International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 6, September-October 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Study on Big Data Analysis using R Studio on COVID 19

Praveen Kumar[1], Jyoti Kataria[2]

[1]M Tech Scholar, [2]Asistant Professor,
[1,2]Department of Computer Science & Engineering,
[1,2]Manav Institute of Technology and Management, Jevra, Haryana, India

ABSTRACT

Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. This is focuses on scalable big-data systems, which include a set of tools and mechanisms to load, extract, and improve disparate data while leveraging the massively parallel processing power to perform complex transformations and analysis. This is based on COVID 19 patients data on accessible big-data systems that include a set of tools and technique to load, extract, and improve dissimilar data while leveraging the immensely parallel processing power to perform complex transformations and analysis.

KEYWORDS: *Big Data, COVID 19, R Studio, etc*

How to cite this paper: Praveen Kumar | Jyoti Kataria "Study on Big Data Analysis using R Studio on COVID 19" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-6, October 2020, pp.414-418, URL: www.ijtsrd.com/papers/ijtsrd33408.pdf

IJTSRD33408

1. INTRODUCTION

Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it.

We can drastically increase this number by supporting many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.[9]

This research is based on the standards of R Programming as direct simulator of R Programming. The analysis of Big Data involves multiple distinct phases, each of which introduces challenges. In this chapter we will describe about the design and evaluation simulation experiments. The main reason behind this is that these services are not fully implemented around the world. In mean time after full implementation, simulators will be easily available. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. [12]

BIG DATA PROBLEM AND CHALLENGES

How can the data be preprocessed in order to improve the quality of data and analysis results before we begin data analysis [1] [2]? As the sizes of dataset are often very large, sometimes several gigabytes or more, and their origin from varied sources, current real-world databases are pitilessly susceptible to inconsistent, incomplete, and noisy data. Therefore, a number of data preprocessing techniques, including data cleaning [11], data integration, data transformation and date reduction, can be applied to remove noise and correct irregularities. Different challenges arise in each sub-process when it comes to data-driven applications.

BIG DATA OPPORTUNITIES

This initiative will also lay the groundwork for complementary "Big Data" activities, such as "Big Data" substructure projects, platforms development, and techniques in settling complex, data-driven problems in sciences and engineering. Researchers, policy and decision makers have to recognize the potential of harnessing "Big Data" to uncover the next wave of growth in their fields. There are many advantages in business section that can be obtained through harnessing "Big Data" increasing operational efficiency, informing strategic direction, developing better customer service, identifying and developing new products and services, identifying new customers and markets, etc.

2. R STUDIO

The R language is well established as the language for doing statistics, data analysis, data-mining algorithm development, stock trading, credit risk scoring, market basket analysis and all [9] manner of predictive analytics. However, given the deluge of data that must be processed and analyzed today, many organizations have been reticent about deploying R beyond research into production applications.[12]

R is a statistical software, and an object-oriented high-level programming language used for data analysis, which includes a large number of statistical procedures such as t-test, chi-square test, standard linear models, instrumental variables estimation, local regression polynomials, etc. Besides, R provides high-level graphics capabilities. R is an object-oriented programming language. This means that everything what is done with R can be saved as an object. Every object has a class. [12]

It describes what the object contains and what each function does. Application of R as a programming language and statistical software is much more than a supplement to Stata, SAS, and SPSS. Although it is more difficult to learn, the biggest advantage of R is its free-of-charge feature and the wealth of specialized application packages and libraries for a huge number of statistical, mathematical and other methods. R is a simple, but very powerful data mining and statistical data processing tool and once "discovered", it provides users with an entirely new, rich and powerful tool applicable in almost every field of research.

3. COVID 19

In the past decades, several new diseases have emerged in new geographical areas, with pathogens including Ebola, Zika, Nipah, and coronaviruses (CoVs). Recently, a new type of viral infection has emerged in Wuhan City, China, and initial genomic sequencing data of this virus does not match with previously sequenced CoVs, suggesting a novel CoV strain (2019-nCoV), which has now been termed as severe acute respiratory syndrome CoV-2 (SARS-CoV-2). Although Coronavirus disease 2019 (COVID-19) is suspected to originate from an animal host (zoonotic origin) followed by human-to-human transmission, the possibility of other routes such as food-borne transmission should not be ruled out. Coronaviruses are large group of viruses that cause illness in humans and animals. Rarely, animal coronaviruses can evolve and infect people and then spread between people such as has been seen with MERS and SARS. The outbreak of Novel coronavirus disease (COVID-19) was initially noticed in a seafood market in Wuhan city in Hubei Province of China in mid-December, 2019, has now spread to 214 countries/territories/areas worldwide.[3] WHO (under International Health Regulations) has declared this outbreak as a "Public Health Emergency of International Concern" (PHEIC) on 30thJanuary 2020. WHO subsequently declared COVID-19 a pandemic on 11th March, 2020.Members of the family Corona virus cause a broad spectrum of animal and human diseases. Uniquely, replication of the RNA genome proceeds through the generation of a nested set of viral mRNA molecules. Human coronavirus (HCoV) infection causes respiratory diseases with mild to severe outcomes. In the last 15 years, we have witnessed the emergence of two zoonotic, highly pathogenic HCoVs: severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). Replication of HCoV is regulated by a diversity of host factors and induces drastic alterations in cellular structure and physiology. In this review all (as we possible) information about Corona viruses are given. KEYWORDS: Corona, respiratory, viruses, Hcov, host, RNA.

SCOPE

The guidelines are in addition to the guidelines on appropriate management of suspect/confirmed case of COVID-19 issued by MoHFW on 7th April, 2020. As per existing guidelines, during the containment phase the patients should be clinically assigned as very mild/mild, moderate or severe and accordingly admitted to (i) COVID Care Center, (ii) Dedicated COVID Health Center or (iii) Dedicated COVID Hospital respectively. Guidelines for home isolation of very mild/pre-symptomatic patients were issued on 27th April 2020. The present guidelines are in supersession of the guidelines issued on 27th April 2020. [1]

4. WHO

World health organizationis providing guidance on early investigations, which are critical in an outbreak of a new virus. The data collected from the protocols can be used to refine recommendations for surveillance and case definitions, to characterize the key epidemiological transmission features of COVID-19, help understand spread, severity, spectrum of disease, impact on the community and to inform operational models for implementation of countermeasures such as case isolation, contact tracing and isolation. Several protocols are available here. One such protocol is for the investigation of early COVID-19 cases and contacts (the "First Few X (FFX) Cases and contact investigation protocol for 2019-novel coronavirus (2019-nCoV) infection"). [2]

Recommended Test

Real time or Conventional RT-PCR test is recommended for diagnosis. SARS-CoV-2 antibody tests are not recommended for diagnosis of current infection with COVID-19. Dual infections with other respiratory infections (viral, bacterial and fungal) have been found in COVID-19 patients. Depending on local epidemiology and clinical symptoms, test for other potential etiologies (e.g. Influenza, other respiratory viruses, malaria, dengue fever, typhoid fever) as appropriate. For COVID-19 patients with severe disease, also collect blood cultures, ideally prior to initiation of antimicrobial therapy.[5]

Management of COVID-19

In the containment phase, patients with suspected or confirmed mild COVID-19 are being isolated to break the chain of transmission. Patients with mild disease may present to primary care/outpatient department, or detected during community outreach activities, such as home visits or by telemedicine. Mild cases can be managed at Covid Care Centre, First Referral Units (FRUs), Community Health Centre (CHC), sub-district and district hospitals or at home subject to conditions stipulated in the home isolation guidelines available at Detailed clinical history is taken including that of co-morbidities. Patient is followed up daily for temperature, vitals and Oxygen saturation (SpO2). [6]

5. Explore Variables

Distribution of every numeric variable can be checked with function summary (), which returns the minimum, maximum, mean, median, and the first and third quartiles. For factors (or categorical variables), it shows the frequency of every level.

The frequency of factors can be calculated with function table(), and then plotted as a line chart of covid 19 data country wise confirmed, Death and Recoverd case on till August 2020

Table 5.1 4M

	Min	Median	Mean	Max
Confirmed	0	1622	15443	801422
Deaths	0	31	287	42072
Recovered	0	692.5	8393	2140614

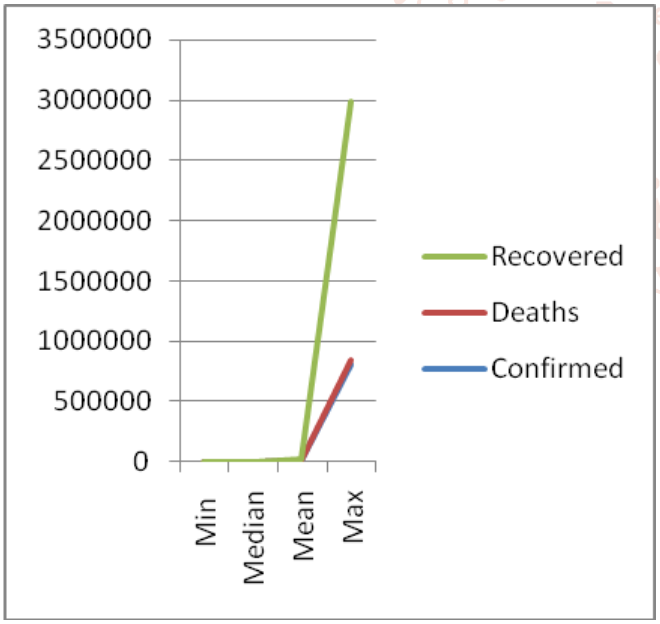

Fig. 5.1 Frequency of Factors (Line Chart)

The above line graph calculated with Min, Median, Mean and Max function (), and then plotted as a line chart of covid 19 data country wise confirmed, Death and Recovered case on till August 2020.

The nest command is used to plot a line graph on covid19 data.

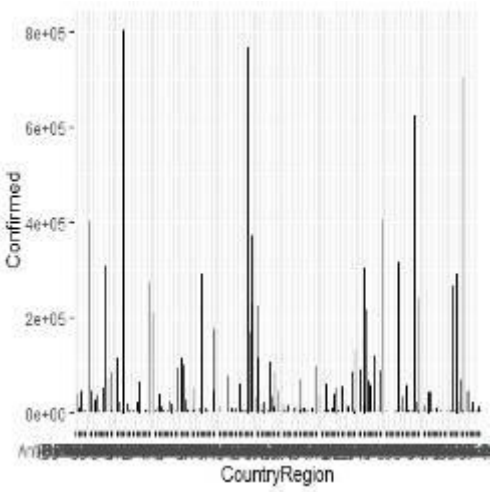

Fig 5.2 Country Wise Confirmed Case

The above line graph calculated with data country wise confirmed case on till August 2020 then plotted as a line chart.

The nest command is used to plot line graph on coviddata.

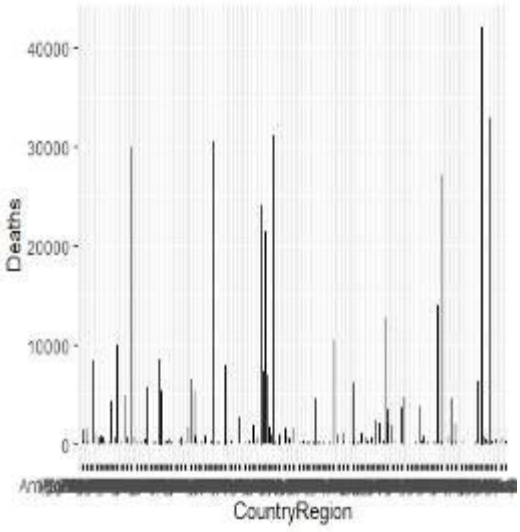

Fig 5.3 Country Wise Deaths Case

The above line graph calculated with data country wise death case on till August 2020 then plotted as a line chart.
The nest command is used to plot line graph on coviddata.

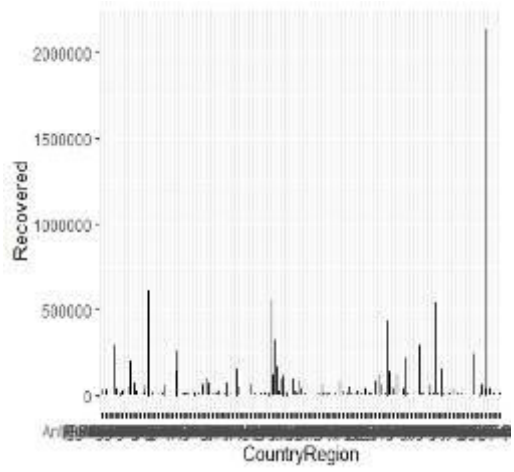

Fig 5.4 Country Wise Recovered Case

The above line graph calculated with data country wise recovered case on till August 2020. and then plotted as a line chart.

The nest command is used to plot line graph on coviddata.

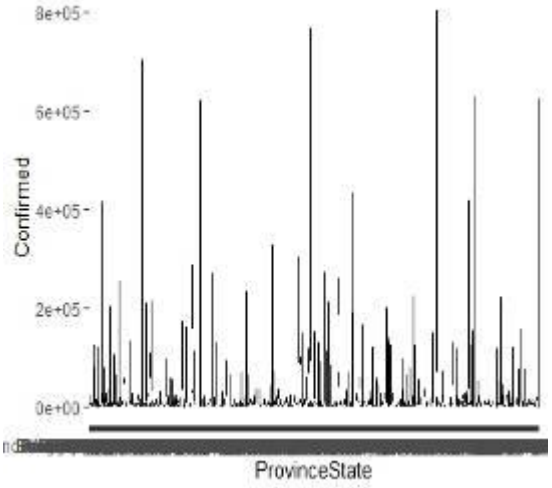

Fig 5.5 State Wise Confirmed Case

The above line graph calculated with data state wise confirmed case on till August 2020 then plotted as a line chart.

The nest command is used to plot line graph on coviddata.

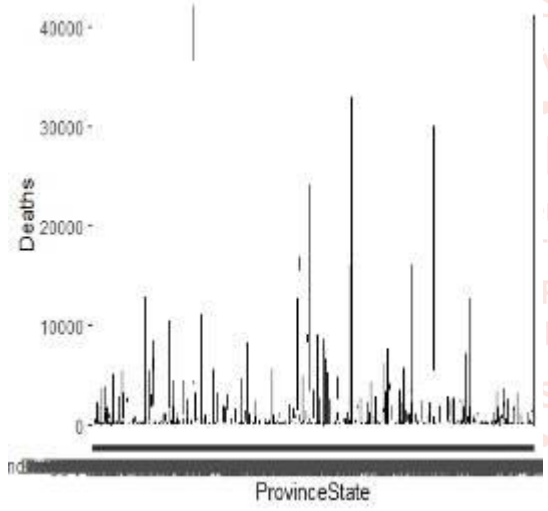

Fig 5.6 State wise death case

The above line graph calculated with data state wise death case on till August 2020. and then plotted as a line chart.

The nest command is used to plot line graph on coviddata.

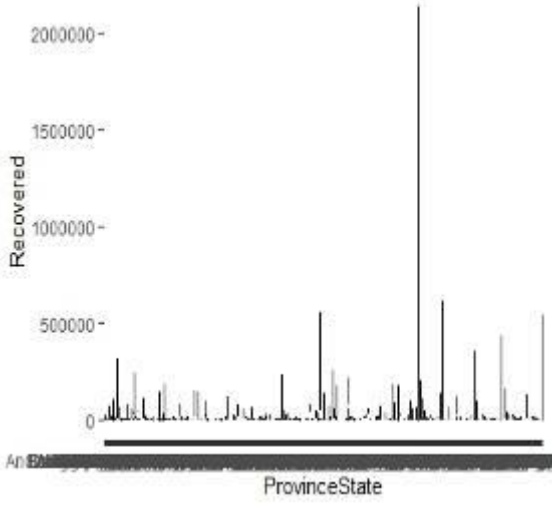

Fig 5.7 State Wise Recovered

The above line graph calculated with data state wise recovered case on till August 2020. then plotted as a line chart.

6. Conclusion

Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone.

Future Work

This research is based on the standards of R Programming and services, as direct simulator of R Programming and services are unavailable, as they are very expensive. The main reason behind this is that these services are not fully implemented around the world. In mean time after full implementation, simulators will be easily available. The future work will be on simulator of R Programming and services, to improve the performance of Big Data Analysis.

REFERENCES

[1] PalashGhosh, RikGhosh and BibhasChakraborty "COVID-19 in India: State-wise Analysis and Prediction" Published in April 29, 2020. https://doi.org/10.1101/2020.04.24.20077792.

[2] Abdul Hafeez, Shmmon Ahmad, Sameera Ali Siddqui, Mumtaz Ahmad, Shruti "A Review of COVID-19 (Coronavirus Disease-2019) Diagnosis, Treatments and Prevention" DOI: 10.14744/ejmo.2020.90853 EJMO 2020;4(2):PP 116–125

[3] Kit-San Yuen1, Zi-Wei Ye2, Sin-Yee Fung1, Chi-Ping Chan1 and Dong-Yan Jin1 Yuen et al. "SARS-CoV-2 and COVID-19: The most important research questions Cell Biosci" (2020) doi.org/10.1186/s13578-020-00404-4

[4] Andrew Crotty, Alex Galakatos, KayhanDursun, Tim Kraska, UgurCetintemel, Stan Zdonik, "Tupleware: "Big" Data, Big Analytics, Small Clusters", 2016

[5] Jennifer Ortiz, Victor Teixeira de Almeida, Magdalena Balazinska, "Changing the Face of Database Cloud Services with Personalized Service Level Agreements", 2015

[6] Rajan Gupta, Saibal K. Pal and GauravPandey"A Comprehensive Analysis of COVID-19 Outbreak situation in India" this version posted April 11, 2020. doi.org/10.1101/2020.04.08.20058347.

[7] Situation Report – 94 "Coronavirus disease (COVID-19)" August 2020

[8] Coronavirus disease (COVID-19) Situation Report – 202 Data as received by WHO from national authorities by CEST, 9 August 2020

[9] COVID-19 for India Updates "Data as received by WHO from international authorities CEST" 23 April 2020

[10] Francesco Di Gennaro, DamianoPizzol, Claudia Marotta, Mario Antunes, Vincenzo Racalbuto, Nicola Veronese and Lee Smith: "Coronavirus Diseases (COVID-19) Current Status and Future Perspectives: A Narrative Review" Published International Journal of Environmental Research and Public Health 14 April 2020

[11] Anant Bhardwaj1, Souvik Bhattacherjee2, Amit Chavan2 Amol Deshpande2, Aaron J. Elmore1,3, Samuel Madden1, AdityaParameswaran, "DataHub: Collaborative Data Science & Dataset Version Management at Scale", 2015

[12] Challenges and Opportunities with Big Data

[13] HongboZou, Yongen Yu, Wei Tang, Hsuan-Wei Michelle Chen, "Flex Analytics: A Flexible Data Analytics Framework for Big Data Application with I/O Performance Improvement", Elsevier 2014

[14] Alekh Jindal, Robust Data Transformations, 2015

[15] RaduTudoran, "High-Performance Big Data Management Across Cloud Data Centers", Jan 2015

[16] Bill Howe, "Big Data Science Needs Big Data Middleware", Jan 2015