1. Create a directory “Covid19Data” in your PC and set the working directory in R to this directory using setwd().

This command is setting the working directory to “Covid19Data”, which contains the CSV files covid\_19\_india.csv, covid\_vaccine\_statewise.csv, StatewiseTestingDetails.csv.

**setwd('Covid19Data')**

A screenshot of a computer

Description automatically generated

1. Save all data files “covid\_19\_india.csv”, “covid\_vaccine\_statewise.csv” and “StatewiseTestingDetails.csv”, which are provided as part of this assessment into the “Covid19Data” directory. Run the command library(tidyverse) in R before moving to (3).

I had already installed the tidyverse library, and with this command I am running it.

**library(tidyverse)**

A screenshot of a computer

Description automatically generated

1. Read the data files “covid\_19\_india.csv”, “covid\_vaccine\_statewise.csv” and “StatewiseTestingDetails.csv” into three tibbles (a modern take on data frames) in R using read\_csv() function. Name the tibbles as follows: indiaCovid, indiaCovidVaccine and indiaCovidTesting. [ ].

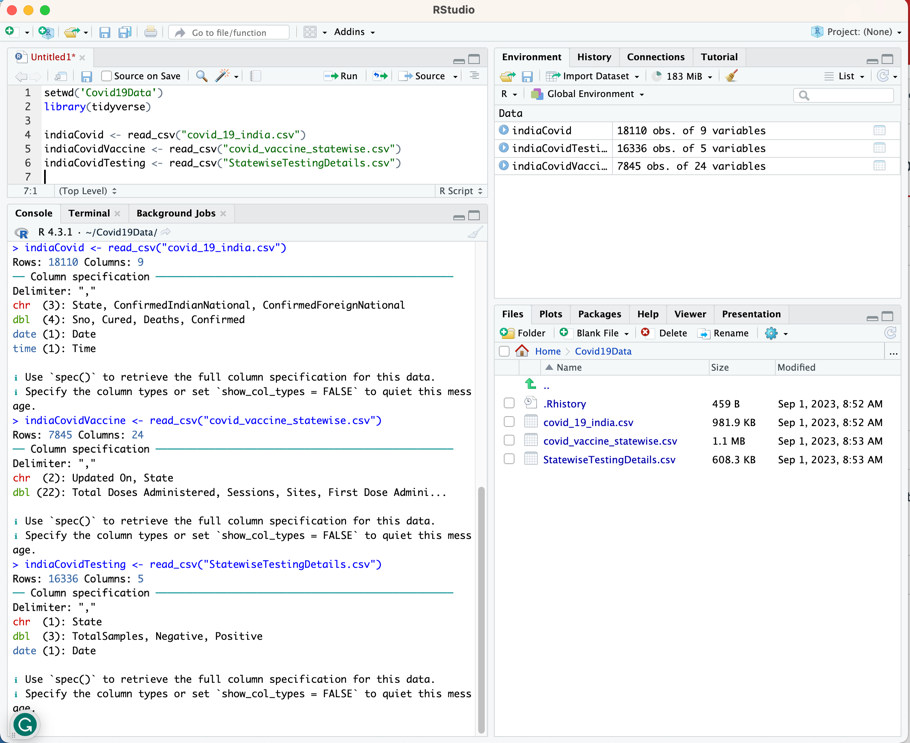
These commands imports the data from each CSV file into a tibble. Spec() shows the data types.

**indiaCovid <- read\_csv("covid\_19\_india.csv")**

**indiaCovidVaccine <- read\_csv("covid\_vaccine\_statewise.csv")**

**indiaCovidTesting <- read\_csv("StatewiseTestingDetails.csv")**

**spec(indiaCovid)**



1. Now, calculate the total number of confirmed cases (people affected by the novel coronavirus), tests (Total Samples collected) and vaccinations (First Dose Administered) conducted in India from these datasets. Some of the columns in these tibbles have running total values state-wise.

The code calculates national totals for confirmed cases, testing, and vaccinations from state-level data in 3 steps:

1. Group the data by State and filter to the most recent row for each state using top\_n(1, column). This gets the latest value per state.
2. Pull just the column of interest from the filtered data into a vector using pull(column). This extracts the state values.
3. Sum the vector using sum() to calculate the national total by aggregating the state-level values.

Screenshot 1 shows the calculation for total confirmed cases, while the second shows the printed totals for all categories.

**# Get total confirmed per state**

**confirmed\_states <- indiaCovid %>%**

**group\_by(State) %>%**

**top\_n(1, Confirmed) %>%**

**pull(Confirmed)**

**# Sum to get national total**

**totalConfirmed <- sum(confirmed\_states)**

**# Get total tested per state**

**tested\_states <- indiaTesting %>%**

**group\_by(State) %>%**

**top\_n(1, TotalSamples) %>%**

**pull(TotalSamples)**

**# Sum to get national total**

**TotalSamples <- sum(tested\_states)**

**# Get total vaccinated per state**

**vaccinated\_states <- indiaVaccine %>%**

**group\_by(State) %>%**

**top\_n(1, `First Dose Administered`) %>%**

**pull(`First Dose Administered`)**

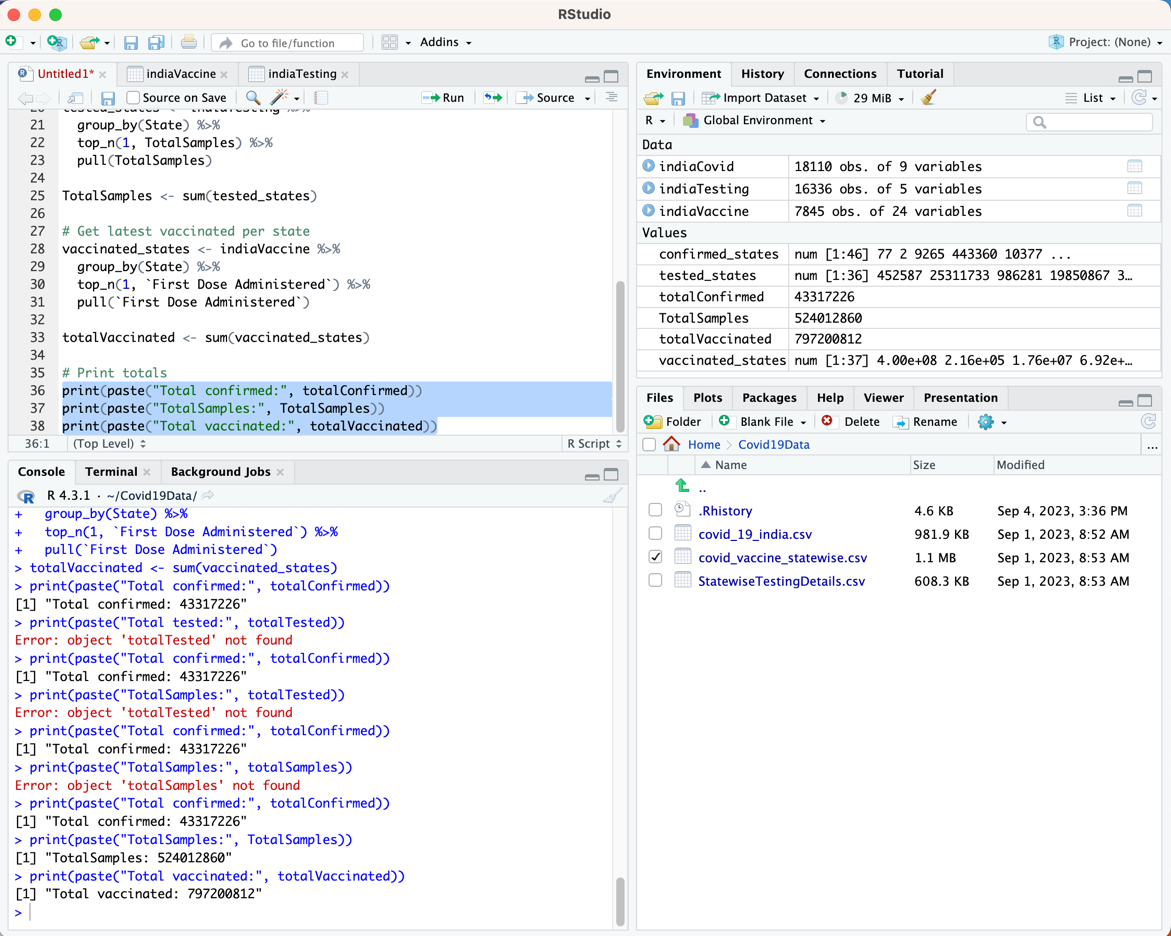
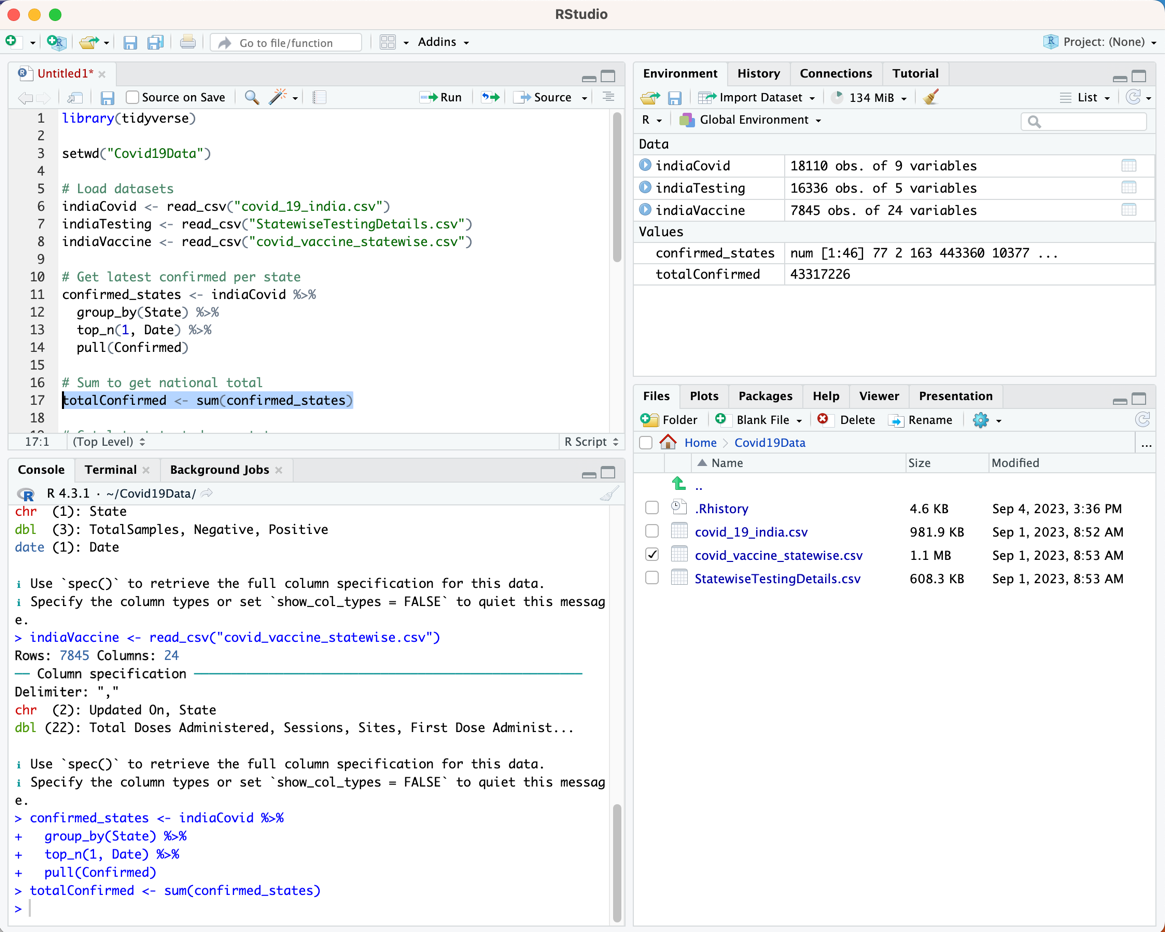
**totalVaccinated <- sum(vaccinated\_states)**

**# Print totals**

**print(paste("Total confirmed:", totalConfirmed))**

**print(paste("TotalSamples:", TotalSamples))**

**print(paste("Total vaccinated:", totalVaccinated))**



1. Now, create a data frame with the total number of confirmed cases (people affected by the novel coronavirus), tests (Total Samples collected) and vaccinations (First Dose Administered) conducted in India as follows:

This code creates a new data frame called indiaCovidData to store the category names and totals. Picture one shows the code and results, while picture two shows the stored values.

**# Calculate totals**

**totalConfirmed <- sum(confirmed\_states)**

**totalTested <- sum(tested\_states)**

**totalVaccinated <- sum(vaccinated\_states)**

**# Create category vector**

**categories <- c("Confirmed", "Tested", "Vaccinated")**

**# Create count vector**

**counts <- c(totalConfirmed, totalTested, totalVaccinated)**

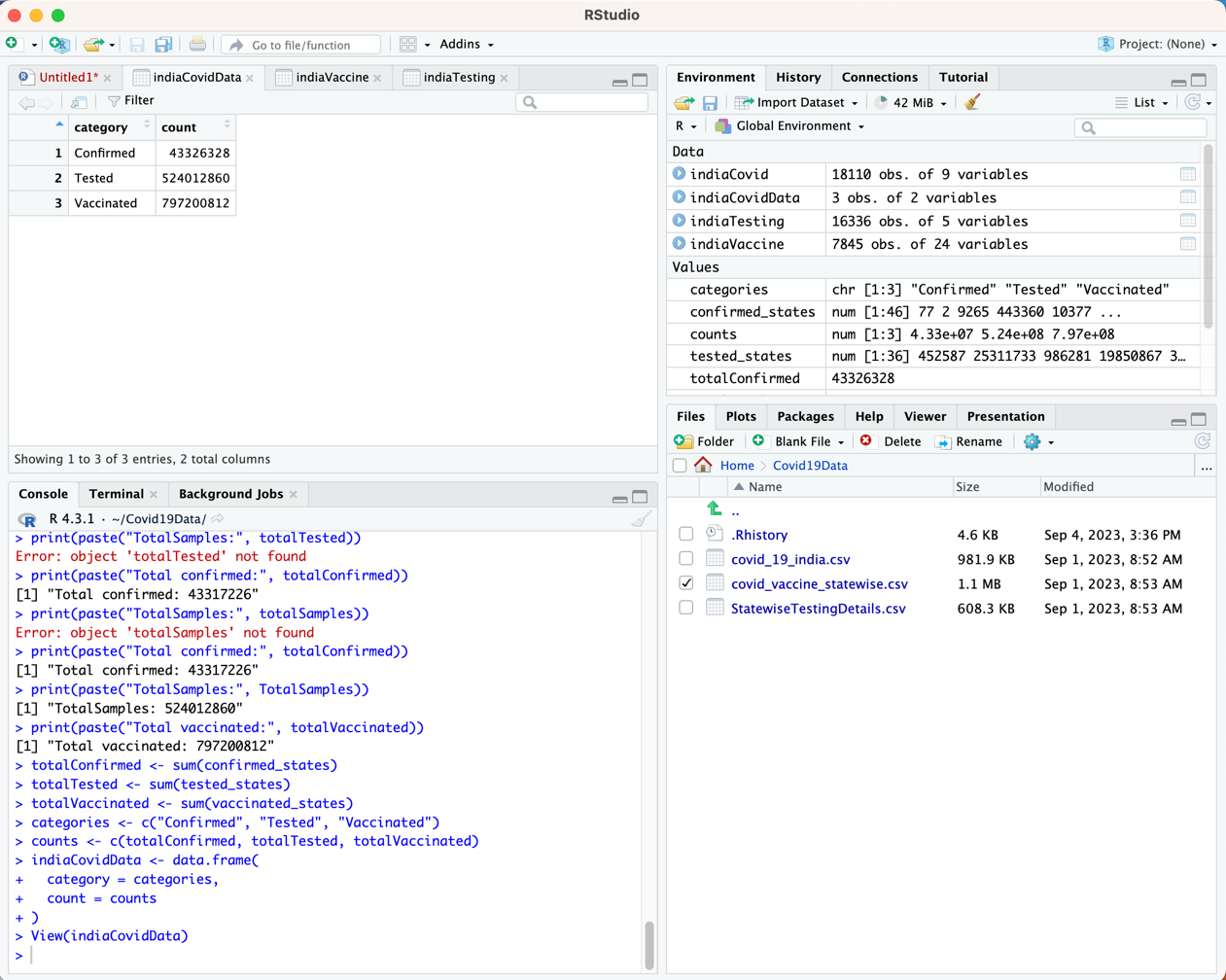
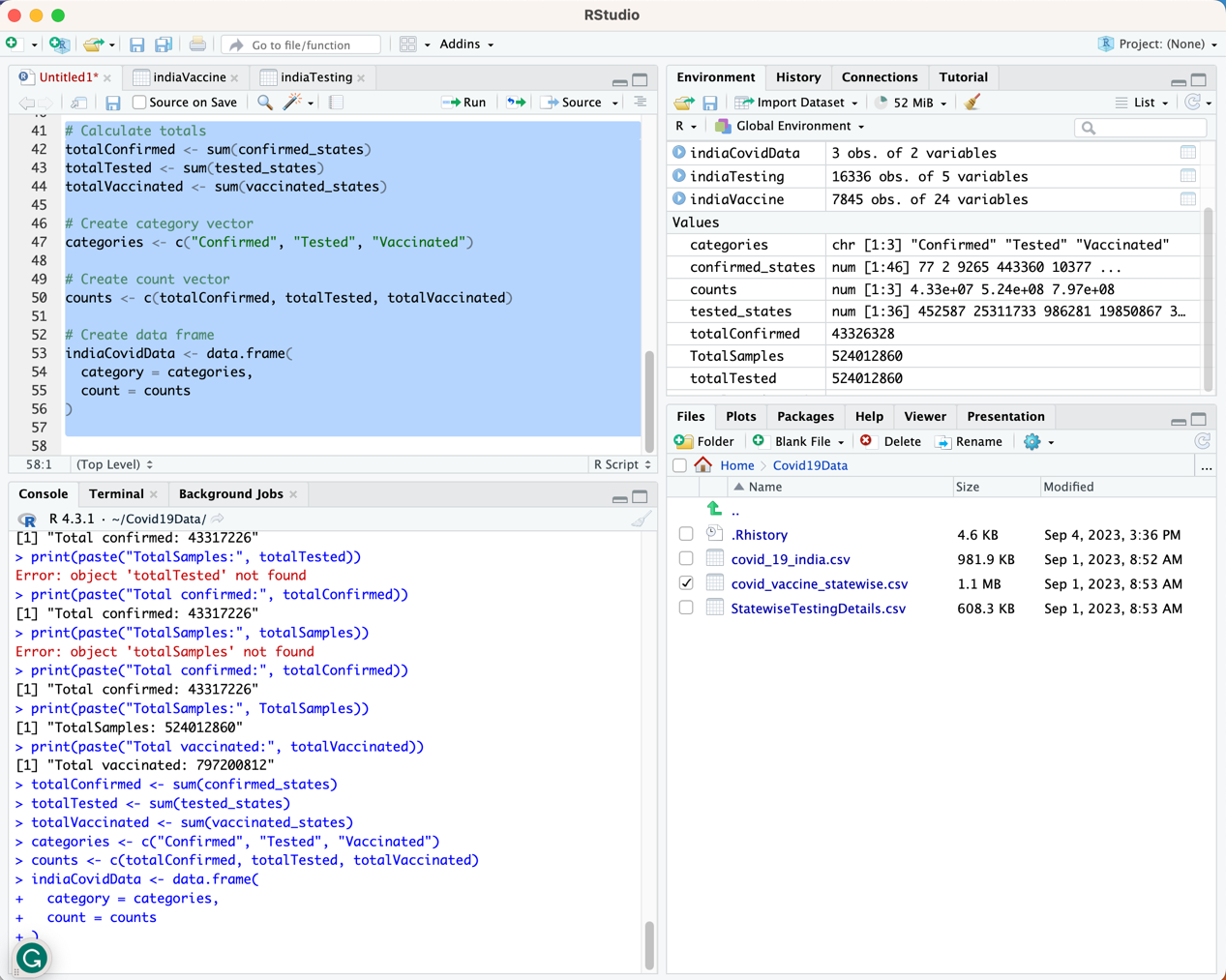
**# Create data frame**

**indiaCovidData <- data.frame(**

**category = categories,**

**count = counts**

**)**



1. Plot the data stored in the “indiaCovidData” as follows: ggplot(indiaCovidData, aes(category, count, colour=category)) + geom\_col(). [Note: Make sure you run library(ggplot2) command before ggpolt()].

This code uses ggplot2 to create a bar chart from the indiaCovidData dataframe, showcasing that the column vaccinated holds the highest number.

**# Load ggplot2**

**library(ggplot2)**

**# Create plot**

**ggplot(indiaCovidData, aes(category, count, color=category)) +**

**geom\_col()**

