DATA ANALYSIS ON COVID-19 DATASET

SUBMITTED BY
DEEPEKA.S(310818104018)
DIVYA.R(310818104023)
KOUSALYA.V(310818104052)

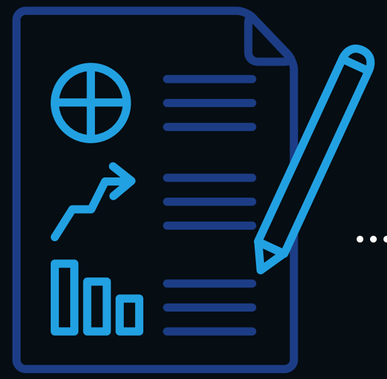
ABSTRACT

DATA ANALYSIS IS A PROCESS OF INSPECTING, CLEANSING, TRANSFORMING, AND MODELING DATA WITH THE GOAL OF DISCOVERING USEFUL INFORMATION, INFORMING CONCLUSIONS, AND SUPPORTING DECISION-MAKING. IN THIS COVID-19 DATA SETS WE HAVE ANALYSED AND MADE CERTAIN VISUALIZATIONS BASED ON OUR ANALYSIS. WE MADE THIS ANALYSIS TO ANSWER CERTAIN QUESTIONS WHICH MOST OF THE PEOPLE ASK WHAT IS THE MAIN REASON OF THIS CORONA VIRUS SPREAD. COMMON ANSWERS WERE LIKE PEOPLE DON’T FOLLOW UP SOCIAL DISTANCING AND NOT WASHING THEIR HANDS PROPERLY. WE WERE NOT SATISFIED WITH THE ANSWERS WE GOT AND WE DECIDED TO ANALYSE WHAT IS THE MAIN CAUSE OF THIS SPREAD. WE MADE CERTAIN VISUALIZATIONS WHICH CLEARLY EXPLAINED OUR QUESTIONS.

PROPOSED METHODOLOGY

WE WILL DISCUSS FEW QUESTIONS AND THE RELATED RESULTS FOR THOSE QUESTIONS AND MAKE A CLEAR CONCLUSION REGARDING THIS SPREAD OF CORONA VIRUS. THROUGHOUT THIS ANALYSIS WE CLEANED, TRANSFORMED, AND INSPECTED EVERY INFORMATION IN OUR DATA SET USING RSTUDIO CLOUD AND FINALLY WE GOT A CLEAR SOLUTION

ARCHITECTURE DIAGRAM



Data
Collection

Data cleaning and
filterning

R STUDIO

Backend

Timeline

R STUDIO

Frontend

LIST OF MODULES

- CLEANING
- FILTERING
- VISUALING
  - -TIME GRAPH
  - -BAR GRAPH

Module 1: Cleaning

IN CLEANING PROCESS WE REMOVE UNWANTED COLUMNS AND CHOOSE AN APPROPRIATE JOIN TO MERGE THE DATA AND KEEP THE 'MISSING' VALUES MINIMAL.

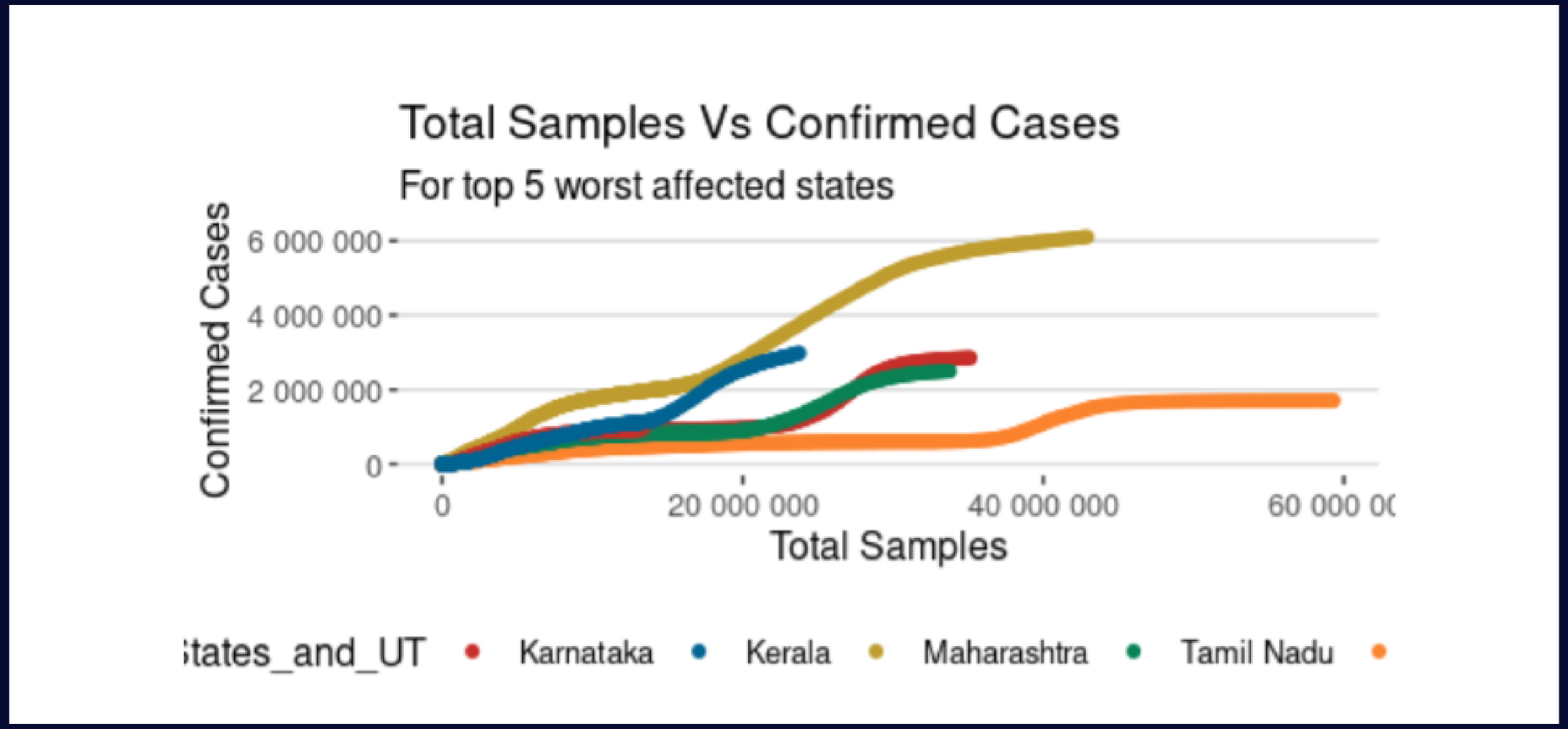
Module 2: Filtering

BY USING THE FLITERING OPTION WE CAN EASILY FILTER OUT DATA ON THE BASIS OF WHICH WE WANT AND THAT MAKE US UNDERSTAND THE DATA EASILY.

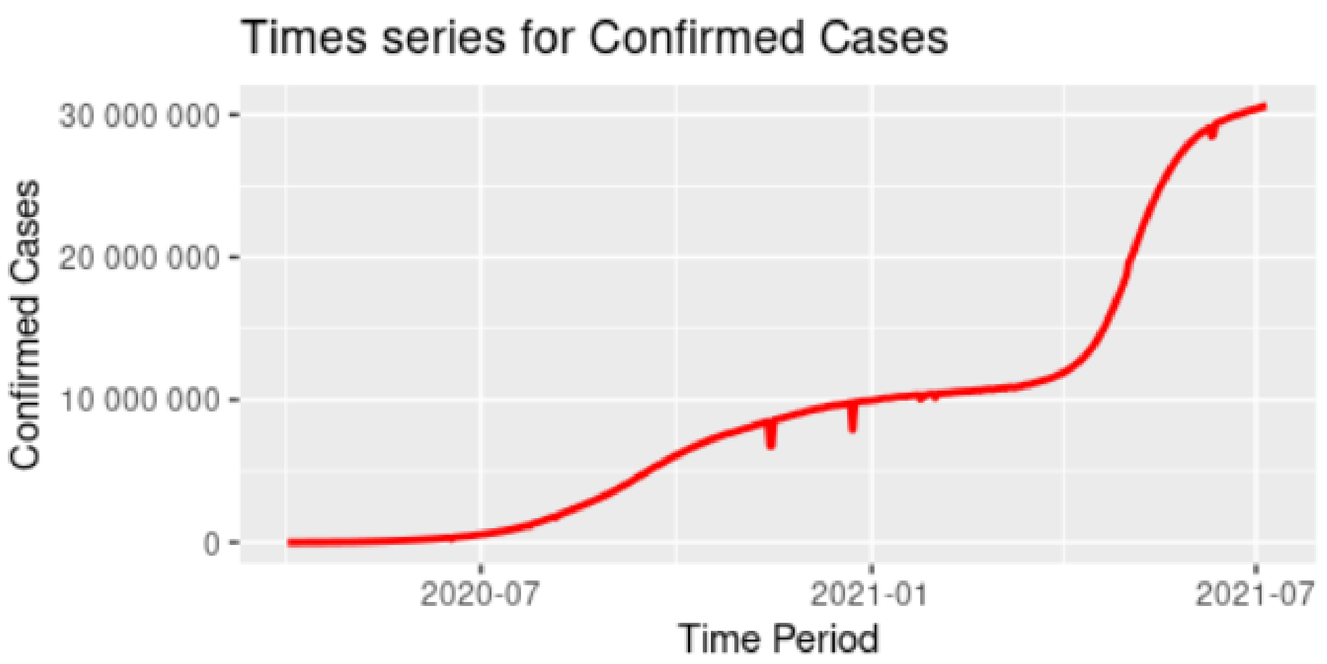
Module 3: Visualising (time graph)

Time series aim to study the evolution of one or several variables through time. A focus is made on the "tidyverse": the "lubridate" package is indeed your best friend to deal with the date format, and ggplot2 allows to plot it efficiently.

Time graph


Total Samples Vs Confirmed Cases
For top 5 worst affected states
Confirmed Cases
6 000 000
4 000 000
2 000 000
0
0
20 000 000
40 000 000
60 000 00
Total Samples
tates_and_UT
Karnataka
Kerala
Maharashtra
Tamil Nadu

Time graph


Times series for Confirmed Cases
30 000 000
20 000 000
10 000 000
0
Confirmed Cases
2020-07
2021-01
2021-07
Time Period

Module 3: Visualising (bar graph)

There are two types of bar charts: geom_bar() and geom_col() . geom_bar() makes the height of the bar proportional to the number of cases in each group (or if the weight aesthetic is supplied, the sum of the weights). If you want the heights of the bars to represent values in the data, use geom_col() instead.

Bar graph

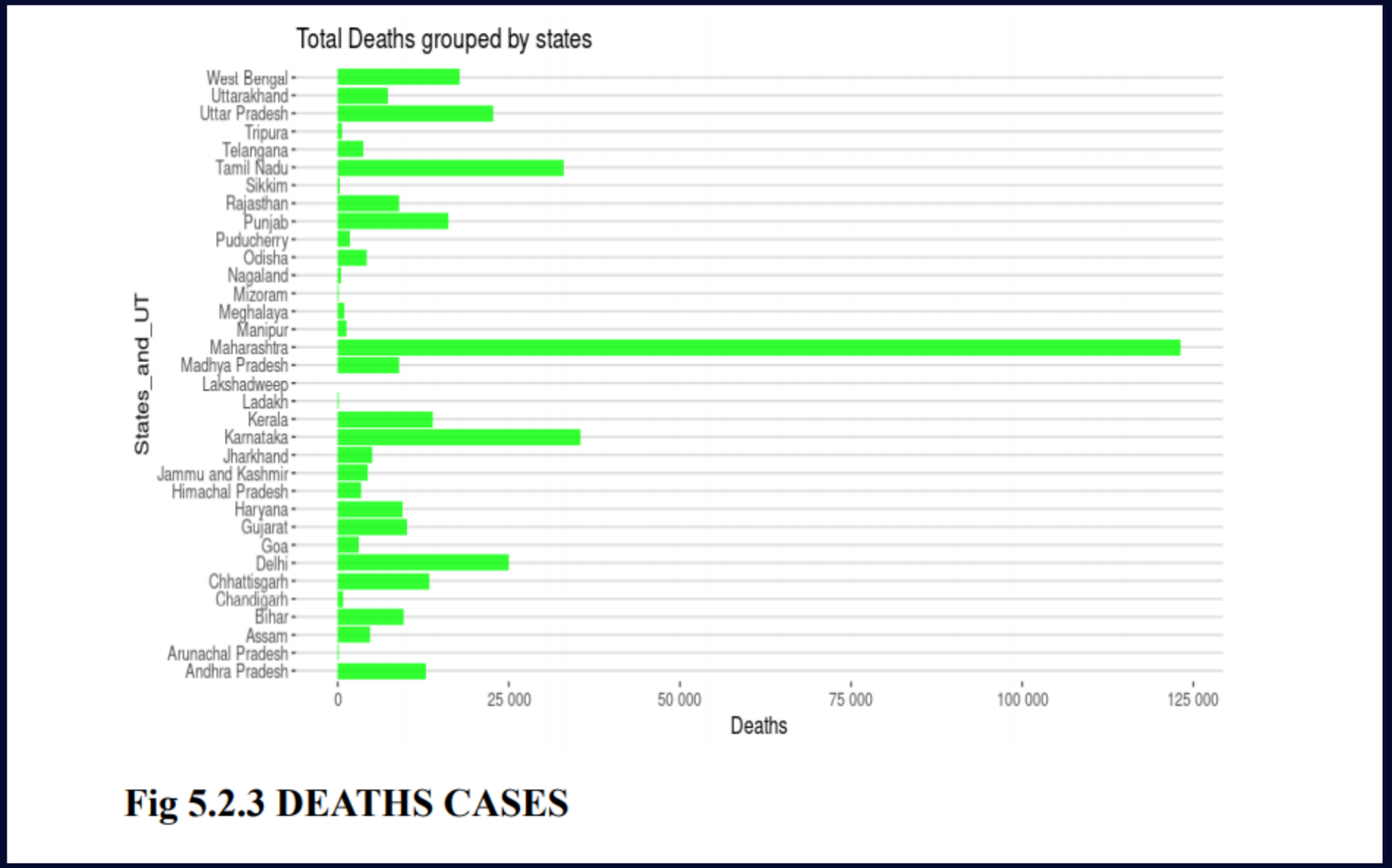


Fig 5.2.3 DEATHS CASES

Result

- INCREASE IN NUMBER OF TOTAL SAMPLES WILL BE THE RESULT OF REDUCED COVIDCASES
- COVISHIELD HAS BEEN VACCINETED
- MEN WERE VACCINATED MORE THAN WOMEN

Conclusion and Future Enhancement

IN THIS FAST GROWING TECHNOLOGY WE CAN UDERSTAND THINGS USING DATA ANALYTICS. SO COVID IS INCRESING AND BY R STUDIO WE CAN MAKE VARIOUS DATA ANALYSIS PROCESS WHICH HELP TO KNOW THE ACTUAL REASON

REFERENCE

HTTPS://WWW.KAGGLE.COM/SUDALAIRAJKUMAR/COVID19-IN-INDIA/TASKS?TASKID=631