UNIVERSITY INSTITUTE of COMPUTING
Asia's Fastest Growing University

NAAC GRADE A+
Accredited University

Analysis of COVID-19 Trends and Patterns

Student Name: Divyanshi Kushwaha **UID: 24MCI10148**
Branch: MCA(AIML) **Section/Group: 24MAM-3A**
Semester: 1st **Date of Performance:17/10/24**
Subject Name: R Programming Lab **Subject Code: 24CAP-614**

1.Aim/Overview of the practical:

The aim of this project is to conduct a comprehensive exploratory data analysis (EDA) on a global COVID-19 dataset, with a focus on understanding the distribution and impact of the pandemic across various countries. The primary goal is to analyze key metrics such as total confirmed cases, deaths, recoveries, and active cases, by aggregating the data at the country level. Through this process, the analysis will aim to uncover patterns, trends, and anomalies in the spread of COVID-19, as well as the varying degrees of severity across different regions.

In addition, the project will leverage visualizations like boxplots and histograms to facilitate a deeper understanding of the distribution of COVID-19 cases across countries, enabling insights into how the pandemic has unfolded. By filtering and sorting the data, the analysis will also focus on identifying countries that have been most severely impacted, highlighting those with the highest number of confirmed cases (e.g., countries with more than 1 million cases).

2.Objective:

- Install and load necessary R packages for data manipulation and visualization.
- Load the COVID-19 dataset from a local CSV file.
- Perform data exploration by checking the structure, dimensions, and summary of the dataset.
- Aggregate the dataset by countries to obtain total confirmed, deaths, recovered, and active cases.
- Create visualizations such as boxplots and histograms to better understand the distribution of COVID-19 cases across countries.
- Filter the dataset to focus on countries with high numbers of confirmed cases (e.g., more than 1 million).
- Sort the dataset to identify the countries with the highest confirmed cases.

3.Task to be done:

- Package Installation & Loading: Install and load required packages like ggplot2, dplyr, tidyr, and readr for data handling and visualization.
- Data Import: Import the dataset from a local file path and inspect its structure using functions like head(), tail(), str(), and summary().
- Data Aggregation: Group the dataset by countries and calculate the total confirmed, deaths, recovered, and active cases.
- Visualizations:
  a) Create a boxplot to compare the total confirmed cases across countries.
  b) Create a histogram to show the distribution of total confirmed cases.
- Filtering: Extract countries with more than 1 million confirmed cases for further analysis.
- Sorting: Sort the countries by the number of total confirmed cases in descending order.

4.Steps/Commands involved to perform practical:

Installing Packages

```
> install.packages("ggplot2")
Installing package into 'C:/Users/divya/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://mirror.niser.ac.in/cran/bin/windows/contrib/4.4/ggplot2_3.5.1.zip'
Content type 'application/zip' length 5009017 bytes (4.8 MB)
downloaded 4.8 MB

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\divya\AppData\Local\Temp\Rtmp6vWsyC\downloaded_packages
> install.packages("dplyr")
Installing package into 'C:/Users/divya/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
trying URL 'https://mirror.niser.ac.in/cran/bin/windows/contrib/4.4/dplyr_1.1.4.zip'
Content type 'application/zip' length 1582782 bytes (1.5 MB)
downloaded 1.5 MB

package 'dplyr' successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package 'dplyr'
Warning: restored 'dplyr'
```

```
> install.packages("tidyr")
Installing package into 'C:/Users/divya/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
trying URL 'https://mirror.niser.ac.in/cran/bin/windows/contrib/4.4/tidyr_1.3.1.zip'
Content type 'application/zip' length 1269828 bytes (1.2 MB)
downloaded 1.2 MB

package 'tidyr' successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package 'tidyr'
Warning: restored 'tidyr'

The downloaded binary packages are in
        C:\Users\divya\AppData\Local\Temp\Rtmp6vWsyC\downloaded_packages
Warning message:
In file.copy(savedcopy, lib, recursive = TRUE) :
  problem copying C:\Users\divya\AppData\Local\R\win-library\4.4\00LOCK\tidyr\libs\x64\tidyr.dll to C:\Users\divya\AppData\Local\R\win-library\4.4\tidyr\libs\x64\tidyr.dll: Permission den$
> install.packages("readr")
Installing package into 'C:/Users/divya/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
trying URL 'https://mirror.niser.ac.in/cran/bin/windows/contrib/4.4/readr_2.1.5.zip'
Content type 'application/zip' length 1205684 bytes (1.1 MB)
downloaded 1.1 MB

package 'readr' successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package 'readr'
Warning: restored 'readr'

The downloaded binary packages are in
        C:\Users\divya\AppData\Local\Temp\Rtmp6vWsyC\downloaded_packages
Warning message:
In file.copy(savedcopy, lib, recursive = TRUE) :
  problem copying C:\Users\divya\AppData\Local\R\win-library\4.4\00LOCK\readr\libs\x64\readr.dll to C:\Users\divya\AppData\Local\R\win-library\4.4\readr\libs\x64\readr.dll: Permission den$
> library(ggplot2)
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':
```

Display number of rows and columns

```
> library(tidyr)
> library(readr)
> data <- read.csv("C:/Users/divya/Downloads/extracted_data/covid_19_clean_complete.csv")
> dim(covid_data)  # Displays the number of rows and columns
Error: object 'covid_data' not found
> dim(data)  # Displays the number of rows and columns
[1] 49068    10
> head(data)
  Province.State      Country.Region       Lat      Long       Date Confirmed Deaths Recovered Active           WHO.Region
1                        Afghanistan  33.93911  67.70995 2020-01-22         0      0         0      0 Eastern Mediterranean
2                            Albania  41.15330  20.16830 2020-01-22         0      0         0      0               Europe
3                            Algeria  28.03390   1.65960 2020-01-22         0      0         0      0               Africa
4                            Andorra  42.50630   1.52180 2020-01-22         0      0         0      0               Europe
5                             Angola -11.20270  17.87390 2020-01-22         0      0         0      0               Africa
6                Antigua and Barbuda  17.06080 -61.79640 2020-01-22         0      0         0      0             Americas
> tail(data)
      Province.State        Country.Region       Lat      Long       Date Confirmed Deaths Recovered Active           WHO.Region
49063                       Western Sahara  24.21550 -12.88580 2020-07-27        10      1         8      1               Africa
49064                Sao Tome and Principe   0.18640   6.61310 2020-07-27       865     14       734    117               Africa
49065                                Yemen  15.55273  48.51639 2020-07-27      1691    483       833    375 Eastern Mediterranean
49066                              Comoros -11.64550  43.33330 2020-07-27       354      7       328     19               Africa
49067                           Tajikistan  38.86100  71.27610 2020-07-27      7235     60      6028   1147               Europe
49068                              Lesotho -29.61000  28.23360 2020-07-27       505     12       128    365               Africa
> str(data)
'data.frame':   49068 obs. of  10 variables:
 $ Province.State: chr  "" "" "" "" ...
 $ Country.Region: chr  "Afghanistan" "Albania" "Algeria" "Andorra" ...
 $ Lat           : num  33.9 41.2 28 42.5 -11.2 ...
 $ Long          : num  67.71 20.17 1.66 1.52 17.87 ...
 $ Date          : chr  "2020-01-22" "2020-01-22" "2020-01-22" "2020-01-22" ...
 $ Confirmed     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Deaths        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Recovered     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Active        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ WHO.Region    : chr  "Eastern Mediterranean" "Europe" "Africa" "Europe" ...
> summary(data)
 Province.State     Country.Region         Lat              Long              Date             Confirmed            Deaths           Recovered           Active         WHO.Region
 Length:49068       Length:49068       Min.   :-51.796   Min.   :-135.00   Length:49068       Min.   :      0   Min.   :     0.0   Min.   :      0   Min.   :    -14   Length:49068
 Class :character   Class :character   1st Qu.:  7.873   1st Qu.: -15.31   Class :character   1st Qu.:      4   1st Qu.:     0.0   1st Qu.:      0   1st Qu.:      0   Class :character
 Mode  :character   Mode  :character   Median : 23.634   Median :  21.75   Mode  :character   Median :    168   Median :     2.0   Median :     29   Median :     26   Mode  :character
                                       Mean   : 21.434   Mean   :  23.53                      Mean   :  16885   Mean   :   884.2   Mean   :   7916   Mean   :   8085
                                       3rd Qu.: 41.204   3rd Qu.:  80.77                      3rd Qu.:   1518   3rd Qu.:    30.0   3rd Qu.:    666   3rd Qu.:    606
                                       Max.   : 71.707   Max.   : 178.06                      Max.   :4290259   Max.   :148011.0   Max.   :1846641   Max.   :2816444
```

View Aggregate data

```
> covid_aggregated <- data %>%
+   group_by(Country.Region) %>%
+   summarise(
+     total_confirmed = sum(Confirmed, na.rm = TRUE),
+     total_deaths = sum(Deaths, na.rm = TRUE),
+     total_recovered = sum(Recovered, na.rm = TRUE),
+     total_active = sum(Active, na.rm = TRUE)
+   )
>
> # View the aggregated data
> head(covid_aggregated)
# A tibble: 6 × 5
  Country.Region      total_confirmed total_deaths total_recovered total_active
  <chr>                         <int>        <int>           <int>        <int>
1 Afghanistan                 1936390        49098          798240      1089052
2 Albania                      196702         5708          118877        72117
3 Algeria                     1179755        77972          755897       345886
4 Andorra                       94404         5423           69074        19907
5 Angola                        22662         1078            6573        15011
6 Antigua and Barbuda            4487          326            2600         1561
> head(covid_aggregated)
# A tibble: 6 × 5
  Country.Region      total_confirmed total_deaths total_recovered total_active
  <chr>                         <int>        <int>           <int>        <int>
1 Afghanistan                 1936390        49098          798240      1089052
2 Albania                      196702         5708          118877        72117
3 Algeria                     1179755        77972          755897       345886
4 Andorra                       94404         5423           69074        19907
5 Angola                        22662         1078            6573        15011
> ggplot(covid_aggregated, aes(x = reorder(Country.Region, total_confirmed), y = total_confirmed)) +
+   geom_boxplot(fill = "lightblue") +
+   coord_flip() +
+   labs(title = "Boxplot of Total COVID-19 Confirmed Cases by Country",
+        x = "Country", y = "Total Confirmed Cases") +
+   theme_minimal()
>
> # Histogram of Total Confirmed Cases
> ggplot(covid_aggregated, aes(x = total_confirmed)) +
+   geom_histogram(binwidth = 100000, fill = "lightgreen", color = "black") +
+   labs(title = "Histogram of Total COVID-19 Confirmed Cases",
+        x = "Total Confirmed Cases", y = "Frequency") +
+   theme_minimal()
```

Total COVID-19 Confirmed Cases by Country

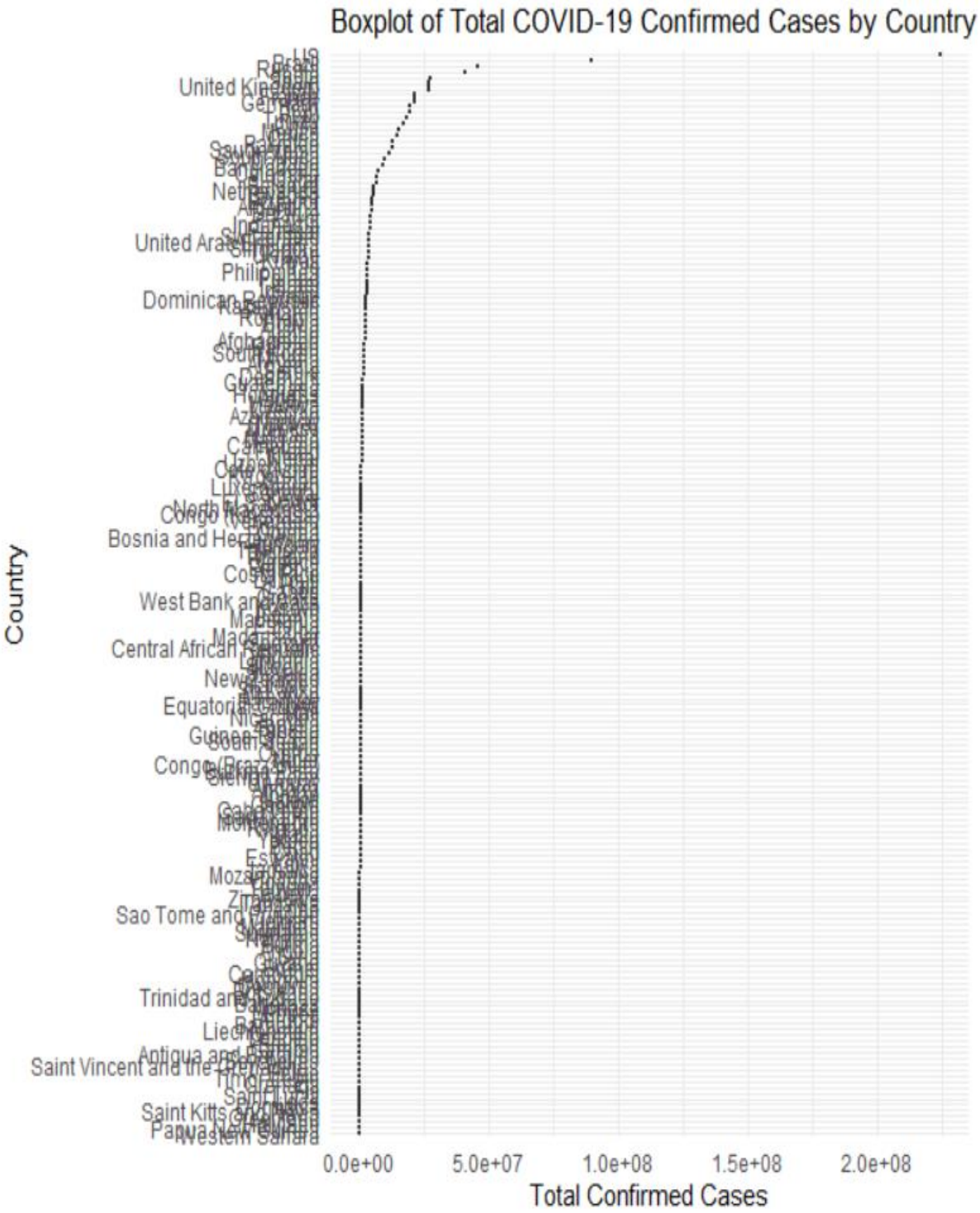

Fig. 01

Histogram of Total COVID-19 Confirmed Cases

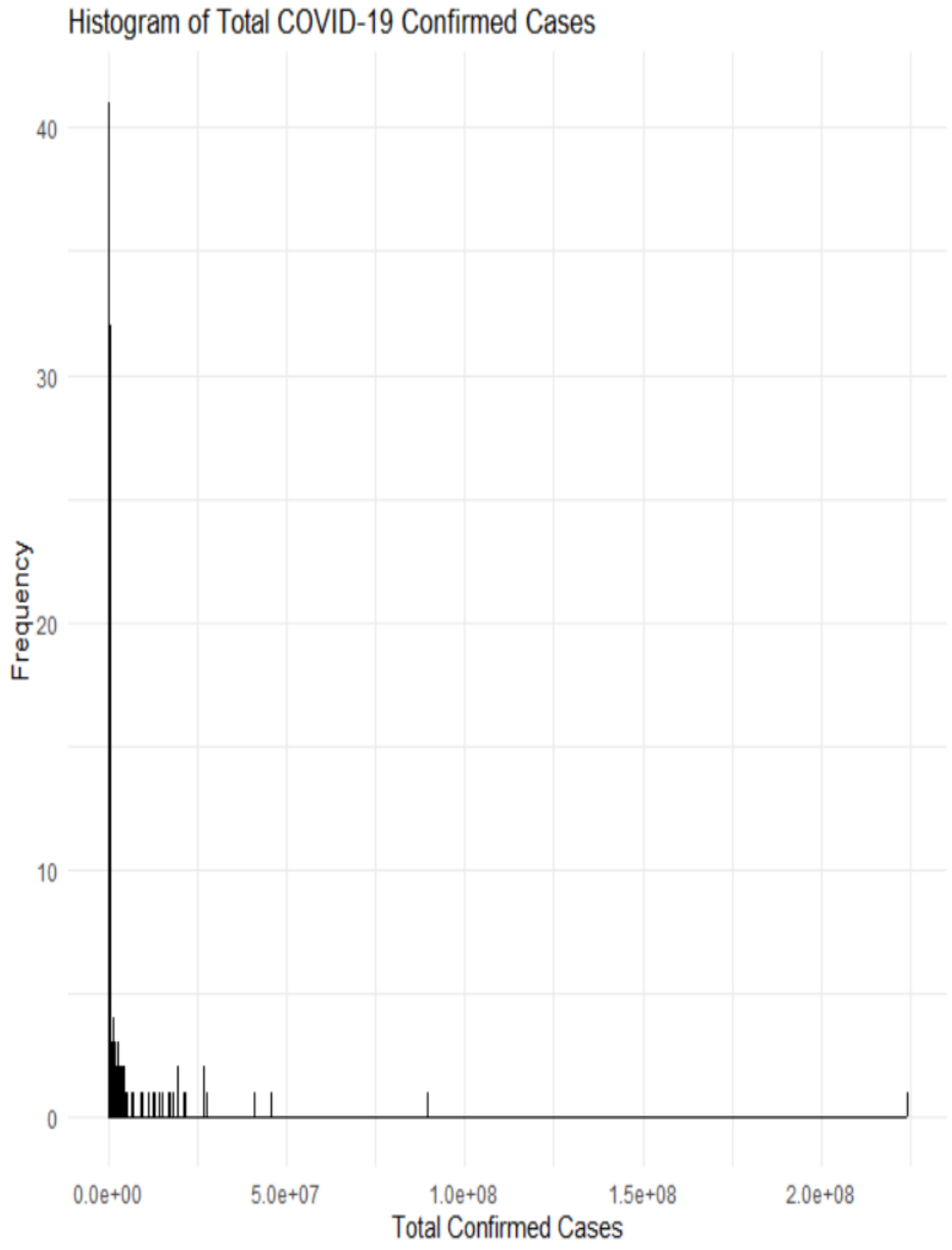

Fig. 02

Filtering

```
> # Filter for countries with more than 1 million total confirmed cases
> high_cases <- covid_aggregated %>%
+   filter(total_confirmed > 1e+06)
>
> # View the filtered data
> head(high_cases)
# A tibble: 6 × 5
  Country.Region total_confirmed total_deaths total_recovered total_active
  <chr>                   <int>        <int>           <int>        <int>
1 Afghanistan           1936390        49098          798240      1089052
2 Algeria               1179755        77972          755897       345886
3 Argentina             4450658        97749         1680024      2672885
4 Armenia               1587173        27089          857482       702602
5 Austria               2034986        71390         1638380       325216
6 Azerbaijan            1134717        14282          703402       417033
```

Sorting

```
> # Sort the aggregated data by total confirmed cases
> sorted_covid <- covid_aggregated %>%
+   arrange(desc(total_confirmed))  # Use total_confirmed here
>
> # View the sorted data
> head(sorted_covid)
# A tibble: 6 × 5
  Country.Region total_confirmed total_deaths total_recovered total_active
  <chr>                   <int>        <int>           <int>        <int>
1 US                  224345948     11011411        56353416    156981121
2 Brazil               89524967      3938034        54492873     31094060
3 Russia               45408411       619385        25120448     19668578
4 India                40883464      1111831        23783720     15987913
5 Spain                27404045      3033030        15093583      9277432
6 United Kingdom       26748587      3997775          126217     22624595
```

5. Learning outcomes (What I have learnt):

- Learn how to install and load R packages for data manipulation and visualization.
- Gain experience in importing datasets in R from local files.
- Understand how to explore the structure and summary statistics of a dataset.
- Practice data aggregation techniques using dplyr to summarize data by group.
- Enhance skills in creating meaningful visualizations (boxplots, histograms) to uncover patterns in the data.
- Learn how to filter and sort data based on specific criteria for more targeted analysis.