**RESEARCH ARTICLE**

Analyzing and Processing the data of COVID-19 disease

**Submitted By:**

Priyam Gupta 19BCS089

Anany Talwad 19BEC004

Kshirsagar Uddesh Shyam 19BEC021

Mane Param Dattatraya 19BEC022

Prashant Sharma 19BEC034

Sanda Nihar Purshottam 19BEC038

Satyam Kumar 19BEC040

**Under the guidance of:**

Dr. Uma Sheshadri Mam

Professor,

Computer Science & Engineering

IIIT Dharwad



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

DHARWAD

**INTRODUCTION:**

The novel coronavirus SARS-CoV-2 was first described in Wuhan, China, in January 2020, as the causative agent of COVID-19. Three coronaviruses, including SARS-CoV-2, are known to have crossed species barriers to infect humans. Besides SARS-CoV-2, SARS- CoV and MERS-CoV are pathogenic in humans, can cause severe illness and death, and have the potential for epidemic and pandemic spread. The World Health Organisation (WHO) has classified these three viruses as priority pathogens to accelerate the development of vaccines and therapeutics to prevent epidemics. SARS-CoV-2 is a beta-coronavirus and shows genetic similarity with SARS-CoV. SARS- CoV-2 is highly infectious and is spreading fast, currently at pandemic proportions. The infection presents with different symptoms and a wide spectrum of severity. Some patients do not have fever, cough, or radiologic abnormalities on initial presentation. Some patients only experience fever and cough, while others display the very severe form of the disease that leads to bilateral pneumonia, which has to be treated in intensive care units, and sometimes leads to death. A better understanding of the immunogenicity and pathobiology of SARS-CoV-2 infections in humans is urgently needed as a basis for the development of diagnostics, therapeutics, and vaccines against SARS-CoV-2. High-density peptide arrays enable the rapid identification of antigen epitopes, recognized by antibodies for many applications. Pathogen-specific peptide arrays help to identify biomarkers for (early) detection of diseases,7 to develop therapeutics, or to rationally design vaccines. Identification of epitopes that are targeted by protective antibodies is critical to develop monoclonal antibodies as therapeutics. Using high-density peptide arrays in combination with clinical and immunological data provides insight into immunogenicity induced by SARSCoV-2 with a focus on antibody development during infection. Screening patient-specific IgA, IgG, and IgM responses to the full viral proteome, mapped as overlapping linear peptides on an array, is fast and provides data to identify new diagnostic markers, and to shed light on unknown immunological

interactions. Here, we report longitudinal B-cell population and virus-specific antibody

response data from two SARS-CoV-2-positive patients: a married couple. While the husband had a more severe course of disease and was hospitalized (no ventilation), his wife began to experience only mild symptoms six days after her husband and was subsequently diagnosed. In addition to identifying biomarkers for disease progression, we discovered important epitopes for vaccine development.

World Health Organization (WHO) has declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. The COVID-19 infectious disease is fast spreading across the countries, impacting the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level across the GLOBE. As a Student community, it is a concern for the community as many of our students travel across the states and countries for their project work, and internship work in future. The result of this RESEARCH cum Data TRENDS study is therefore crucial. The final outcome is aimed towards understanding the infectious disease data pattern in the GLOBAL population such that the impact of COVID-19 has minimum effect on our students Project/Internship selection planning process. The 2019 novel coronavirus (2019-nCoV), officially named as COVID-19 pandemic by the WHO, has spread to more than 180 countries including China. Confirmed novel coronavirus cases increased ten-fold in less than a month, from 100,000 in the first week of March to more than one million on 02 April and today we can see that more than 2.5 million confirmed cases in the world, while more than 1,80,000 deaths have been reported across the world. Europe has become the new epicentre of coronavirus. More than 97.5% of the global COVID-19 cases are currently outside China. In this research work we used various data structure concepts to understand the distribution of data in COVID-19 Cases and performs various data processing operations. Using the concepts, we classify the countries according to HIGH RISK LEVEL and LOW RISK LEVEL for upcoming two years also we formed the list of top HIGH-RISK LEVEL countries according to the particular age group and plotted the corresponding graphs. It is also important to identify that which country or state take how may days for confirmed case turned into death and turned into recovery case. One of the advantages of this research is that we will give prior information about COVID-19 disease country wise and state wise to the students such that they will prepare according to the situation exist in that particular country or state. Through this research we will alert the people who falls in HIGH RISK LEVEL age group by providing the data set of that particular age group. As the data set is growing dynamically, we can able to track the latest information about confirmed cases, recovered cases, death rate, average number of days for confirmed case to become a death case or recovered case, age group wise distribution of confirmed cases and recovered cases of COVID-19 disease.

**DATA FILES:**

**About Time Series** - A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Time series analysis is use in order to understand the underlying structure and function that produce the observations. Understanding the mechanisms of a time series allows a model to be developed that explains the data in such a way that prediction, monitoring, or control can occur. Time series are used in Statistics, Signal Processing, Pattern Recognition, Econometrics, Mathematical Finance, Weather Forecasting, Earthquake Prediction, Control Engineering, Astronomy, Communication Engineering and largely in any domain of Applied Science and engineering which involves temporal measurements. Here we used time series to analyse the COVID-19 confirmed cases, recovered case and death rate also for the analysing the age wise group data on specified time interval as the data is changing dynamically so time series helps us to catch up this data on different dates and time.

**Following Data Files are used in the Research Study:**

  

 

**Libraries, Data Types and Algorithm of the Programs** –

**Libraries and Modules Used:**

1. **Kivy** - Kivy is a free and open source Python library for developing mobile apps and other multitouch application software with a natural user interface. It is distributed under the terms of the MIT License, and can run on Android, iOS, GNU/Linux, OS X, and Windows. For the research the package includes App, Label, GridLayout, Text Input, SoundLoader, runTouchApp, Button, Popup, Config, Window.
2. **Time** – Time module handles the time related tasks, to use functions which are defined in this module we have to first import this module in our code.
3. **pandas** - In computer programming, pandas are a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. From pandas we imported the ExcelWriter and ExcelFile.
4. **pprint** - Data pretty printer. The pprint module provides a capability to “pretty-print” arbitrary Python data structures in a form which can be used as input to the interpreter. If the formatted structures include objects which are not fundamental Python types, the representation may not be loadable.
5. **openpyxl** - openpyxl is a Python library to read/write Excel 2010 xlsx/xlsm/xltx/xltm files.
6. **matplotlib** - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
7. **pathlib** - From this library we imported path. Pathlib module in Python provides various classes representing file system paths with semantics appropriate for different operating systems.
8. **easygui** - EasyGUI is a module for very simple, very easy GUI programming in Python. EasyGUI provides an easy-to-use interface for simple GUI interaction with a user. From this we imported choicebox. Choicebox provides a way for a user to select from a list of choices. The choices are specified in a sequence (a tuple or a list).
9. **Heapq** - Heapq is a Python module which provides an implementation of the Min heap.
10. **random** - Generate pseudo-random numbers.
11. **xlrd** - xlrd is a module that allows Python to read data from Excel files.

**Data Types Used:**

1. **Class** - It is a user defined Data Type. Class is similar to Structures, user defined data type in c with a difference that members of class are private by default while public in structures. Further class is better as it allows to store both data and functions.
2. **Lists** - All the items in a list do not need to be of the same type. Lists are mutable, meaning, the value of elements of a list can be altered.
3. **String** - Array of characters, immutable in nature.
4. **Array** – We can store multiple items of the same type and arrays are immutable in nature. Array can be handled in python by module name array.

**Etc.**

We also used **if-else** conditional statements and **for** loop in the programs.

**Python Functions:**

A Function is the block of code which only runs when it is called. A function can return data as a result. In a python a function is defined using a *def* keyword. To call a function use the function name followed by parenthesis. You can send any data types of argument to a function (string, number, list, dictionary etc.), and it will be treated as the same data type inside the function. We used the functions in our research work to make result more efficient.

**Algorithm of Programs (IDEA):**

1. In first problem, we used two for loops, one to get into the country and other to get into the dates of confirmed or recovered or death cases. We also used df.loc() function from the pandas library to access the location we wanted.
2. In second problem, first we find the total number of confirmed cases of every state then we put them sequence wise into the one list then we created the another list and we putted the index numbers of the top five sates in that newly created list after that we increment those indexes by one because we have to compare those indexes with the serial numbers in the data file agedata.xlsx and printed the corresponding state.
3. In third problem, We have the date wise column in the data file, first we calculate the total sum of all the date columns and placed those each column sum into one list then we merge the total sum of seven columns and forms the another list which consist of week wise sums and then took the average by dividing with seven and printed the result.

**References:**

* <https://github.com/CSSEGISandData/COVID-19>
* <https://www.worldometers.info/coronavirus/country/us/>
* <https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-Fatalities?%3Aembed=yes&%3Atoolbar=no&%3Atabs=n>
* <https://www.mass.gov/doc/covid-19-dashboard-april-25-2020/download>
* <https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/>
* <https://coronavirus.illinois.gov/s/>
* <https://txdshs.maps.arcgis.com/apps/opsdashboard/index.html#/ed483ecd702b4298ab01e8b9cafc8b83>
* <https://coronavirus.maryland.gov/>
* <https://www.doh.wa.gov/Emergencies/Coronavirus>
* [https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/key-metrics/caseshttps://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/key-metrics/cases](https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/key-metrics/caseshttps:/coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/key-metrics/cases)
* <http://ldh.la.gov/Coronavirus/>
* <https://www.michigan.gov/coronavirus/>
* <https://portal.ct.gov/-/media/Coronavirus/CTDPHCOVID19summary4252020.pdf?la=en>
* <https://www.geeksforgeeks.org/>
* <https://towardsdatascience.com/>