scientists are changing all that. By uncovering unexpected relationships in disparate sets of data, they are uncovering correlations that allow public health officials to work more proactively

In the future, data scientists might not only look at such inadvertent self-reporting, but also to other information posted publicly to social media or other sources that could simply describe risk factors rather than actual infection. During the 2014 Ebola outbreak in Africa, the International Telecommunication Union [rapidly developed methods to track movements and Call Data Records in the outbreak area](#) to assist

epidemiologists in developing spread models to anticipate where the disease would head next..

Thus this project intends to analyse the epidemic using data science

1.2.

MODERN TECHNOLOGY AND EPIDEMIC-

Technologic advances during this century have led to unparalleled improvements in comfort, productivity, and life span. In the field of medicine, changes in technology have provided immeasurable benefits to society. Catastrophic and epidemic diseases have been prevented and/or controlled by improvements in sanitation and hygiene and by introduction and widespread use of vaccines and antimicrobial and other drugs. Improvements in intensive care, surgical techniques, cancer therapy, and therapy for a variety of chronic diseases (like diabetes mellitus and heart disease) have led to prolonged survival and better quality of life for millions of people. The impact of technology on the practice of medicine is among the most salutary changes that has occurred during the twentieth century. While some of the most spectacular successes of technology have been in the area of prevention and treatment of infectious diseases, there have been negative effects (often indirect) as well. Societal changes resulting from mass urbanization are clearly linked to the age of technology. urbanization are clearly linked to the age of technology. These events have led to the widespread use of illicit intravenous drugs, which in turn has contributed to the spread of serious blood. borne pathogens like human immunodeficiency virus and hepatitis viruses, and to increased incidence of subacute bacterial endocarditis. The same changes have led to increasing numbers of persons who, without permanent homes, are often squeezed into crowded and unclean living spaces facilitating transmission of tuberculosis and diarrheal and respiratory infections. Accelerating technologic advancement is an ineluctable part of our balance of life, and, up to now at least, the benefits appear to outweigh the negatives. The focus of this presentation is to consider ways that technology has had a direct effect on the potential for causing epidemic infectious diseases and has contributed to the emergence of some diseases by providing recommendations to **THE GOVT OF INDIA** by **ICMR-NIE**.

During an outbreak, clear messaging to the local population is vital to ensure they are informed and reminded to use appropriate precautions, getting clear information to the point most required. Thanks to Africa's technological [push for mobile phones](#), several NGOs have been able to

spread the message highlighting the hazards of Ebola and how to avoid contracting the virus.

In a recent project, Text To Change was able to work in conjunction with Airtel, sending more than 1 million text messages to inform the Sierra Leonean population of ways to prevent the spread of Ebola.

Programmes such as this are able to utilize the available technology to quickly reach the population with up-to-date information and, more importantly, quickly update it as the hazards change. It is through the use of these tools that we are able to reach a disparate population to make them aware of the issues and treatments for some of these diseases. It is important to note that, although technology and mobile phones are becoming more prevalent in Africa, the widespread use of smartphones is limited. The TTC and other programmes focus on using SMS texts to spread the message.

Training

Technological advances in mobile computing are [used to train health workers in the field](#). Tools such as the Oppia browser and e-buddi from the MMEI have been able to provide training directly to health workers on the ground. The benefit of tools such as this is that they can quickly deliver standardized learning, with the ability to be updated in the field as new procedures are validated. This replaces potentially conflicting training that is not widely disseminated or tested. Additionally, by using established communications links, training material can be tailored and tested with the health workers on site.

By using these tools, we are able to focus efforts on training the existing staff, health workers and local population to treat themselves long after the emergency support has ended. This will also have an appropriate [knock-on effect on the general medical infrastructure](#) in the region. At present, the penetration rate for mobile devices is much lower in the rural areas most in need of such advice and treatment, but this will change.

Connectivity

Much has been written about the growing connectivity capabilities within Africa and the developing world. Improved connectivity facilitates a number of technological advances and is a key enabler of the technical

capabilities discussed in this article. There have been significant advances in this area, but access in rural locations is still limited. From our work in the region, this is the key tool that can bring both economic and health benefits.

During an epidemic, communication is a critical tool to ensure we can both spot the virus spreading early and make attempts to contain it by educating the local population. From an NGO perspective, it also provides a means of maintaining medical services through the use of telemedicine. In the recent outbreak, a number of facilities [were effectively shut down](#) as staff were evacuated, creating a secondary and potentially more damaging issue, as basic healthcare was no longer available.

Anything we can do now to ensure a stable communications network in region will have a significant impact in a future epidemic. The legacy these solutions provide to the beneficiary hospitals is also significant. As evidence that such capabilities [can be provided in harsh environments](#), recently VSEE provided telemedicine access to refugees in Syrian refugee camps. By utilizing this technology *in situ*, remote areas will have the capability to continue providing health services following an outbreak.

Traditional means of providing bandwidth to remote areas using satellite technology are being augmented by Wi-Fi and 3G or 2.5G technologies. This is not the traditional solution in the Western world but, due to the expense of satellite-based solutions, these technologies are economically viable. Hardware solutions such as the BRCK have been developed and deployed in Africa specifically to provide mobile internet capabilities in situations with limited power supplies. This device provides internet access when the power is out, utilizing any remaining 3G communications links.

Real time monitoring

[Several innovative solutions](#) such as Epi-info VHF and EVDNet have been trialled to assist in the monitoring of any new virus outbreak. This is an area that has seen real innovation in using the available mobile capabilities in the region. The viability of these approaches will only grow with the increase in mobile penetration.

Innovations that are being trialled include examining data from mobile phone towers to track users who have been close to a known case of the virus. Other initiatives are being trialled that allow staff to electronically enter patients' details as they are received at a medical facility, [thus helping to provide a clear picture of the outbreak itself](#). These technologies leverage the communication capabilities others are trialling to allow clinics to be linked together and tracked centrally. Data from these tools are now being analysed [with at least one hackathon being aimed at using the data](#).

These trials highlight the potential usage of technology, and the results are starting to show promise. With the next epidemic potentially just around the corner, it is imperative to continue pursuing these technical evaluations to ascertain which tools can be effectively used in the next outbreak. Technology will not stop the outbreak alone but, as has been shown, it can be an effective force multiplier that can be leveraged to aid in the detection and treatment of the next epidemic. We must continue which approaches are the most successful if we are to avert widespread casualties in the future.

2. AIM AND SCOPE OF THE PROJECT

2.1 ANALYSIS OF EPIDEMIC

A large number of real world networks exhibit community structure, and different communities may often possess heterogeneity. In this paper, considering the heterogeneity among communities, we construct a new community network model in which the communities show significant differences in average degree. Based on this heterogeneous community network, we propose a novel mathematical epidemic model for each community and study the epidemic dynamics in this network model. We find that the location of the initial infection node only affects the spreading velocity and barely influences the epidemic prevalence. And the epidemic threshold of entire network decreases with the increase of heterogeneity among communities. Moreover, the epidemic prevalence increases with the increase of heterogeneity around the epidemic threshold, while the converse situation holds when the infection rate is much greater than the epidemic threshold.

The mathematical models have become important tools in analyzing the spread and control of infectious diseases. The development of such models is aimed at both understand observed epidemiological patterns and predicting the consequences of the introduction of public health interventions to control the spread of diseases. Some diseases not confer immunity against the disease but other diseases confer immunity so recovered individuals gain immunity against disease. These types of disease can be modifications by SI and SIS where S susceptible and I infective respectively. Both epidemic models (SI and SIS) are one of the most basic and most important models in describing of many diseases. Therefore, it attracted many authors attention and a number of papers have been published. For example Gao and Hethcote [1] considered an SIS model with a standard disease incidence and densitydependent demographics. Li and Ma [2] studied an SIS model with vaccination and temporary immunity. Kermack and McKendrick [3] proposed a simple SIS model with infective immigrants. In recent years, many papers found treatment function for example, Li et al [4] proposed the SIS model with a limited resource for treatment. Shurowq k. Shafeeq [5] studied the effect of treatment, immigrants and vaccinated on the dynamic of SIS epidemic model. In this paper we proposed and studied a mathematical model consisting of SII S* epidemic model with treatment, in which it is

assumed that the disease transmitted by contact as well as external sources in the environment. The local as well as global stability analysis.

2.2 DEVELOP MEASURES-

Modelling of disease transmission via compartmental models is well established and generally highly effective. However, the differential equations of these models depend on good estimates of underlying rate parameters and will then provide a continuous solution under the assumption that the population is well-mixed and homogeneous (i.e. all individuals have equal contact with all others). Under these assumptions disease propagation is driven by the parameter R_0 —the ratio of the rate of new infections to the rate of removal of infectious individuals from the transmission pool. Typically, and particularly for contemporary and evolving transmission, these parameters can be somewhat difficult to estimate

The emergence of an earlier coronavirus, associated with the Severe Acute Respiratory Syndrome (SARS) in 2003, provided an opportunity to apply the structures and concepts of complex systems to the modelling of infectious diseases. Small and Tse [14] introduced a complex network based model of propagation and showed good agreement between simulations of that model and available case data. They found that epidemic parameters widely quoted in the literature were only consistent with observed case data when including significant nosocomial transmission [15]. Finally, and most importantly for the current discussion, the scale-free topology of the model [14] explained super-spreader events through contact rather than requiring pathologically highly infectious individuals [16].

Both the network-based models used to model SARS in 2003 [14]–[15][16], and the model we describe here are network models of contact between individuals. Unlike what we will propose in this current communication, the model of SARS in 2003 was topologically stationary [14]–[15][16]. The model assumed a lattice with long-range (i.e. small-world) connections following a power-law degree distribution. In those papers [14]–[15][16] time varying control strategies were reflected only in changes of the rate parameters. The current coronavirus outbreak (that is, COVID-19) poses a different and unique challenge. Since February 2020 (and up to the time of writing) global transport networks and daily movement of individuals have been disrupted on a global scale. Entire cities and countries have engaged in various levels of “lockdown”. We argue that it is

neither appropriate nor sufficient to model this simply by modifying the rate of transmission or rate of removal. In the current work we propose a network switching model through which the topology of the network changes to reflect various changes in government and community mitigation and control strategies.

3. EXPERIMENTAL PROCEDURE-

3.1 EPIDEMIC DATA-

DATA Is very important for the every aspect of the research.the icmr - national institute of epidemiology does test, surveys all over india collects the data of the disease which is predominantly spread across the country.

- Provides a systematic approach for dissecting a health problem into its component parts.
- Ensures that you are fully versed in the basic dimensions of a health problem.
- Identifies populations at increased risk for the health problem under investigation.
- Provides timely information for decision-makers, the media, the public, and others about ongoing investigations.
- Supports decisions for initiating or modifying control and prevention measures.
- Measures the progress of control and prevention programs.
- Enables generation of testable hypotheses regarding the etiology, exposure mode, control measure effectiveness, and other aspects of the health problem.
- Helps validate the eventual incrimination of causes or risk factors.

For the project a total of 1,27,998 data sets of epidemic disease is collected.in .sql file from the ICMR-NATIONAL INSTITUTE OF EPIDEMIOLOGY.

The .sql file is studied using db forgesql then file is converted into .xls file.

FileHomeInsertDrawPage LayoutFormulasDataReviewViewHelp

Tell me what you want to do

NIE-FILE [Protected View] - Excel

VASANTH PANDURANGAN

PROTECTED VIEWBe careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View.

Enable Editing

A2

virus_id

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																		
2	virus_id	pat_info	test_no	virus_name	virus_type	virus_test	virus_sample	virus_test	virus_other	virus_resu	virus_stat	virus_date	iden_id	old_id	pati_id	cen_id	lab_code	
3	4	4	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	12-01-2015	4	0	9	1	1	
4	5	4	2	Dengue	#####	Blood - Serum	Other Test			2	2 Y	12-01-2015	4	0	9	1	1	
5	31	20	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	20	0	28	1	1	
6	32	20	2	Dengue	#####	Blood - Serum	Antigen Detection (N			2	2 Y	0000-00-00	20	0	28	1	1	
7	33	21	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	21	0	29	1	1	
8	34	22	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	22	0	30	1	1	
9	36	24	1	Chikungunya	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	24	0	33	1	1	
10	37	25	1	Measles	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	25	0	20	1	1	
11	38	25	2	Not in the FTD Brain	#####	CSF	RTPCR			2	2 Y	0000-00-00	25	0	20	1	1	
12	39	26	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	26	0	37	1	1	
13	40	26	2	Dengue	#####	Blood - Serum	Antigen Detection (N			2	2 Y	0000-00-00	26	0	37	1	1	
14	41	27	1	Chikungunya	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	27	0	38	1	1	
15	42	28	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	28	0	39	1	1	
16	44	30	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	26-10-2018	30	0	47	1	1	
17	45	30	2	Dengue	#####	Blood - Serum	Antigen Detection (N			1	2 Y	26-10-2018	30	0	47	1	1	
18	46	31	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	31	0	50	1	1	
19	47	31	2	Dengue	#####	Blood - Serum	Antigen Detection (N			2	2 Y	0000-00-00	31	0	50	1	1	
20	48	32	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	32	0	55	1	1	
21	49	32	2	Dengue	#####	Blood - Serum	Antigen Detection (N			2	2 Y	0000-00-00	32	0	55	1	1	
22	50	33	1	Herpes simplex virus	#####	Others	vesicle scr PCR			1	2 Y	0000-00-00	33	0	355	1	1	
23	52	35	1	Dengue	#####	Blood - Serum	IgM			2	2 Y	0000-00-00	35	0	352	1	1	
24	53	36	1	Herpes simplex virus	#####	Others	vesicle scr PCR			2	2 Y	0000-00-00	36	0	32	1	1	

Sheet1

Average: 68656.23053Count: 125615Sum: 8624183742

Ready

32°CMostly sunny

ENG IN

29-04-2022

Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing

VASANTH PANDURANGAN

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
127966	131535	67451		4 Sapovirus		##### Stool			PCR			2	2 Y						
127967	131536	67452		1 Enterovirus		##### Stool			PCR			2	2 Y	18-06-2015	67451	0	1408	2	
127968	131537	67452		2 Norovirus		##### Stool			PCR			2	2 Y	0000-00-00	67452	0	1409	2	
127969	131538	67452		3 Astrovirus		##### Stool			PCR			2	2 Y	18-06-2015	67452	0	1409	2	
127970	131539	67452		4 Sapovirus		##### Stool			PCR			2	2 Y	18-06-2015	67452	0	1409	2	
127971	131540	67453		1 Dengue		##### Blood - Serum			IgM			2	2 Y	18-06-2015	67453	0	1409	2	
127972	131541	67453		2 Chikungunya		##### Blood - Serum			IgM			3	2 Y	18-06-2015	67453	0	1409	2	
127973	131542	67453		3 Dengue		##### Blood - Serum			IgM			3	2 Y	18-06-2015	67453	0	1410	2	
127974	131543	67454		1 Dengue		##### Blood - Serum			IgM			1	2 Y	18-06-2015	67453	0	1410	2	
127975	131544	67455		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67453	0	1410	2	
127976	131545	67456		1 Dengue		##### Blood - Serum			IgM			3	2 Y	0000-00-00	67454	0	1411	2	
127977	131546	67457		1 Dengue		##### Blood - Serum			IgM			1	2 Y	0000-00-00	67455	0	1412	2	
127978	131547	67458		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67456	0	1413	2	
127979	131548	67458		2 Dengue		##### Blood - Serum			IgM			1	2 Y	0000-00-00	67457	0	1414	2	
127980	131549	67459		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67458	0	1415	2	
127981	131550	67460		1 Dengue		##### Blood - Serum			IgM			2	2 Y	19-06-2015	67458	0	1415	2	
127982	131551	67461		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67459	0	1416	2	
127983	131552	67462		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67460	0	1417	2	
127984	131553	67463		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67461	0	1418	2	
127985	131554	67464		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67462	0	1419	2	
127986	131555	67465		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67463	0	1420	2	
127987	131556	67466		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67464	0	1421	2	
127988	131557	67467		1 Dengue		##### Blood - Serum			IgM			2	2 Y	0000-00-00	67465	0	1422	2	
127989						##### Blood - Serum			IgM			2	2 Y	0000-00-00	67466	0	1423	2	
												2	2 Y	0000-00-00	67467	0	1424	2	

Sheet1

Ready

32°C Mostly sunny

ENG IN 09:04 29-04-2022

3.2 DATA MODIFICATION-

What is data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If

data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn't mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA

At the end of the data cleaning process, you should be able to validate the data

Components of quality data

Determining the quality of data requires an examination of its characteristics, then weighing those characteristics according to what is most important to your organization and the application(s) for which they will be used.

5 characteristics of quality data

1. **Validity.** The degree to which your data conforms to defined business rules or constraints.
2. **Accuracy.** Ensure your data is close to the true values.
3. **Completeness.** The degree to which all required data is known.
4. **Consistency.** Ensure your data is consistent within the same dataset and/or across multiple data sets.
5. **Uniformity.** The degree to which the data is specified using the same unit of measure.

Benefits of data cleaning

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

- Removal of errors when multiple sources of data are at play.
- Fewer errors make for happier clients and less-frustrated employees.
- Ability to map the different functions and what your data is intended to do.

- Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
- Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

ORGANISING EPIDEMIC DATA-n

Organizing descriptive data into tables, graphs, diagrams, maps, or charts provides a rapid, objective, and coherent grasp of the data. Whether the tables or graphs help the investigator understand the data or explain the data in a report or to an audience, their organization should quickly reveal the principal patterns and the exceptions to those patterns. Tables, graphs, maps, and charts all have four elements in common: a title, data, footnotes, and text ([Box 6.2](#)). In this chapter, additional guidelines for preparing these data displays will appear where the specific data display type is first applied.

Tables are commonly used for characterizing disease cases or other health events and are ideal for displaying numeric values. In addition to the previously mentioned elements in common to all data displays, tables have column and row headings that identify the data type and any units of measurement that apply to all data in that column or row. A well-structured analytical table that is organized to focus on comparisons will help you understand the data and explain the data to others. In arranging analytical tables, you should begin with the arrangement of the data space by following a simple set of guidelines

Cases are customarily organized in a table called a *line-listing* . This arrangement facilitates sorting to reorganize cases by relevant characteristics. The line-listing in has been sorted by days between vaccination and onset to reveal the pattern of this important time–event association. Commonly in descriptive epidemiology, you organize cases by frequency of clinical findings . If the disease cause is unknown, this arrangement can assist the epidemiologist in developing hypotheses regarding possible exposures. For example, initial respiratory symptoms might indicate exposure through the upper airways

Components of Statistical Data Displays

A statistical data display should include, at a minimum,

- A title that includes the what, where, and when that identifies the data it introduces.
- A data space where the data are organized and displayed to indicate patterns.
- Footnotes that explain any abbreviations used, the data sources, units of measurement, and other necessary details or data.
- Text that highlights the main patterns of the data (this text might appear within the table or graphic or in the body of the report).

Guidelines for Graphical Data Presentation

- Take care in selecting a graph type in computer graphics programs. In Microsoft Excel (Microsoft Corporation, Redmond, WA), for example, you should use “scatter,” not “line” to produce numerically scaled line graphs.
- Adhere to mathematical principles in plotting data and scaling axes.
- On an arithmetic scale, represent equal numerical units with equal distances on an axis.

- When using transformed data (e.g., logarithmic, normalized, or ranked), represent equal units of the transformed data with equal distances on the axis.
- Represent dependent variables on the vertical scale and independent variables on the horizontal scale.
- Use alternatives to joining data points with a line. Consider instead
 - No line at all (use data markers only).
 - A trend line of best fit underlying the data markers.
 - A moving average line underlying the data markers.
- Aspect ratios (data space width to height) of approximately 2:1 work well. Extreme aspect ratios distort data.
- Scale the graph to fill the data space and to improve resolution. If this means that you must exclude the zero level, exclude it, but note for the reader that this has been done.
- Do not insist on a zero level unless it is an integral feature of the data (e.g., an endpoint).
- Use graphic designs that reveal the data from the broad overview to the fine detail.
- To compare two lines, plot their difference directly.
- Use visually prominent symbols to plot and emphasize the data.
- Make sure overlapping plotting symbols are distinguishable.
- When two or more data sets are plotted in the same data space
 - Design point markers and lines for visual discrimination; and
 - Differentiate them with labels, legends, or keys.
- To avoid clutter and maintain undistorted comparisons, consider using two or more separate panels for different strata on the same graph.
- When comparing two graphs of the same dependent variable, use scaling that improves comparison and resolution.
- Clearly indicate scale divisions and scaling units.

- Minimize frames, gridlines, and tick marks (6–10/axis is sufficient) to avoid interference with the data.
- Use six or fewer tick mark labels on the axes. More than that becomes confusing clutter.
- Keep keys, legends, markers, and other annotations out of the data space. Instead, put them just outside the data region.
- Proofread your graphs.

Contact Diagrams

Contact diagrams are versatile tools for revealing relationships between individual cases in time. In contact diagrams ([Figure 6.2](#), panel A) ([5](#)), which are commonly used for visualizing person-to-person transmission, different markers are used to indicate the different groups exposed or at risk.

Epidemic Curves

Epidemic curves ([Box 6.5](#)) are histograms of frequency distributions of incident cases of disease or other health events displayed by time intervals. Epidemic curves often have patterns that reveal likely transmission modes. The following sections describe certain kinds of epidemic situations that can be diagnosed by plotting cases on epidemic curves.

3.3 BIG DATA TOOL – TABLEAU

Tableau empowers people throughout the organization to answer questions of their data, large or small, in real-time. The more questions they ask, the more value they extract from the data, leading to smarter business decision every day. Tableau allows you to have one interface for all of your data, regardless of where that data is stored.

No matter what kind of data consumption or storage software you're using, from various Hadoop distributions to NoSQL

databases to Spark, Tableau integrates seamlessly with your infrastructure. And while Big Data can be inherently messy and complex, the ecosystem around it is quickly evolving to enable you to structure that data for easy exploration or accelerate the performance of the NoSQL and Hadoop databases so they feel much more like the traditional databases

Big Data platforms enable you to collect, store and manage more data than ever before. Previously unseen patterns emerge when we combine and cross-examine very large data sets. These patterns contain critical business insights that allow for the optimization of business processes that cross department lines

Tableau reduces the need to pull individual reports from multiple software or databases. See it all in one place and discover new patterns you might never have seen using separate individual reports. Furthermore, explore this data through the eyes of your entire organization, allowing for more questions to be asked and more discoveries to be made.

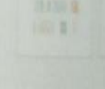
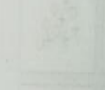
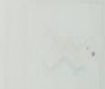
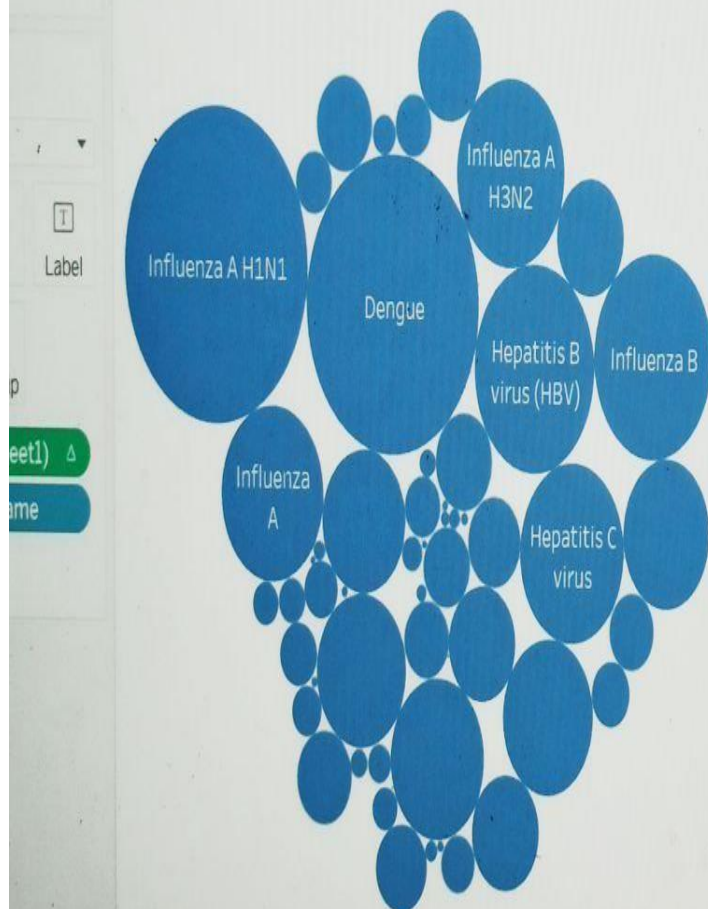
The converted .xls file is imported to tableau and attributes are selected to get the chart for the wide understanding of diseases.

4.RESULTS, ANALYSIS AND DISCUSSION-

4.1- ANALYTICAL RESULTS-

- <https://public.tableau.com/app/profile/vasanth.p1001/viz/ICMR-NIE-SURVEY/Dashboard1>

ICMR-NIE- SURVEY



For horizontal

0 or more

1 or more

SURVEY

Virus Name

Null	0.04%
---- Select ----	0.06%
Adeno Virus	1.03%
Astrovirus	0.60%
BK Virus	0.00%
Blood - Plasma"	0.02%
Blood - Serum"	1.49%
Chandipura virus..	0.03%
Chikungunya	1.30%
Coronavirus	0.13%
Coxsackievirus	0.01%
Crimean Congo v..	0.00%
CSF	0.20%
Cytomegalovirus..	3.18%
Dengue	13.55%
Enterovirus	2.07%
Epstein-Barr-vir..	0.95%
Hantavirus	0.05%
Hepatitis A virus..	3.70%
Hepatitis B virus..	6.68%
Hepatitis C virus..	5.08%
Hepatitis D virus..	0.00%
Hepatitis E virus..	3.85%
Herpes simplex v..	0.00%
Herpes simplex v..	4.11%
Herpes simplex v..	2.13%

Sheet 1 Dashboard 1 Sheet 2 Sheet 3

column SUM of % of Total CNT (Sheet1): 100.00%



Following components are considered for analysis of data to give recommendation.

Recognizing disease patterns by personal attributes (e.g., age, sex, education, income, or immunization status) constitutes the fifth element in descriptive epidemiology. Two important qualifications apply to person data assessments. First, determining rates is more often necessary than for time and place. Second, age is a strong independent determinant for many causes of morbidity and mortality.

Social Groupings and Personal Contact

Social groupings might be as compact as a household or as diffuse as a social network linked by a common interest. The underlying epidemiologic process might produce disease distributions within and among social groupings that range from strong aggregation to randomness or uniformity. Clustered distributions might result from common exposures of group members, an agent that is transmissible through personal contact, an environmental exposure in the living or meeting areas, or localization of houses near or within an environmental area of high risk. Random or uniform distributions indicate that the exposure lies outside the group.

For diseases or behaviors spread through personal contact or association, contact diagrams reveal the pattern of spread plus such key details as index cases and outliers. In the example diagram, closeness and quality of relationships, timing between onsets, and places of contact are all displayed through different symbols and shading ([Figure 6.2](#)) (5). To support person-to-person transmission, you should also see that the timing between onsets of cases approximates the known incubation periods for the disease ([Figure 6.8](#)) (5).

Age

In most descriptive analyses, the epidemiologist will determine disease rates by age. This can be as simple as finding that a health event is affecting only a limited age group or as complicated as comparing age-specific rates among multiple groups. Age represents three different categories of determinants of disease risk ([Box 6.11](#)). Because age is a pervasive determinant of disease and because **population groups often differ in their age structures, age adjustment** (standardization) is a useful tool for comparing rates between population groups ([17](#)). Age-adjusted rates can be used for comparing populations from different areas, from the same area at different times, and among other characteristics

4.2 RECOMMENDATIONS-

Recommendations for early recognition and source control Early recognition of ARIs and application of source control, including respiratory hygiene, are administrative control measures aimed at reducing or preventing the dissemination of infectious agents from the source. The early identification, isolation and reporting of ARIs of potential concern are therefore central to effective containment and treatment.

- Use clinical triage for early identification of patients with ARIs to prevent the transmission of ARI pathogens to health-care workers and other patients (Strong recommendation, very low to low quality of evidence) (27, 51) (Annex K, Table K.1). Regularly monitor and evaluate the clinical triage system to ensure effectiveness (52- 55).
- Place ARI patients in an area separate from other patients, and evaluate clinical and epidemiological aspects of the case as soon as possible (51, 52, 56). Complement investigation with laboratory evaluation if applicable (57, 58).
- In people with ARIs, encourage the use of respiratory hygiene (i.e. covering the mouth and nose during coughing or sneezing with a medical mask [surgical or procedure mask], cloth mask, tissue, sleeve or flexed elbow), followed by hand hygiene, to reduce the dispersal of respiratory secretions containing potentially infectious particles (Strong recommendation, very low quality of evidence) (27, 51, 59-63) (Annex K, Table K.2).
- Implement additional IPC precautions promptly according to the suspected pathogen (Table 2.1) (64).
- Report all available essential information regarding episodes of ARIs of potential concern to public health authorities via the local surveillance system. This is in line with the requirements of the IHR (2005) (6),

which have been in force since June 2007. The IHR (2005) require the international notification to WHO by States Parties of events that may constitute a public health emergency of international concern. Public health authorities • Establish channels to inform health-care facilities and the community about ongoing epidemic ARIs, so that the facilities will be aware of the extent and types of problems likely to be encountered.

Effective IPC programmes can reduce the frequency and financial burden of health-care associated infections (76-78). The 10-year long SENIC (Study on the Efficacy of Nosocomial Infection Control) study in the United States of America showed that organized IPC programmes are both effective and cost effective (77). Currently, IPC programmes are considered an integral part of the delivery of patient care¹ (79, 80). In addition to the recommendations for early recognition and source control described in Section 2.1, the following administrative control strategies for IPC programmes in health-care facilities outlined below are recommended. For all ARIs • Strengthen or establish an IPC committee and IPC programmes with trained personnel to keep policies current (52, 53, 69-75, 79, 81, 82). • Monitor and increase compliance with IPC precautions using evidence-based methods, including multimodal strategies (e.g. change in infrastructure, education, posters, reminders, senior management engagement and performance feedback) (83-85). • Educate health-care workers about ARIs, including the IPC precautions to be used for patients who present with a febrile ARI (55, 86, 87). • Ensure that adequate IPC supplies are provided (55, 87-89), for example: – hand-hygiene facilities (e.g. soap and clean running water, alcohol-based hand rub, and paper or single-use towels); – PPE for patient care (e.g. masks, respirators, gowns, gloves and eye protection); – PPE for heavy duties (e.g. closed protective footwear, waterproof aprons and rubber gloves); and – an adequate supply of appropriate materials for cleaning and disinfection.

IPC precautions are measures designed to minimize the risk of transmission of infections. Such precautions are typically separated into Standard Precautions and additional precautions, such as Contact, Droplet and Airborne Precautions. Annex B summarizes the application and principles of Standard and additional precautions in health care. Additional precautions may be needed depending on: • the suspected or confirmed causative agents of the ARIs (53, 65, 67-69, 94); • the presence of epidemiological and clinical clues suggesting that patients have ARIs of potential concern; and • the types of contact and procedures that are undertaken with patients with ARIs

Surveillance -

- As a priority, establish within the health-care facility processes for the early recognition and investigation of possible pandemic ARI patients (57, 58).
- Connect the hospital and public-health infectious diseases surveillance systems, and immediately report any essential information about possible pandemic ARI cases to public health authorities. The reporting should occur through the local surveillance system, as per Annex 1 of the IHR (2005) (6).
- Public-health authorities should keep health-care facilities informed about ongoing epidemics.
- In the case of pandemic influenza: – enhance ILI surveillance (Annex D) (185, 191); – define criteria that would shift surveillance of episodes of influenza of potential concern (e.g. human cases of avian influenza) from passive to active.

SUMMARY:

A survey of both non communicable and communicable diseases has been conducted with the National Institute of Epidemiology (NIE), ICMR, which included 54 various diseases, and more than 127800 data has given us a clear statement that the most commonly infected diseases are Influenza A H1N1 (Swine flu), followed by Dengue. As a data scientist, we have clumped the data into an easily accessible one using Structural Query Language (SQL) to communicate with the data interpreted and we organised them using Tableau. Using Tableau, it has become an easier task to find out the percentage effect of the even rare diseases in the survey like Hepatitis D Virus, Herpes Simplex Virus, Crimean Congo Virus, and many more.

CONCLUSION:

By organising the data using SQL and Tableau, our has become more accessible to public and hopefully an awareness on public hygiene would be created. The current work has opened the gate for future data scientists to get involved with other biological scientists and modernise the world of epidemiology and medicine with Data sciences combined with medicine and providing a better knowledge to the public using Tableau

REFERENCE-

1. Ehrenberg AC. The problem of numeracy. *Am Stat.* 1981;35:67–71.
2. Centers for Disease Control and Prevention. Intussusception among recipients of rotavirus vaccine—United States, 1998–1999. *MMWR.* 1999;48:577–81.
3. Hawley B, Casey ML, Cox-Ganser JM, Edwards N, Fedan KB, Cummings KJ. Notes from the field: respiratory symptoms and skin irritation among hospital workers using a new disinfection product—Pennsylvania, 2015. *MMWR.* 2016;65:400–1.
4. Centers for Disease Control and Prevention. Injuries and deaths associated with use of snowmobiles—Maine, 1991–1996. *MMWR.* 1997;46:1–4.
5. Xie S, Zeng G, Lei J, et al. A highly efficient transmission of SARS among extended family and hospital staff in Beijing, China, April 2003. Presented at the 2nd Southeast Asian and Western Pacific Bi-

- Regional TEPHINET Scientific Conference, November 24–28, 2003, Borocay, Philippines.
6. Tauxe RV, Tormey MP, Mascola L, Hargrett-Bean NT, Blake PA. Salmonellosis outbreak on transatlantic flights; foodborne illness on aircraft: 1947–1984. *Am J Epidemiol*. 1987;125:150–7.
 7. Abu-Sbeih A, Fontaine RE. Secondary water-borne hepatitis E outbreak from water storage in a Jordanian town [abstract LB01]. Presented at the TEPHINET 2000 First International Conference, April 17–21, 2000, Ottawa, Canada.
 8. Dou F, Sun H, Wang ZJ. Dengue outbreak at a fishing port: Guangdong Province, China, 2007. Presented at the International Conference on Emerging Infectious Diseases, March 16–19, 2008, Atlanta, Georgia.
 9. Ko AI, Galvao Reis M, Ribeiro Dourado CM, Johnson WD, Jr., Riley LW. Urban epidemic of severe leptospirosis in Brazil. Salvador Leptospirosis Study Group. *Lancet*. 1999;354:820–5.
 10. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System. Death rates for selected causes by 10-year age groups, race, and sex: death registration states, 1900–32, and United States, 1933–98. <https://www.cdc.gov/nchs/nvss/mortality/hist290.htm>
 11. National Institutes of Health, National Cancer Institute, Surveillance Epidemiology, and End Results Program. SEER*Explorer. <http://seer.cancer.gov/explorer/>
 12. Fontaine RE, van Severin M, Houg A. The stratification of malaria in El Salvador using available malaria surveillance data [Abstract 184]. Presented at the XI International Congress for Tropical Medicine and Malaria, September 16–22, 1984, Calgary, Alberta, Canada.
 13. Tobler W. A computer movie simulating urban growth in the Detroit region. *Econ Geogr*. 1970;46(Suppl):234–40.
 14. Vazquez-Prokopec GM, Kitron U, Montgomery B, Horne P, Ritchie SA. Quantifying the spatial dimension of dengue virus epidemic spread within a tropical urban environment. *PLoS Negl Trop Dis*. 2010;4:e920.
 15. Biggs HM, Behravesh CB, Bradley KK, et al. Diagnosis and management of tickborne rickettsial diseases: Rocky Mountain

spotted fever and other spotted fever group rickettsioses, ehrlichioses, and anaplasmosis—United States. *MMWR Recomm Rep*. 2016;65(No. RR-2):1–44.

- 16.Registrar-General. *Report on the mortality of cholera in England 1848-49*. London, UK: Her Majesty's Stationery Office; 1852.
- 17.Fleiss JC. *Statistical methods for rates and proportions*. New York:John Wiley & Sons; 1987