Progressive Education Society's

**Modern College of Engineering**

**MCA Department**

**A.Y. 2024-25**

**Subject Code: 410908: Data Science Laboratory**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Class: SY MCA Div: A Batch: S3 Roll Number:52062

Name: Laxman Shinde Assignment No: 06 Date of Implementation: 17/10/2024

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Q. **Design and Develop real-time Data Science  
Application (e.g. Image Recognition/ Intelligent**  
  
**Assistant/ Recommendation System/ Fake News  
Detection/Emotion Recognition/Chatbot/Other)**

Project Introduction :

**1. Investing Coding 19 virus Trends**

**Overview**

Leveraging R and the powerful dplyr library, you'll manipulate, filter, and aggregate a comprehensive dataset containing information on COVID-19 cases, tests, and hospitalizations across different countries. By applying data wrangling techniques such as grouping and summarizing, you'll uncover which countries have the highest rates of positive COVID-19 tests relative to their testing numbers. This hands-on project will not only strengthen your R programming skills and analytical thinking but also provide valuable experience in deriving actionable insights from real-world health data – a crucial skill in today's data-driven healthcare landscape.

**Tools and Technologies**

* R
* dplyr
* readr
* tibble

**Prerequisites**

To successfully complete this project, you should be comfortable with [data structures in R](https://www.dataquest.io/course/datastructure-in-r-rewrite/) such as:

* Creating and working with vectors, matrices, and lists in R
* Indexing data structures to extract elements for analysis
* Applying functions to data structures to perform calculations
* Manipulating and analyzing data using dataframes

**Step-by-Step Instructions**

1. Load and explore the COVID-19 dataset using readr and tibble
2. Filter and select relevant data using dplyr functions
3. Aggregate data by country and calculate summary statistics
4. Identify top countries by testing numbers and positive case ratios
5. Create vectors and matrices to store key findings
6. Compile results into a comprehensive list structure

**Expected Outcomes**

* Analyzing a real-world COVID-19 dataset using R and dplyr
* Applying data manipulation techniques to filter and aggregate data
* Identifying trends and insights from data using grouping and summarizing
* Creating and manipulating different R data structures (vectors, matrices, lists)
* Interpreting results to answer specific questions about COVID-19 testing and positive rates

**Code :**

# Understanding the Data

## Loading the dataset from the `covid19.csv` CSV file and quick explorati

library(readr)

# Loading the dataset

covid\_df <- read\_csv("covid19.csv")

# Displaing the dimension of the data:

dim(covid\_df)

# Storing the column names in a variable

vector\_cols <- colnames(covid\_df)

# Displaing the variable vector\_cols

vector\_cols

# Showing the first few rows of the dataset

head(covid\_df)

# Showing a global view of the dataset.

library(tibble)

glimpse(covid\_df)

library(dplyr)

# Filter the "All States" Province states and remove the `Province\_State` column

covid\_df\_all\_states <- covid\_df %>%

filter(Province\_State == "All States") %>%

select(-Province\_State)

# Selecting the columns with cumulative numbers

covid\_df\_all\_states\_daily <- covid\_df\_all\_states %>%

select(Date, Country\_Region, active, hospitalizedCurr, daily\_tested, daily\_positive)

head(covid\_df\_all\_states\_daily)

# Extracting the Top Ten countries in the number of tested cases

## Su

covid\_df\_all\_states\_daily\_sum <- covid\_df\_all\_states\_daily %>%

group\_by(Country\_Region) %>%

summarise(tested = sum(daily\_tested),

positive = sum(daily\_positive),

active = sum(active),

hospitalized = sum(hospitalizedCurr)) %>%

arrange(desc(tested)) #this is equivalent to `arrange(-tested)`

covid\_df\_all\_states\_daily\_sum

covid\_top\_10 <- head(covid\_df\_all\_states\_daily\_sum, 10)

covid\_top\_10

countries <- covid\_top\_10$Country\_Region

tested\_cases <- covid\_top\_10$tested

positive\_cases <- covid\_top\_10$positive

active\_cases <- covid\_top\_10$active

hospitalized\_cases <- covid\_top\_10$hospitalized

names(positive\_cases) <- countries

names(tested\_cases) <- countries

names(active\_cases) <- countries

names(hospitalized\_cases) <- countries

positive\_cases

sum(positive\_cases)

mean(positive\_cases)

positive\_cases/sum(positive\_cases)

positive\_cases/tested\_cases

positive\_tested\_top\_3 <- c("United Kingdom" = 0.11, "United States" = 0.10, "Turkey" = 0.08)

# Creating vectors

united\_kingdom <- c(0.11, 1473672, 166909, 0, 0)

united\_states <- c(0.10, 17282363, 1877179, 0, 0)

turkey <- c(0.08, 2031192, 163941, 2980960, 0)

# Creating the matrix covid\_mat

covid\_mat <- rbind(united\_kingdom, united\_states, turkey)

# Naming columns

colnames(covid\_mat) <- c("Ratio", "tested", "positive", "active", "hospitalized")

#d Displaying the matrix

covid\_mat

question <- "Which countries have had the highest number of positive cases against the number of tests?"

answer <- c("Positive tested cases" = positive\_tested\_top\_3)

datasets <- list(

original = covid\_df,

allstates = covid\_df\_all\_states,

daily = covid\_df\_all\_states\_daily,

top\_10 = covid\_top\_10

)

matrices <- list(covid\_mat)

vectors <- list(vector\_cols, countries)

data\_structure\_list <- list("dataframe" = datasets, "matrix" = matrices, "vector" = vectors)

covid\_analysis\_list <- list(question, answer, data\_structure\_list)

covid\_analysis\_list[[2]]

**Output :**









