ISSN 0974-763X

SOUTH ASIAN JOURNAL OF MANAGEMENT RESEARCH (SAJMR)

Volume 12 Number 1 **July, 2022**

Chhatrapati Shahu Institute of Business Education and Research (CSIBER)
(An Autonomous Institute)
University Road, Kolhpaur – 416 004 Maharashtra State.

SOUTH ASIAN JOURNAL OF MANAGEMENT RESEARCH (SAJMR)

ISSN 0974 – 763X

(An International Peer Reviewed Research Journal)

Published by

Chhatrapati Shahu Institute of Business Education & Research (CSIBER)

University Road, Kolhapur – 416 004, Maharashtra State

Ph: 91-231-2535706/07, Fax: 91-231-2535708, Website : www.siberindia.edu.in

Email: sajmr@siberindia.edu.in, sibersajmr@gmail.com

Time Series And Cluster Analysis Of Covid-19 Growth Curve In India

S.S. Jamsandekar, R.S. Kamath, M.B. Patil
Department of Computer Studies,
Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, Maharashtra**.**

Abstract: The study is an exploratory data analysis and cataloguing of regions from India based on the varieties of COVID-19 infected cases using unsupervised learning approach, World Health Organization's COVID-19 Situation Reports are refereed for this study. These reports provide worldwide details of the current COVID-19 epidemiological situation, present official case, death counts, and transmission classifications and so on. It is observed from the dataset that the highest number of active cases i.e. 10,17,754 on 18th September 2020. In this context,we model the trajectory of the cumulative confirmed cases and deaths of COVID-19 via a piecewise time series model. Weanalyze the trajectory of the cumulative COVID-19 cases and deaths for majorregions of India and discover interesting patterns with potentially relevant implications of the pandemic responses by different regions. We further have applied k-means clustering approach for the visualization of Indian regions. The clusters of the regions are formed based on the parameters such as Confirmed cases, Active cases, Cured cases, and Total deaths. This work exhibits the performance of the unsupervised learning technique by experimenting with several iterations and clusters. The sum of square errors is considered from each cluster to measure the performance of the k-means clustering.

Keywords: clustering; kmeans; time series; exploratory analysis;

1.0 Introduction

Artificial Intelligence and Machine Learning are playing vital role in resolving outbreak of the COVID-19. Researchers Wongvibulsinet al have designed machine learning calculator that provides the COVID-19 hospitalized patients predictions of severe illness or death [3]. A team of scientists at the University of Liverpool, UK, has used a combination of fundamental biology and machine learning for the prediction of new virus strains [4]. This finding could help to target the surveillance for new diseases.BECKER'S HEALTH IT journal has reported Artificial Intelligence tool by Los Angeles-based University of Southern California to remove uncertainty from medical diagnosis [5].

Pandemic time series forecasting strategy using Recurrent Neural Networks (RNN) for prediction of the future number of confirmed cases of COVID-19based on several factors such as transmission rate, temperature, and humidity has been proposed in [9]. Such time series analysis and visualization can give better insights of the covid 19 spread out over a tie period.

In this context, authors have presented the cluster analysis and visualization of Indian regions based on COVID-19 metrics. The dataset for the present study is retrieved from an open access data platform [1]. The cluster analysis is carried out by applying K-means clustering approach. It is observed from the dataset that the highest number of active cases i.e. 10,17,754 on 18th September 2020 [2].Hence the region wise infected cases as on 18th September was considered for the analysis. The clusters of the regions are formed based on the parameters such as Confirmed cases, Active cases, Cured cases, and Total

deaths. In this division regions are clustered into five groups of size 10,18, 3, 1 and 3.

2.0 Methods and Materials

In continuation to the scenario mentioned in the introduction, Indian regions COVID-19 metrics are taken into account to analyse similarities of the regions and assigned them to the clusters. The parameters such as Confirmed cases, Active cases, Cured cases, and Total deaths of these regions as on 18th September 2021 taken from COVID-19 Situation report. The screenshot of the partial dataset is shown in the figure 1. It has found that the infected cases were 0 for Laksdweep, so it is excluded from the dataset. The basic statistical analysis of this dataset is shown in the figure 2. It's observed that Confirmed cases, Active cases, Cured cases, and Total deaths are highest in Maharashtra. Where as the Confirmed cases and Total deaths are minimum in Mizoram.

Region	Confirmed Cases	Active Cases	Cured/Discharged	Death
Andaman and Nicobar	3604	174	3378	52
Andhra Pradesh	601462	88197	508088	5177
Arunachal Pradesh	6851	1871	4967	13
Assam	150349	28208	121613	528
Bihar	164051	13156	150040	855
Chandigarh	9256	3085	6062	109
Chhattisgarh	77775	36036	41111	628
Dadra and Nagar Have	2831	221	2608	2
Delhi	234701	31721	198103	4877
Goa	26783	5612	20844	327
Gujarat	118926	15975	99681	3270
Haryana	103773	21014	81690	1069
Himachal Pradesh	11190	4146	6946	98
Jammu and Kashmir	59711	20239	38521	951
Jharkhand	67100	13703	52807	590
Karnataka	494356	103650	383077	7629
Kerala	122214	34380	87345	489
Ladakh	3576	972	2558	46
Lakshadweep	0	0	0	0
Madhya Pradesh	97906	21631	74398	1877
Maharashtra	1145840	302135	812354	31351
Manipur	8430	1841	6538	51
Meghalaya	4356	1983	2342	31
Mizoram	1534	585	949	0
Nagaland	5306	1193	4098	15
Odisha	167161	33026	133466	669
Puducherry	21428	4744	16253	431
Punjab	90032	21568	65818	2646
Rajasthan	109473	17495	90685	1293

Figure 1: Screenshot of the partial dataset

Confirmed.Cases	Active.Cases	Cured.Discharged	Death
Min. : 1534	Min. : 174	Min. : 949	Min. : 0
1st Qu.: 8843	1st Qu.: 2534	1st Qu.: 6300	1st Qu.: 75
Median : 77775	Median : 15975	Median : 52807	Median : 590
Mean : 148991	Mean : 29079	Mean :117502	Mean : 2411
3rd Qu.: 165549	3rd Qu.: 31197	3rd Qu.:134412	3rd Qu.: 2262
Max. :1145840	Max. :302135	Max. :812354	Max. :31351

Figure 2: Basic Statistical Summary of COVID-19 as on 18th September 2021

The K-means clustering, an unsupervised classification technique is applied for the present study [3]. The K-means techniquederives a collection of k clusters using a heuristicsearch starting with a selection of k randomly chosen clustersearch of which in the beginning represents a cluster mean. Further the clustering depends on estimating likeness between regions by figuring the separation between each pair. The similarity is estimated concerning the mean estimation of the regions in a group.

3.0 K-means Clustering for COVID-19 Metrics Analysis

The region wise metrics of COVID-19 cases are analysed and clusters are formed in R platform, an opensource data mining platform. State wise COVID-19 records are clustered based on four parameters. A single record $C_i$represented as multidimensional data vector and is defined as:

C_i= [Confirmed cases$_i$, Active cases$_i$, Cured cases$_i$, and Total deaths$_i$]
Where i = 1 to 35

It is required to specify the number of clusters in K-means clustering. Figure 3 shows the plot for “Number of clusters” against "within groups total of squares". This parameter should be minimized for the selection of number of clusters [4]. Plot shows that “within groups sum of squares” decreases to a certain point 5, and afterwardthere is no much significant decrease in value. Hence it’s decided to design 5 clusters.

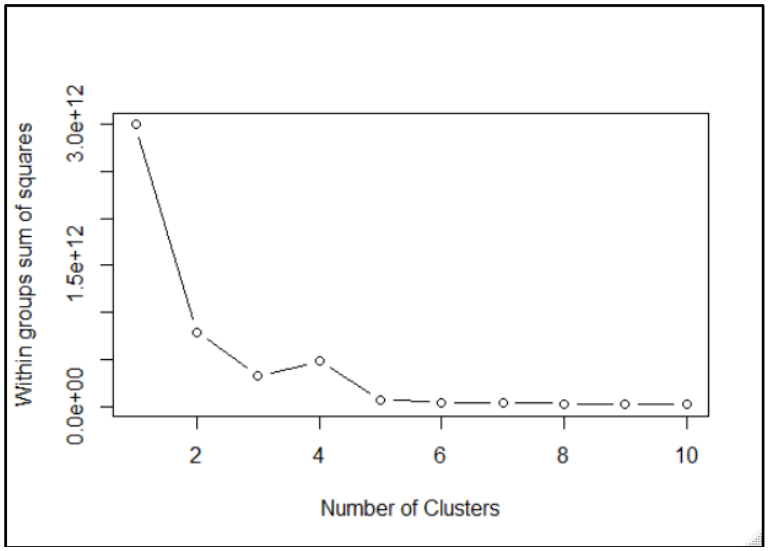

Figure 3: Number of clusters against within clusters sum of squares

The R function kmeans() is applied to get clusters of Indian regions based on COVID-19 metrics. The k-means analysis is experimented by changing number of clusters by keeping number of iterations 15, as constant. The experiment is evaluated using within clusters sum of square errors and BetweenSS/TotalSS [5].

The ideal value for BSS/TSS i.e. the properties of internal cohesion and external separation should approach 1. The summary of k-means clustering shown in the figure 4. The ratio BSS/TSS 98% shows the accuracy of the clustering. K-means forms 5clusters of size 10, 18, 3, 1 and 3 based on COVID-19 metrics.

```
K-means clustering with 5 clusters of sizes 10, 18, 3, 1, 3

Cluster means:
  Confirmed.Cases Active.Cases Cured.Discharged      Death
1       129093.10    23712.600         104009.30  1371.2000
2        20560.72     6430.222          13905.28   225.2222
3       540412.67    79485.667         453785.67  7141.3333
4      1145840.00   302135.000         812354.00 31351.0000
5       262191.67    41430.667         216150.67  4610.3333

Clustering vector:
 [1] 2 3 2 1 1 2 2 2 5 2 1 1 2 2 2 3 1 2 1 4 2 2 2 2 1 2 1 1 2 3 1 2 5 2 5

Within cluster sum of squares by cluster:
[1] 16328557957 15546878379 16037533181           0 12918652281
 (between_SS / total_SS =  98.0 %)
```

Figure 4: K-means clustering summary for Indian regions based on COVID-19 metrics

4.0 Time Series Analysis

In extension to cluster analysis, authors have carried out time series analysis and visualization of covid 19 virus outbreak in different regions of India to study the trajectory of covid 19 growth curve from March 2020 to Feb 2021.

In current analysis month wise the average number of confirmed cases and deaths are compared in different regions of India, this visualization presents a snapshot indicating the trend of spread over time.

Figure 5. shows the R script for averaging the confirmed cases, active case and deaths in each region grouped by month is computed and then visualized the growth using ggplot with facet wrap.

```
library(ggplot2)
library(dplyr)
library(hrbrthemes)

library(lubridate)
set.seed(2017)
options(digits=4)

covid_data <- read.csv("E:/covid_19_data_set/covid_19.csv")
covid_data$S..No.=NULL
covid_data


region<- unique(covid_data$Region)
region

no_region<-length(region)
no_region

# declare and initialize empty covid_region_month dataframe
covid_region_month <- data.frame( Region = "X", year = 2020, month = 0, Confirmed_cases= 0, Active_cases=0, Cured_cases=0, Deaths=0,stringsAsFactors=FALSE)


#
for(j in 1:no_region)
{
  covid_region1<-subset(covid_data, Region== region[j])
covid_region<-covid_region1

for(i in 1: nrow(covid_region))
 {
  date_str= covid_region[i,1]
  dy= substr(date_str,1,2)
  mm= substr(date_str,4,5)
  yr= substr(date_str,7,10)
  new_dt<- paste(yr, mm, dy, sep ="-")
  covid_region[i,7]<-new_dt
```

```
}

as.Date(covid_region[,7])
# replace values in date column
covid_region$Date<-covid_region[,7]
# delete column 7
covid_region[,7]=NULL

covid_month<- covid_region %>% mutate(year = year(as.Date(Date)), month = month(as.Date(Date))) %>%
  group_by(Region, year, month) %>%
  summarise(Confirmed_cases= mean(Confirmed.Cases), Active_cases= mean(Active.Cases), Cured_cases= mean(Cured.Discharged), Deaths= mean(Death))

covid_region_month = rbind(covid_region_month, data.frame( covid_month))

}


covid_region_month <- covid_region_month[-1,]
# plot of region wise all months covid 19 confirmed cases

  covid_region_month %>%
  na.omit() %>%
  ggplot(aes(x = month, y = Confirmed_cases)) +
  geom_point(color = "darkorchid4") +
  facet_wrap( ~ Region ) +
  labs(title = "covid 19 confirmed cases ",
       subtitle = "Use facets to plot by a variable - region in this case",
       y = "confirmed casees",
       x = "month") + theme_bw(base_size = 10)+
 # adjust the x axis breaks
 scale_x_continuous(breaks = c(1,2,3,4,5,6,7,8,9,10,11,12))
 #scale_x_date(date_breaks = "months", date_labels = "%m")


# plot of region wise all months covid 19 confirmed cases amd Cured cases


  p = ggplot() +
    geom_point(data = covid_region_month, aes(x = month, y = Confirmed_cases), color = "blue") +
    geom_point(data = covid_region_month, aes(x = month, y = Cured_cases), color = "green") +
    facet_wrap( ~ Region ) +
    labs(title = "covid 19 confirmed cases and Cured cases  ",
         subtitle = "Use facets to plot by a variable - region in this case",
         y = "confirmed cases and deaths",
         x = "month") + theme_bw(base_size = 10)+
    # adjust the x axis breaks
    scale_x_continuous(breaks = c(1,2,3,4,5,6,7,8,9,10,11,12))

  print(p)

  # plot of region wise all months covid 19 confirmed cases, Active cases and deaths
  q = ggplot() +
    geom_point(data = covid_region_month, aes(x = month, y = Confirmed_cases), color = "blue") +
    geom_point(data = covid_region_month, aes(x = month, y = Cured_cases), color = "green") +
    geom_point(data = covid_region_month, aes(x = month, y = Deaths), color = "red") +
    facet_wrap( ~ Region ) +
    labs(title = "covid 19 confirmed cases, cured cases and  deaths ",
         subtitle = "Use facets to plot by a variable - region in this case",
         y = "confirmed cases and deaths",
         x = "month") + theme_bw(base_size = 10)+
    # adjust the x axis breaks
    scale_x_continuous(breaks = c(1,2,3,4,5,6,7,8,9,10,11,12))

  print(q)
```

Figure 5: R script of Time series analysis and visualization

5.0 Results and Discussions

The Indian regions as on 18th September 2021 are clustered into five groups based on COVID-19 metrics. The K-means clustering technique employed for this purpose. The clusters and the participating regions are tabulated in table 1. Analysing the cluster means, we can relate each group with each of the 5 classes of regions:

- Only Maharashtra in Cluster 4 with maximum number of Confirmed cases, Active cases, Cured cases, and Total deaths
- The regions falls under cluster 2 have minimum count Confirmed cases, Active cases, Cured cases, and Total deaths
- Cluster 3 formed by Andhra Pradesh, Karnataka and Tamil Nadu is second highest
- The regions Delhi, Uttar Pradesh and West Bengal forms Cluster 5 falls at 3rd position

Table 1: Clusters and participating regions from India

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Assam	Andaman and	Andhra Pradesh	Maharashtra	Delhi
Bihar	Nicobar Islands	Karnataka		Uttar Pradesh
Gujarat	Arunachal Pradesh	Tamil Nadu		West Bengal
Haryana	Chandigarh			
Kerala	Chhattisgarh			
Madhya Pradesh	Dadra and Nagar			
Odisha	Haveli and Daman			
Punjab	and Diu			
Rajasthan	Goa			
Telangana	Himachal Pradesh			

	Jammu and Kashmir Jharkhand Ladakh Manipur Meghalaya Mizoram Nagaland Pondicherry Sikkim Tripura Uttarakhand			

Figure 6 shows cluster plot for 1st and 2nd principle components. Plot reveals that cluster 2 is dense and some of the data points of cluster 2 are closer to cluster 1. There is no overlapping of cluster 3 and 5 with others. Cluster 4 with single data point away from the other clusters.

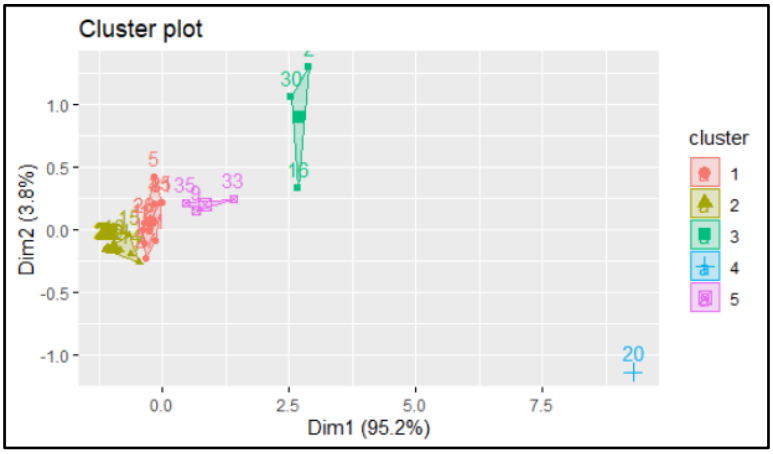

Figure 6: Cluster plot for 1st and 2nd principle components

As per the Time series analysis done the average covid confirmed cases and its growth curve are shown in Figure 7

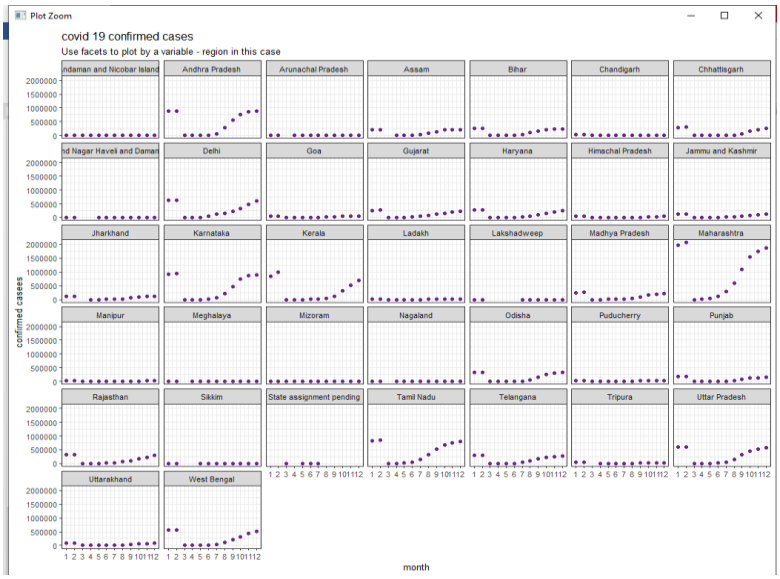

Figure 7: Region wise growth curve of covid 19 confirmed cases

Plot in Figure 7 shows change in the rate of spread from Feb 2020 to March 2021 in different Indian states/ regions. Maharashtra state showed an exponential growth curve between period July to Nov whereas region like Andhra Pradesh, Delhi, Karnataka , Uttar Pradesh , Tamil Nadu, West Bengal shows a logarithmic growth curve from month of July 2020 onwards, It is also observed that growth rate remains same in Gujarat, Goa, Madhya Pradesh Rajasthan, Odisha, Telegana from July 2020.

Figure 8 shows month wise plot of average number of confirmed cases and cured cases grouped by region,Blue colour indicates confirmed cases and Green colour for cured cases.It is observed that number of cured cases are inline with the number of confirmed cases.

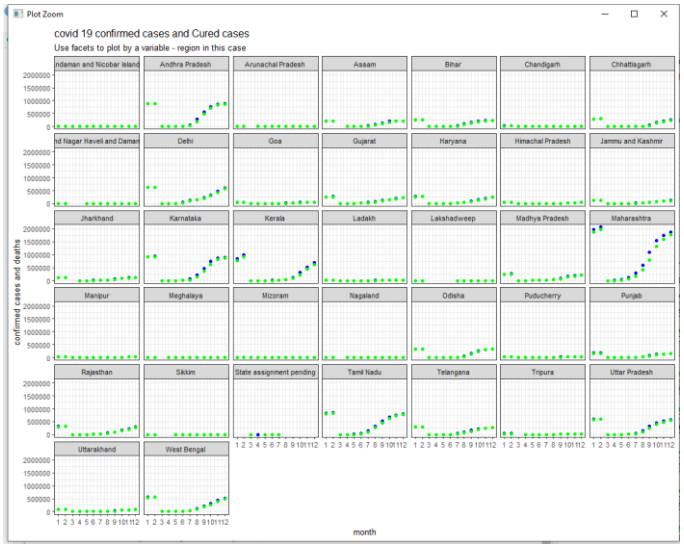

Figure 8: Average confirmed cases and cured cases grouped by region

Plot in Figure 9 shows month wise average confirmed cases, cured cases, deaths grouped by region, this plot clearly indicates that even in regions with exponential growth of confirmed cases like Maharashtra, Delhi, Karnataka, Andhra Pradesh the death rate is steady compared with confirmed cases.

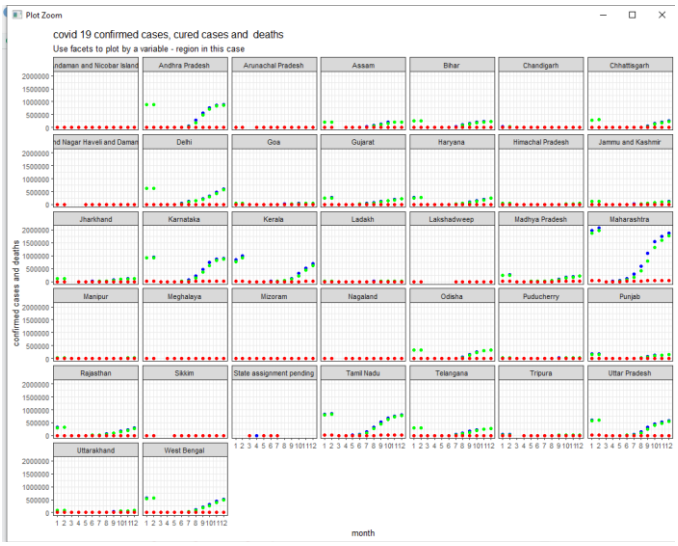

Figure 9: Region wise comparative plot of covid 19 confirmed case, cured cases and deaths

6.0 Conclusion

This study has demonstrated the K-means clustering as tool for analysing Indian regions based on COVID-19 metrics. The dataset for the present analysis retrieved from COVID-19 situation report on 18th September 2021. The 35 Indian regions are clustered into five groups based on the status of infectedas on 18th September 2021. This analysis has illustrated cluster formation and visualization by modulating number of clusters. Thus unsupervised learningtechnique results into 5groups of regions of sizes 10, 18, 3, 1 and 3. It's observed that Confirmed cases, Active cases, Cured cases, and Total deaths are highest in Maharashtra. Where as the regions falls under cluster 2 have minimum count for these parameters. The result depicts that k-means has the potential for Indian regions clustering based on COVID-19 metrics. The time series analysis and its visualization has shown an interesting trend of number of confirmed cases and deaths. The case fatality rate or CFR, which measures deaths among Covid-19 patients is comparatively very low as compared to confirmed cases. Even in badly-hit state like Maharashtra the death rates have kept low all along, even as cases rose. There are many attributes which could contribute to this relatively low fatality ratelike immunity deriving from previous infections but moreover high demographic of younger population in India could be one of the factor as elderly are typically more vulnerable.

7.0 References

COVID-19 State wise data, https:// prsindia.org/covid-19/cases, retrieved on 27th Feb 2021

Coronavirus Disease (COVID-19) Situation Reports, https://www.who.int/ emergencies/diseases/novel-coronavirus-2019/situation-reports, retrieved on 27th Feb 2021

Shannon Wongvibulsin, Brian T. Garibaldi, Annukka A.R. Antar, Jiyang Wen,Mei-Cheng Wang, Amita Gupta, Robert Bollinger, YanxunXu, Kunbo Wang, Joshua F. Betz, John Muschelli, Karen Bandeen-Roche, Scott L. Zeger, Matthew L. Robinson, Development of Severe COVID-19 Adaptive Risk Predictor (SCARP), a Calculator to Predict Severe Disease or Death in Hospitalized Patients With COVID-19, Annals of Internal Medicine, https://www. acpjournals.org/doi/pdf/10.7326/M20-6754, retrieved on 20th March 2021

Sudhir Chowdhary, Tech tactics: How Artificial Intelligence is aiding the fight against Covid-19, https://www. financialexpress.com/industry/technology/tech-tactics-how-artificial-intelligence-is-aiding-the-fight-against-covid-19/ 2214094/, retrieved on 20th March 2021

Hannah Mitchell, AI tool uses EHR data mining to support diagnostic decision-making, https://www.beckershospitalreview.com/artificial-intelligence/ ai-tool-uses-ehr-data-mining -to-support-diagnostic -decision-making.html, retrieved on 20th March 2021

R.S. Kamath, S.S. Jamsandekar, K.G. Kharade, R.K.Kamat, Data Analytics in R: A Case Study Based Approach, Himalaya Publishing House Pvt. Ltd, For MHRD-PMMMNMTT FDC in Cyber Security and Data Science, Shivaji University, Kolhapur, 2019, ISBN: 978-93-5367-791-6

R.S. Kamath, R.K. Kamat, Visualization of University Clusters based on NIRF and NAAC Scores: K-means Algorithm Approach, University News, A Weekly Journal of Higher Education, Vol. 57, No. 03, Jan 21-27, 2019, ISSN: 05662257

R.K. Kamat, R.S. Kamath, Visualization of Earthquake Clusters over Space: K-Means Approach, Journal of Chemical and Pharmaceutical Sciences (JCHPS), Volume 10, Issue 1, Jan-March 2017, Page 250-53, ISSN: 0974-2115

Mohsen Mousavi, RohitSalgotra, Damien Holloway, Amir H. Gandomi, "COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features", IEEE Computational Intelligence Magazine · October 2020 , DOI: 10.1109/MCI.2020.3019895