SIMPLE EXPLORATORY DATA ANALYSIS OF COVID-19 USING GGPLOT2

A PROJECT REPORT

Submitted by

MITHRAVASAN VBH (2303811724321065)

in partial fulfillment of requirements for the award of the course

AGI1252-FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE, 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **" SIMPLE EXPLORATORY DATA ANAYSIS OF COVID-19 DATA USING GGPLOT2"** is the bonafide work of **VBH MITHRAVASAN (2303811724321065)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

HEAD OF THE DEPARTMENT,

Department of Artificial Intelligence,
K. Ramakrishnan College of Technology,
Samayapuram, Trichy -621 112.

SIGNATURE

Mrs. S.MURUGAVALLI, AP/AI

SUPERVISOR,

Department of Artificial Intelligence,
K. Ramakrishnan College of Technology,
Samayapuram, Trichy -621 112.

Submitted for the viva-voce examination held on **02.06.2025**

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report on "**SIMPLE EXPLORATORY DATA ANALYSIS OF COVID-19 DATA USING GGPLOT2**" is the result of original work done by me and best of my knowledge, similar work has not been submitted to "**ANNA UNIVERSITY CHENNAI**" for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfillment of the requirement of the award of the **AGI1252 – FUNDAMENTALS OF DATA SCIENCE USING R**

Signature

MITHRAVASAN VBH

Place: Samayapuram

Date: 02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express our deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.

2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.

3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.

4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.

5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.

6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws.

8. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

9. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.

10. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

11. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

ABSTRACT

The outbreak of COVID-19 has presented unprecedented challenges across the globe, prompting the need for thorough data analysis to understand its spread, impact, and trends. Exploratory Data Analysis (EDA) plays a crucial role in summarizing the main characteristics of data and gaining insights before formal modeling. This study focuses on conducting a simple EDA of COVID-19 data using the powerful visualization capabilities of the ggplot2 package in R.Using publicly available COVID-19 datasets, we examine key variables such as the number of confirmed cases, recoveries, deaths, and active cases across various countries and over time. The analysis involves cleaning and preprocessing the data to handle missing values, standardizing formats, and filtering relevant fields. Once the data is structured, we employ ggplot2 to create a series of informative visualizations.Line charts are used to track the evolution of cases over time, helping identify peaks, trends, and recovery periods. Bar plots illustrate country-wise comparisons in case counts and mortality rates. Scatter plots explore relationships between variables such as population density and infection rates. Additionally, heatmaps highlight regional variations, offering a geographical perspective of the pandemic's impact.The use of ggplot2 enhances the interpretability of complex datasets through aesthetically pleasing and customizable plots.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATA SCIENCE USING R PROGRAMMING FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
The outbreak of COVID-19 has presented unprecedented challenges across the globe, prompting the need for thorough data analysis to understand its spread, impact, and trends. Exploratory Data Analysis (EDA) plays a crucial role in summarizing the main characteristics of data and gaining insights before formal modeling. This study focuses on conducting a simple EDA of COVID-19 data using the powerful visualization capabilities of the ggplot2 package in R.Using publicly available COVID-19 datasets, we examine key variables such as the number of confirmed cases, recoveries, deaths, and active cases across various countries and over time	**PO1 -3** **PO2 -3** **PO3 -3** **PO5 -3** **PO9 -2**	**PSO1 -3** **PSO2 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Introduction

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has significantly impacted global health, economy, and daily life. Since its emergence in late 2019, researchers and policymakers have relied heavily on data to monitor and understand the spread and consequences of the virus. With the increasing availability of public COVID-19 datasets, data analysis has become a key tool in tracking the pandemic's progress. Exploratory Data Analysis (EDA) serves as the foundation for any data science project, helping to uncover trends, detect anomalies, and summarize main characteristics of datasets. In this study, we use the R programming language and its ggplot2 package to conduct a simple yet effective EDA of COVID-19 data.

1.2 Objective

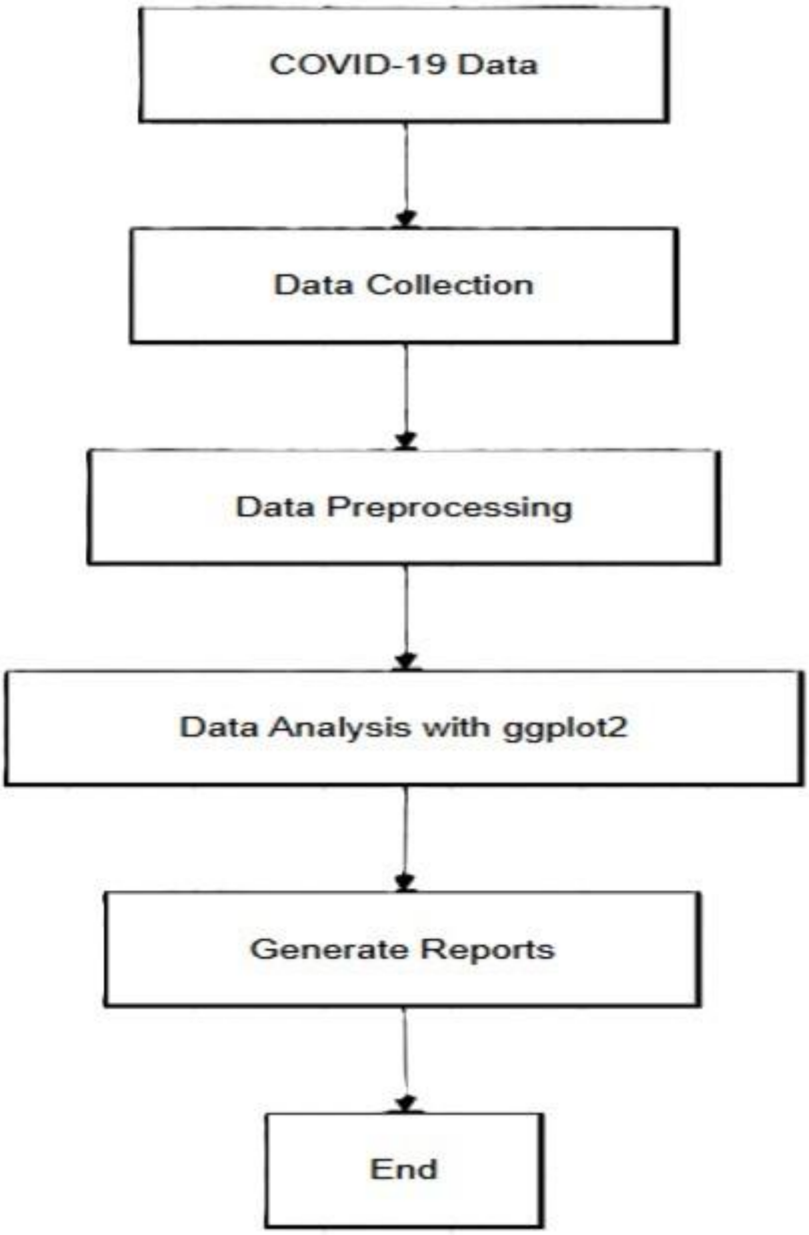
The primary objective of this project is to perform a simple Exploratory Data Analysis (EDA) of COVID-19 data using the ggplot2 package in R. The analysis aims to uncover significant patterns, trends, and relationships within the dataset to better understand the spread and impact of the pandemic. Specific goals include visualizing daily and cumulative case trends, comparing country-wise statistics, and analyzing the distribution of confirmed, recovered, and death cases over time. Another objective is to make complex data more understandable through visual tools, thus supporting data-driven decisions. The use of ggplot2 enables the creation of clear, customizable, and interactive graphics that can highlight insights not immediately apparent through raw data alone

CHAPTER 2
PROJECT METHODOLOGY

2.1 Proposed Work

- **Data Collection**:
  - COVID-19 data is obtained from public sources such as the Johns Hopkins repository, WHO datasets, or Kaggle CSV files.
- **Data Preprocessing**:
  - The raw data is cleaned to handle missing values, convert date formats, and filter necessary columns for analysis.
- **Data Exploration**:
  - Initial investigation using R functions like summary(), str(), and table() to understand the dataset structure and distribution.
- **Data Visualization with ggplot2**:
  - Visual plots (line, bar, scatter, and heatmaps) are created to explore trends, regional differences, and correlations between variables.
- **Insight Generation**:
  - Key insights are derived such as case peaks, recovery trends, and mortality rates. Comparative analysis is performed across countries and time periods.
- **Conclusion & Reporting**:
  - Final interpretations are drawn, and the outcomes are presented in a report or dashboard, highlighting the importance of data in understanding pandemics.

2.2 Block Diagram



CHAPTER 3
R PROGRAMMING CONCEPTS

3.1 Data Manipulation With Dplyr Package

The dplyr package is one of the most efficient tools in R for data manipulation. It simplifies and speeds up common data transformation tasks such as filtering, selecting, arranging, mutating, and summarizing. In this COVID-19 data analysis project, dplyr was used extensively to clean and prepare the data for visualization. The pipe operator %>% allows for chaining multiple functions, making code more readable and organized. For example, the dataset was filtered by country, grouped by date, and summarized to compute daily cases using just a few lines of code. This modular approach ensures that each transformation step is clear and easy to debug. dplyr also supports working with large datasets efficiently, which is crucial when analyzing real-world COVID-19 case data spanning hundreds of thousands of records.

3.2 Data Visualization With Ggplot2 Package

The ggplot2 package is an essential part of data visualization in R. It follows the grammar of graphics framework, allowing users to build complex plots using a layered approach. In this project, ggplot2 was used to create various charts including line graphs, bar plots, and scatter plots to visualize trends in COVID-19 data such as confirmed cases, deaths, and recoveries across different countries and dates. The use of aes() to map variables and geom_*() functions to add layers makes it flexible and highly customizable. For example, geom_line() was used to draw daily case trends, while facet_wrap() helped in creating multi-panel plots for comparing countries. Titles, labels, and color schemes were easily added to improve clarity.

CHAPTER 4

MODULE DESCRIPTION

4.1 Data Collection

This module focuses on collecting COVID-19 data from trusted sources such as Johns Hopkins University and WHO. The data is downloaded in CSV format and includes key variables like confirmed cases, deaths, recoveries, and dates across various countries.

4.2 Data Cleaning And Preprocessing

Raw data often contains missing values, inconsistencies, or redundant entries. This module involves cleaning the dataset by handling missing values, converting date formats, and organizing columns. Functions from base R and the dplyr package are used.

4.3 Data Manipulation

This module includes transforming the data using dplyr functions such as filter(), mutate(), group_by(), and summarise() to prepare subsets for analysis. Grouped summaries and new variables (e.g., daily cases) are calculated here.

4.4 Statistical Summary

Descriptive statistics such as mean, median, standard deviation, and quantiles are computed. This module helps to understand the distribution and spread of COVID-19 cases over time and across

4.5 Data Visualization

. Using ggplot2, various plots such as line charts, bar graphs, and scatter plots are created to visually interpret the trends in the data. These visualizations provide actionable insights and support effective storytelling through data.

CHAPTER 5
CONCLUSION

This project effectively demonstrates how **R programming** combined with the power of the **ggplot2** visualization package can be utilized for conducting **Exploratory Data Analysis (EDA)** on real-world datasets, particularly COVID-19 statistics. From data collection to transformation, visualization, and interpretation, each step was handled using efficient R tools and packages such as dplyr, tidyr, and ggplot2.The analysis offered key insights into the spread and trends of COVID-19 cases across countries and over time. Through grouped summaries and visual plots, patterns in confirmed cases, recoveries, and fatalities were clearly understood. These visualizations made complex information accessible and comprehensible to both technical and non-technical audiences.One of the key strengths of this project was its modular and reproducible workflow. Each component — from data cleaning to plotting — was implemented using reusable R functions, supporting extensibility and real-time analysis. Furthermore, the use of open datasets and open-source tools emphasizes transparency and collaboration in data science.The findings highlight the critical importance of **data-driven decision-making** during public health crises. The ability to visualize and interpret large datasets in an efficient manner enables faster and more informed responses from policymakers and health organizations.In conclusion, this project not only strengthens understanding of R programming and data visualization techniques but also demonstrates the value of EDA in solving real-world problems. It encourages further exploration into predictive modeling and machine learning approaches for more advanced pandemic analysis and forecasting

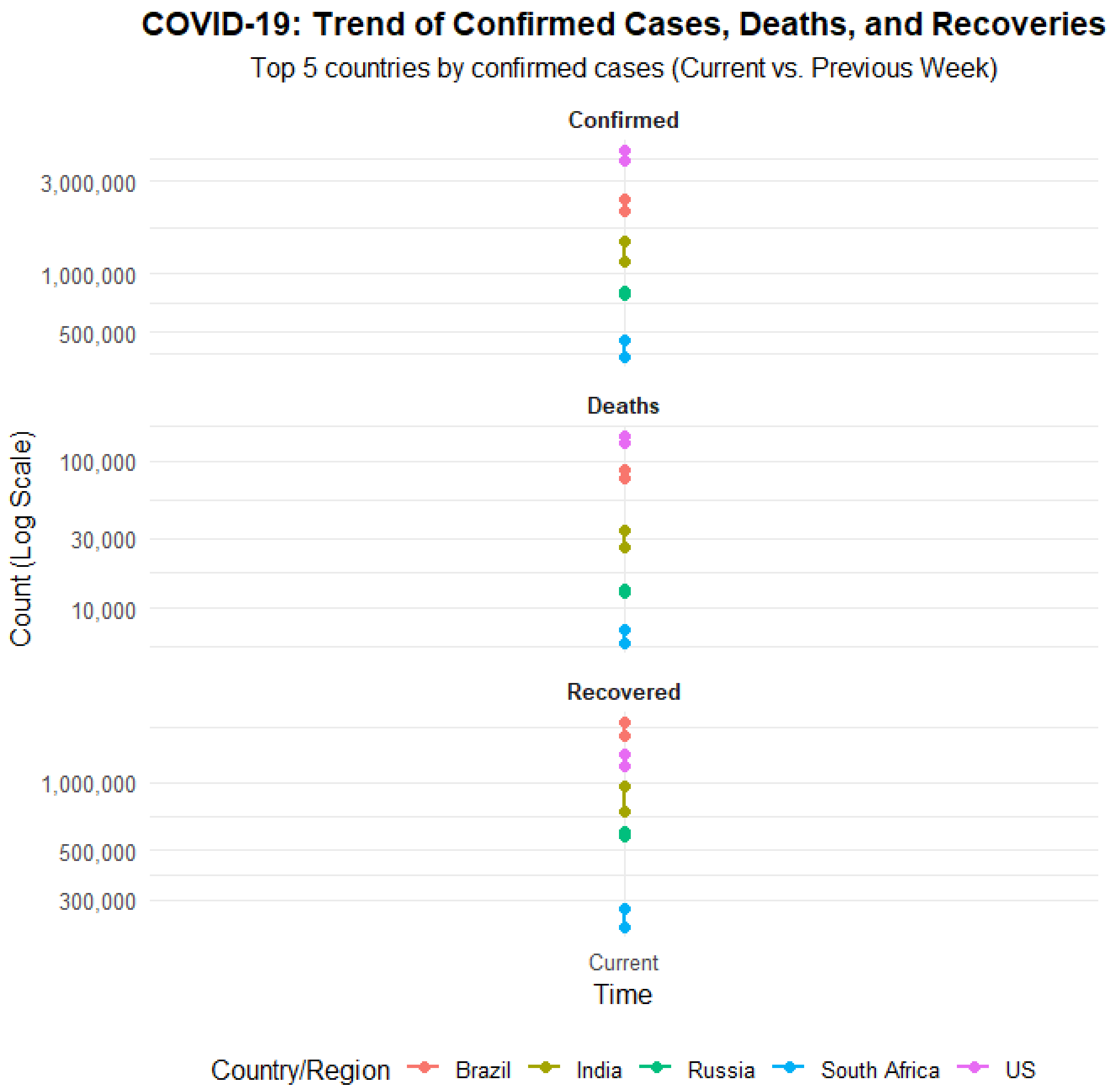
APPENDICES

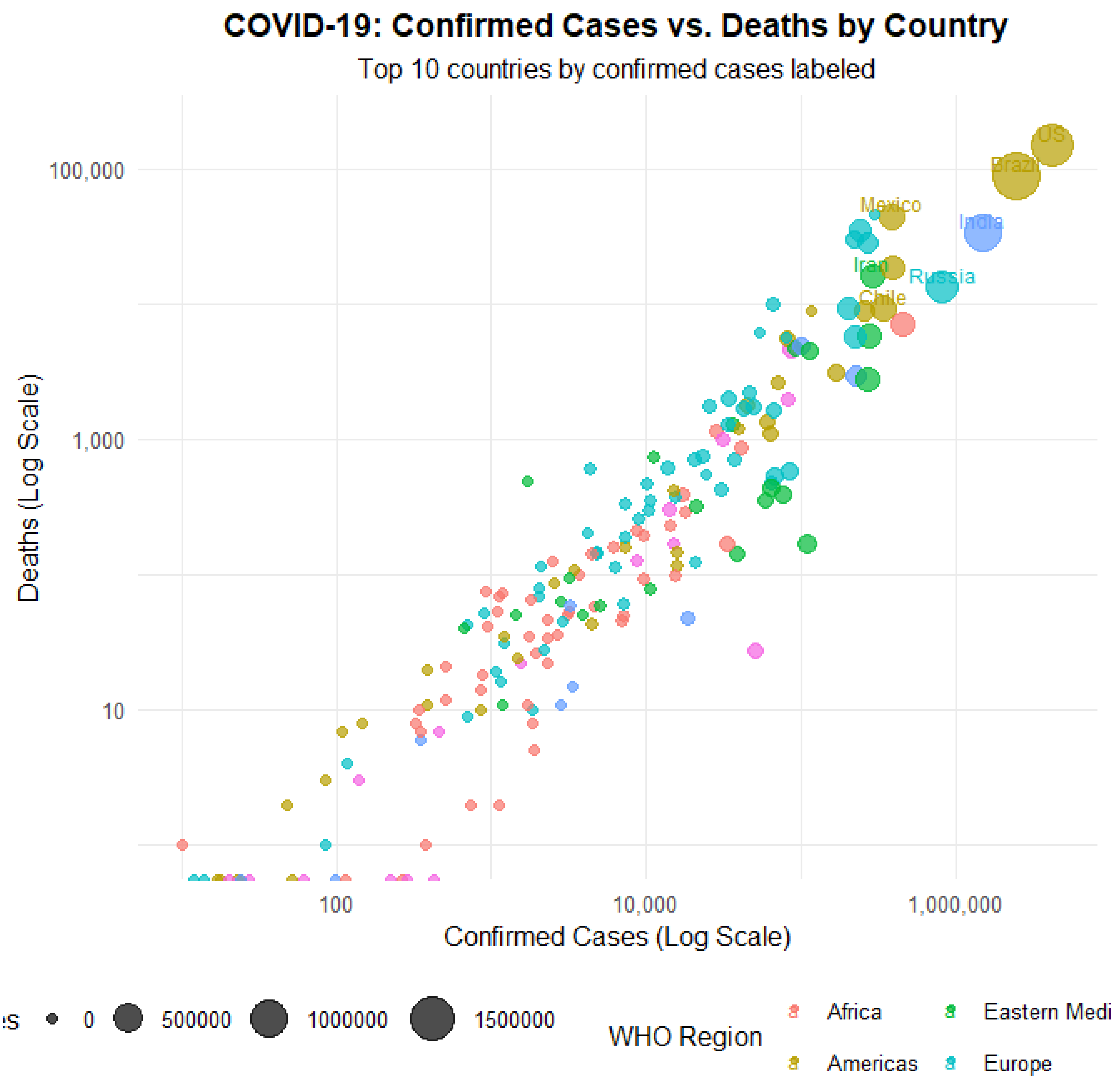
APPENDIX A – SOURCE CODE

```
library(readr)
library(ggplot2)
library(dplyr)
library(tidyr)

# Set the file path (adjust if necessary)
file_path <- "C:/Users/ADMIN/OneDrive/Documents/country_wise_latest.csv"

# Read the CSV file
covid_data <- read_csv(file_path)

# Sample summary statistics
summary_stats <- covid_data %>%
  summarise(
    Total_Confirmed = sum(Confirmed, na.rm = TRUE),
    Total_Deaths = sum(Deaths, na.rm = TRUE),
    Total_Recovered = sum(Recovered, na.rm = TRUE),
    Mean_Death_Rate = mean(`Deaths / 100 Cases`, na.rm = TRUE),
    Max_Confirmed = max(Confirmed, na.rm = TRUE)
  )

# Print summary statistics
print("Summary Statistics:")
print(summary_stats)
```

```
# Identify top 5 countries by confirmed cases
top_countries <- covid_data %>%
  arrange(desc(Confirmed)) %>%
  select(`Country/Region`, Confirmed, Deaths, Recovered) %>%
  head(5)
print("Top 5 Countries by Confirmed Cases:")
print(top_countries)

# Scatter plot with labels for top 10 countries
top_10_countries <- covid_data %>%
  arrange(desc(Confirmed)) %>%
  select(`Country/Region`) %>%
  head(10) %>%
  pull(`Country/Region`)

p_scatter <- ggplot(covid_data, aes(x = Confirmed, y = Deaths, size = Recovered, color = `WHO Region`)) +
  geom_point(alpha = 0.7) +
  geom_text(
    data = covid_data %>% filter(`Country/Region` %in% top_10_countries),
    aes(label = `Country/Region`),
    size = 3, nudge_y = 0.1, check_overlap = TRUE
  ) +
  scale_x_log10(labels = scales::comma) +
  scale_y_log10(labels = scales::comma) +
  scale_size_continuous(range = c(2, 10)) +
  labs(
    title = "COVID-19: Confirmed Cases vs. Deaths by Country",
    subtitle = "Top 10 countries by confirmed cases labeled",
    x = "Confirmed Cases (Log Scale)",
    y = "Deaths (Log Scale)",
    size = "Recovered Cases",
    color = "WHO Region"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5)
  )

# Display and save scatter plot
print(p_scatter)
ggsave("covid_scatter_labeled.png", plot = p_scatter, width = 8, height = 6)

# Prepare data for line chart (time-over-trend analysis)
# Select top 5 countries for line chart
top_5_countries <- top_countries %>% pull(`Country/Region`)

# Create a dataset with current and previous week data
# Estimate previous week deaths and recoveries assuming proportional change
time_data <- covid_data %>%
  filter(`Country/Region` %in% top_5_countries) %>%
  select(`Country/Region`, Confirmed, Deaths, Recovered, `Confirmed last week`, `1
week change`) %>%
  mutate(
    # Previous week confirmed cases
    Prev_Confirmed = `Confirmed last week`,
    # Estimate previous week deaths and recoveries (proportional to confirmed cases
change)
    Prev_Deaths = Deaths * (`Confirmed last week` / Confirmed),
    Prev_Recovered = Recovered * (`Confirmed last week` / Confirmed)
  ) %>%
  select(`Country/Region`, Confirmed, Deaths, Recovered, Prev_Confirmed,
Prev_Deaths, Prev_Recovered) %>%
  pivot_longer(
    cols = c(Confirmed, Deaths, Recovered, Prev_Confirmed, Prev_Deaths,
Prev_Recovered),
    names_to = c("Metric", "Time"),
    names_pattern = "(Confirmed|Deaths|Recovered)(.*)",
    values_to = "Value"
  ) %>%
  mutate(
    Time = if_else(Time == "", "Current", "Previous Week"),
    Metric = factor(Metric, levels = c("Confirmed", "Deaths", "Recovered"))
  )

# Create line chart
p_line <- ggplot(time_data, aes(x = Time, y = Value, color = `Country/Region`, group =
`Country/Region`)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  facet_wrap(~ Metric, scales = "free_y", ncol = 1) +
  scale_y_log10(labels = scales::comma) +
  labs(
    title = "COVID-19: Trend of Confirmed Cases, Deaths, and Recoveries",
    subtitle = "Top 5 countries by confirmed cases (Current vs. Previous Week)",
    x = "Time",
    y = "Count (Log Scale)",
    color = "Country/Region"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5),
    strip.text = element_text(face = "bold")
  )

# Display and save line chart
print(p_line)
ggsave("covid_trend_line.png", plot = p_line, width = 8, height = 8)
```

APPENDIX B - SCREENSHOTS



COVID-19: Confirmed Cases vs. Deaths by Country
Top 10 countries by confirmed cases labeled
Deaths (Log Scale)
100,000
1,000
10
100
10,000
1,000,000
Confirmed Cases (Log Scale)
US
Brazi
Mexico
India
Iran
Russia
Chile
0
500000
1000000
1500000
WHO Region
Africa
Eastern Medi
Americas
Europe

REFERENCES:

BOOK

1. Wickham, H., & Grolemund, G. (2016). *R for Data Science*. O'Reilly Media. https://r4ds.had.co.nz

2. Grolemund, G. (2014). *Hands-On Programming with R*. O'Reilly Media.

3. Teetor, P. (2011). *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. O'Reilly Media.

4. Lander, J. P. (2017). *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley.

ONLINE TUTORIALS AND DOCUMENTATION

1. R Programming for Beginners - Full Course by freeCodeCamp
   https://www.youtube.com/watch?v=_V8eKsto3Ug
2. ggplot2 Tutorial - Data Visualization with ggplot2 in R (DataCamp)
   https://www.youtube.com/watch?v=HeqHMM4ziXA
3. Exploratory Data Analysis in R (DataCamp)
   https://www.datacamp.com/courses/exploratory-data-analysis-in-r
4. Tidyverse Official Site – R packages for data science
   https://www.tidyverse.org/
5. ggplot2 Documentation
   https://ggplot2.tidyverse.org/
6. R Documentation
   https://www.rdocumentation.org/
7. R Project Official Site
   https://www.r-project.org/

COVID-19 DATA SOURCES

1. Johns Hopkins University COVID-19 Data Repository
   https://github.com/CSSEGISandData/COVID-19

2. World Health Organization (WHO) Dashboard
   https://covid19.who.int/

3. Kaggle Datasets - COVID-19
   https://www.kaggle.com/datasets

4. CDC COVID Data Tracker
   https://covid.cdc.gov/covid-data-tracker

ARTICLES & BLOGS

1. Hadley Wickham (2011). *The Split-Apply-Combine Strategy for Data Analysis*. Journal of Statistical Software. https://www.jstatsoft.org/
2. *Towards Data Science* - Blogs on R and COVID-19 Analysis
   https://towardsdatascience.com/

R-bloggers – Latest articles and tutorials on R
https://www.r-bloggers.com/