Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766409

DATA ANALYSIS AND VISUALIZATION OF COVID-19 USING R-LANGUAGE

Ronak Vadiwala
Research Scholar, Department of Computer Science, PAHER University, Udaipur, India
vadiwalaronak@gmail.com

Dr. Neetu Agarwal
Assistant Professor, Department of Computer Science, PAHER University, Udaipur, India
neetu.agarwal1508@gmail.com

Abstract :
R is a software environment and computer language for statistical analysis, visual representation, and reporting. In addition to reporting and web development, R is growing in popularity. In India, there is a severe scarcity of trained R analysts. As firms want to expand their R capabilities, the few who already have the necessary skills are in high demand. If you want to have a successful career in data science, it's worth studying R programming. This FDP has assisted us in learning how to use R's graphics library. Work with datasets in an efficient manner. Our own analysis reports should be generated. Improve our ability to hire by learning in a creative and self-directed manner. On completion of the course, you will receive a certificate. In this paper, we try to concentrate on the visualisation of covid-19 as well as the data analysis using R.

Keywords : R language , Data analysis , Covid-19 and visualization .

1.0. Introduction:

The SARS-Cov2 virus is about to unleash a global COVID-19 pandemic. Every news storey features grim, ever-increasing global figures of COVID-19 cases and deaths. Like Christmas trees, worldwide spread dashboards are starting to light up[1].

R is one of the tools of choice for outbreak epidemiologists, and a fast search will turn up a slew of R libraries dedicated to outbreak management and analysis on CRAN and elsewhere[2]. Obtaining extensive, reliable, and up-to-date data on the COVID-19 outbreak is not as simple as it may appear. Various national and provincial/governmental online sites in afflicted nations provide extensive summary data on incident cases, recovered cases, and deaths caused by the virus, but these data are typically presented as tallies interspersed in (mostly non-English) language[3].

There are a number of potential data sources that have been extracted and compiled from government websites. A widely-used source is a government-provided dataset that serves as the dashboard's source. It's simple to use: simply read CSV files from the relevant GitHub URL[4].

R is free and open-source software distributed under the GNU General Public License, with pre-compiled binary versions available for Linux, Windows, and Mac. R was named from the initial letter of the first names of the two R authors (Robert Gentleman and Ross Ihaka), as well as a play on the Bell Labs Language S. This FDP is for software programmers, statisticians, and data miners who want to use R programming to create statistical software. If you're new to

R programming, this curriculum will provide you with a solid foundation in practically all of the language's fundamentals, allowing you to progress to higher levels of proficiency[5,6].
R is the most useful programming language, with a wide range of uses. For statistics, data analysis, and machine learning, R programming is the ideal technique[7]. The best way to create repeatable, high-quality analysis is to use R programming. R programming makes it simple to apply algorithms to data, and there are numerous packages available to help you execute your algorithmic approaches and analyses. In addition to reporting and web development, R is growing in popularity[8,9].

2.0. Purpose of Work:

We are utilising R programming concepts to analyse and visualise the data in the COVID-19 Data Analysis and Visualization using R Language Project Report. As COVID-19 is a pandemic disease, we are using various levels of data from various nations throughout the world.

3.0. Methodology:

Importing Data Analysis Packages

We will import the necessary packages for our data analysis project in the first step of our R project. We will utilise the following R libraries:

• ggplot2

This is where the project's foundation lies. The most extensively used data visualisation library for creating beautiful visualisation plots is ggplot2.

• ggthemes

This is a complement to our core ggplot2 library. We can now use the mainstream ggplot2 software to better generate additional themes and scales.

• lubridate There are several time frames in our dataset. The lubridate programme will be used to help us interpret our data in different time groups.

• dplyr

This package is R's data manipulation lingua franca.

• tidyr

This software will assist you with data cleanup. The primary idea behind tidyr is to clean up the columns, where each variable is represented by a column, each observation by a row, and each value by a cell.

• DT

We'll be able to interact with the Datatables JavaScript Library with the help of this package.

• scales

We can map data to the correct scales with well-placed axes and legends using graphical scales.

4.0. Results and Discussion:

```
jhu_url <- paste("https://raw.githubusercontent.com/CSSEGISandData/",
  "COVID-19/master/csse_covid_19_data/", "csse_covid_19_time_series/",
  "time_series_19-covid-Confirmed.csv", sep = "")
us_confirmed_long_jhu <- read_csv(jhu_url) %>% rename(province = "Province/State",
  country_region = "Country/Region") %>% pivot_longer(-c(province,
  country_region, Lat, Long), names_to = "Date", values_to = "cumulative_cases") %>%
```

```
mutate(Date = mdy(Date) - days(1)) %>% filter(country_region ==
"US") %>% arrange(province, Date) %>% group_by(province) %>%
mutate(incident_cases = c(0, diff(cumulative_cases))) %>%
ungroup() %>% select(-c(country_region, Lat, Long, cumulative_cases)) %>%
filter(str_detect(province, "Diamond Princess", negate = TRUE))
```

US case COVID-19

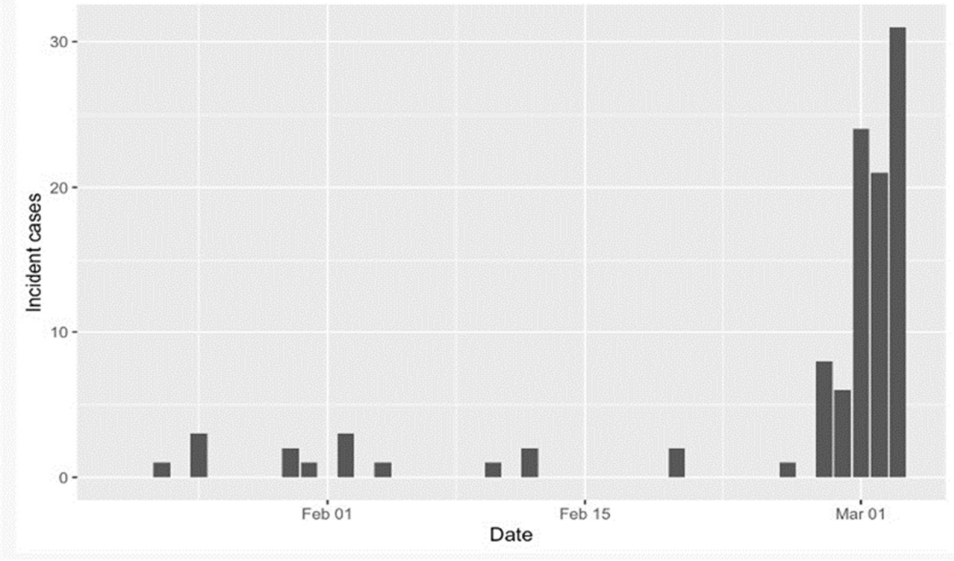


Figure: US COVID-19 Case [1]

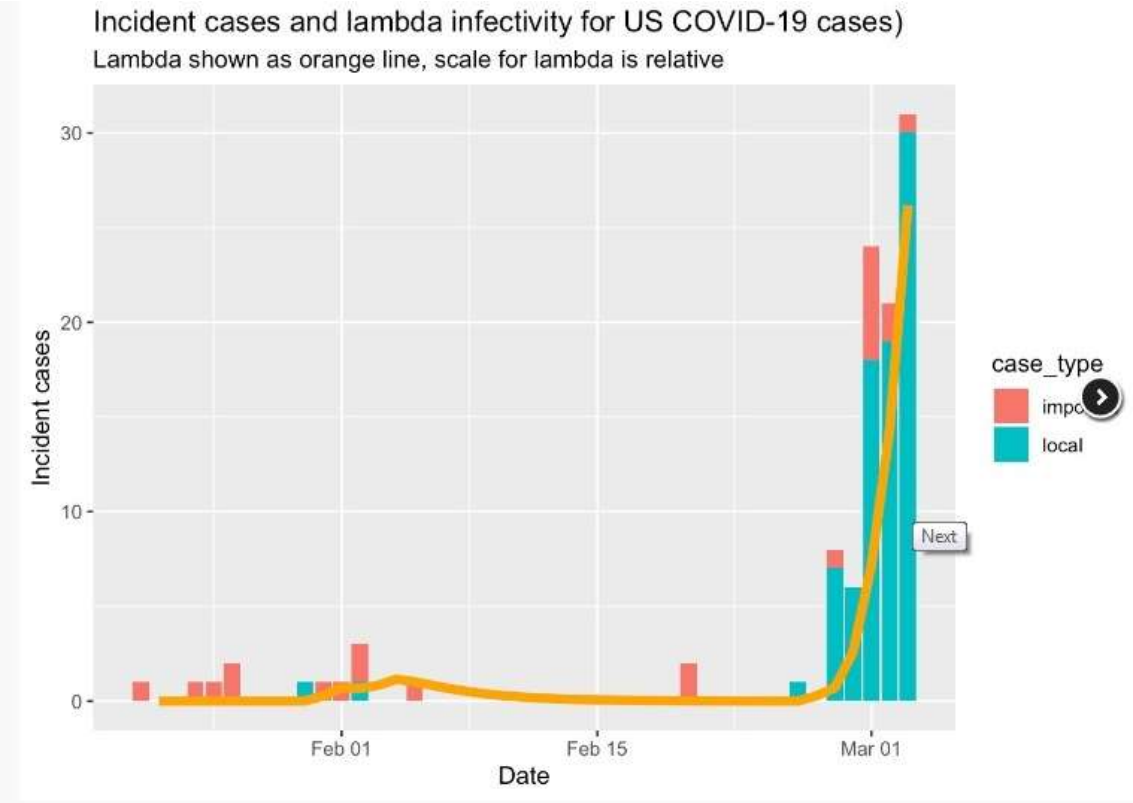


Figure 2 . US COVID-19 Analysis [1]

In this step of data science project, we will create a vector of our colors that will be included in our plotting functions.

Code:

1. colors = **c**(""#CC1011", "#665555", "#05a399", "#cfcaca", "#f5e840", "#0683c9", "#e075b0"")

Input Screenshot 4:

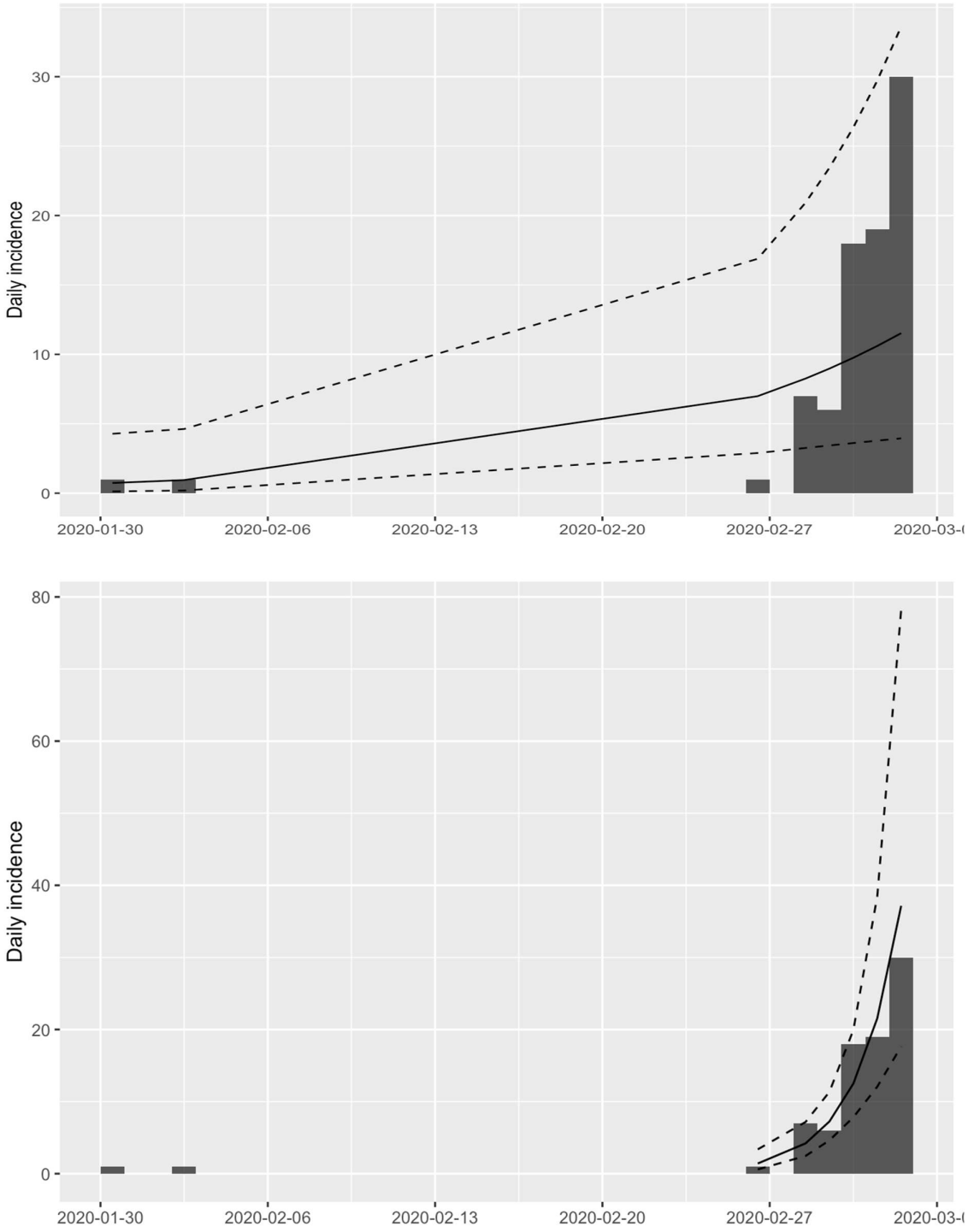
```
colors = c("#CC1011", "#665555", "#05a399", "#cfcaca", "#f5e840", "#0683c9", "#e075b0")
```

3. Reading the Data into their designated variables

Now, we will read several csv files that contain the data. We will store these in corresponding data frames like apr_data, may_data, etc. After we have read the files, we will combine all of this data into a single data frame.

Code:

```
1.  ggplot(hour_data, aes(hour, Total)) +
2.  geom_bar( stat = "identity", fill = "steelblue", color = "red") +
3.  ggtitle("Every Day") +
4.  theme(legend.position = "none") +
5.  scale_y_continuous(labels = comma)
6.  month_hour <- data_2020 %>%
8.  group_by(month, hour) %>%
9.  dplyr::summarize(Total = n())
10. ggplot(month_hour, aes(hour, Total, fill = month)) +
12. geom_bar( stat = "identity") +
13. ggtitle("Case Month") +
14. scale_y_continuous(labels = comma)
```



Estimating the instantaneous *effective reproduction ratio[1]*

The graphic display of data is known as date visualisation. It's used to create visuals that show how various parameters interact. The parameters' values can be determined by a thorough

evaluation of the factors required for visualisation production. To explain information clearly, tools such as statistical graphics, charts, information graphics, and other graphs are utilised. Individuals will be able to grasp the event, analyse it, uncover causes, and gather proof for a problem with the use of visualisation.

It's both an art and a science to visualise data. One of the issues in data visualisation is processing, analysing, and sharing data ethically in the current era.

• Show the data

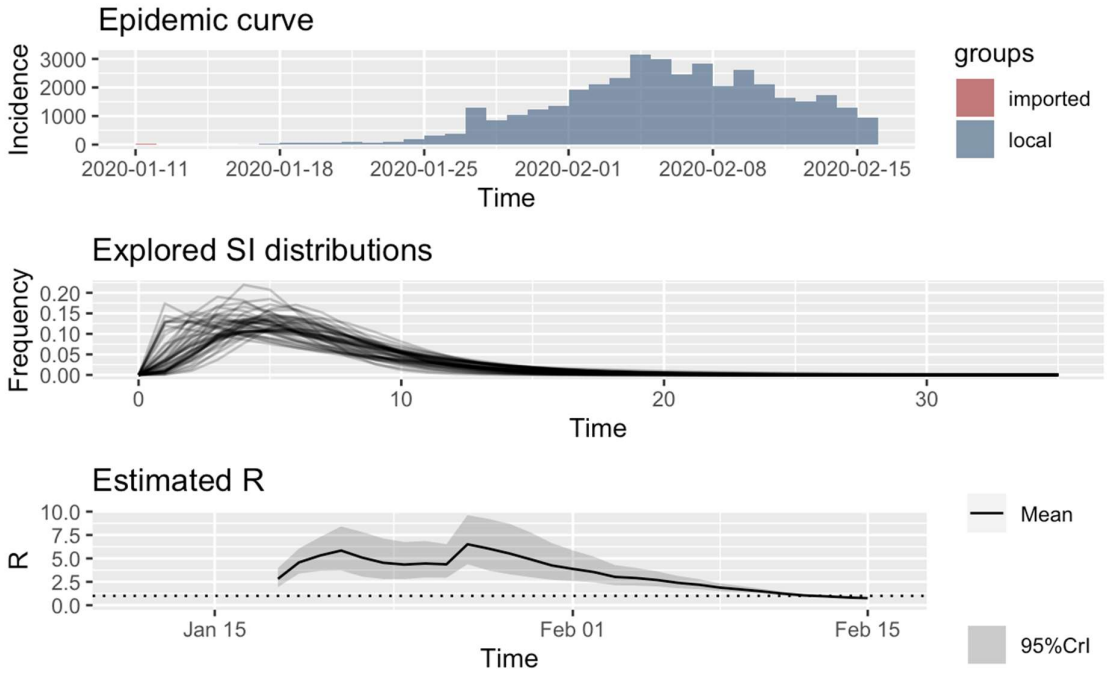
• Induce the viewer to think about the substance rather than methodology, graphic design, graphic production technology, or something else

• Avoid distorting what the data is trying to say

• Present many numbers in a small space

• Make large data sets coherent

• Encourage the eye to compare different pieces of data.



5.0. Effect of COVID-19 Pandemic:

• The SARS-COV-2 virus causes COVID 19 (Coronavirus Disease 2019). It was discovered for the first time in December 2019 in Wuhan, China's Hubei province capital. COVID-19 has been spreading at a quicker rate since its discovery. There have been about 3.77 million cases documented throughout 187 nations and territories, resulting in over 264,000 deaths. Despite this, more than 1.25 million patients have made a full recovery. The virus has no vaccinations or antiviral therapies, according to the WHO.

•In our study, we used R programming to show the behaviour of COVID-19 cases dependent on geographic location.

•R is a free statistical software environment. Statisticians and data miners utilise this to create statistical tools like SAS, SPSS, and Stata. User-created packages enhance Ra's capabilities, allowing for specialised statistical techniques, graphical devices, import/export capabilities,

reporting tools, and so on. R comes with a basic set of packages, and there are over 15,000 more on the R achieve network.

6.0. Conclusion:

- This project examines how COVID-19 has impacted the globe and how the information gained might be applied to further research. They start by developing a fresh suggestion based on those unanticipated results. If you believe you require extra testing, say what you believe should occur next. The charts can also be used to conclude crucial data insights in different scenarios..

7.0. Future work:

• To gain a better understanding of other factors that influence the fatality rate. This project is currently in progress, as there are still a lot of questions to be answered about this condition. Nonetheless, it provides as a springboard for further investigation into issues relating to the global pandemic. A dashboard with interactive graphs to provide a high-level overview.

8.0. Acknowledgment:

I am grateful to God and my guide for providing me with such a significant opportunity in my life. I'd also like to express my gratitude to all who have assisted me in my work, whether directly or indirectly.

9.0.Reference :

[1] Lin S, Pan H, Wu H, Yu X, Cui P et al. Epidemiological and clinical characteristics of 161 discharged cases with coronavirus disease 2019 in Shanghai, China. BMC Infect Dis. 2020 Oct 20;20(1):780. doi: 10.1186/s12879-020-05493-7. PMID: 33081711; PMCID: PMC7573864.

[2] Sanche S, Lin YT, Xu, C et al. The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. medRxiv; 2020. DOI: 10.1101/2020.02.07.20021154.

[3] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet. 2020 Feb 29;395(10225):689-697. doi: 10.1016/S0140-6736(20)30260-9. Epub 2020 Jan 31. Erratum in: Lancet. 2020 Feb 4;: PMID: 32014114; PMCID: PMC7159271.

[4] Lai S, Bogoch I, Ruktanonchai N, Watts A, Lu X et al. Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study. medRxiv [Preprint]. 2020 Feb 5:2020.02.04.20020479. doi: 10.1101/2020.02.04.20020479. PMID: 32511631; PMCID: PMC7276059.

[5] Jandrić P. Postdigital Research in the Time of Covid-19. Postdigit Sci Educ 2, 233–238 (2020). https://doi.org/10.1007/s42438-020-00113-8.

[6] Breslin N, Baptiste C, Gyamfi-Bannerman C et al. Coronavirus disease 2019 infection among asymptomatic and symptomatic pregnant women: two weeks of confirmed presentations to an affiliated pair of New York City hospitals. Am J Obstet Gynecol MFM. 2020 May;2(2):100118. doi: 10.1016/j.ajogmf.2020.100118. Epub 2020 Apr 9. PMID: 32292903; PMCID: PMC7144599.

[7] Huang L, Zhang X, Zhang X, Wei Z, Zhang L et al. Rapid asymptomatic transmission of COVID-19 during the incubation period demonstrating strong infectivity in a cluster of youngsters aged 16-23 years outside Wuhan and characteristics of young patients with COVID- 19: A prospective contact-tracing study. J Infect. 2020 Jun;80(6):e1-e13. doi: 10.1016/j.jinf.2020.03.006. Epub 2020 Apr 10. PMID: 32283156; PMCID: PMC7194554.

[8] Pan L, Mu M, Yang P, Sun Y, Wang R, Yan J, Li P, Hu B et al. Clinical Characteristics of COVID-19 Patients With Digestive Symptoms in Hubei, China: A Descriptive, Cross-Sectional, Multicenter Study. Am J Gastroenterol. 2020 May;115(5):766-773. doi: 10.14309/ajg.0000000000000620. PMID: 32287140; PMCID: PMC7172492.

[9] Wu Z and McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020 Apr 7;323(13):1239-1242. doi: 10.1001/jama.2020.2648. PMID: 32091533.