Investigation: COVID-19 Virus Trends

Mariby Cruz

2025-03-13

Introduction

This project will be looking at COVID-19 Virus data trends from January 20th - June 1st, 2020. We're working with a dataset pulled from Kaggle which has the number of tests conducted over time to help make sense of how the virus is spreading in each country. To fully understand how the virus is spreading, and not be misled by only positive cases reported, the following question will be the focus of this analysis: **Which countries have reported the highest number of positive cases in relation to the number of tests conducted?**

Summary of the Results

The top 3 countries to have the highest number of positive cases relative to the number of test given were: 1. UK 2. USA 3. Turkey

The importance of looking at the positive cases in relation to tests conducted lets us comprehend the rate at which the virus spreads. This can be useful information to confirm containment efforts put in place by each country.

Exploring the Data

The data collected for each observation contains information about: date, regions, diagnostic result, and patient conditions (ie. recovered, death, hospitalized).
I took a quick look at the data frame to check if there is data that will need to be manipulated or updated using the `glimpse()` function. It's also a good time to see if there are inconsistencies in the data like misspelling or combining data when it should be its own column.

```
covid_df <- read.csv("~/R/covid19.csv")
vector_cols <- colnames(covid_df)
glimpse(covid_df)
```

```
## Rows: 10,903
## Columns: 14
## $ Date                    <chr> "2020-01-20", "2020-01-22", "2020-01-22", "202~
## $ Continent_Name          <chr> "Asia", "North America", "North America", "Nor~
## $ Two_Letter_Country_Code <chr> "KR", "US", "US", "US", "US", "KR", "US", "US"~
## $ Country_Region          <chr> "South Korea", "United States", "United States~
## $ Province_State          <chr> "All States", "All States", "Washington", "All~
## $ positive                <int> 1, 1, 1, 1, 1, 2, 1, 1, 4, 0, 3, 0, 0, 0, 0, 1~
## $ hospitalized            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ recovered               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
## $ death                  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ total_tested           <int> 4, 1, 1, 1, 1, 27, 1, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ active                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ hospitalizedCurr       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ daily_tested           <int> 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ daily_positive         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Keeping the rows we need

There were data from both 'provinces' and 'all states' under one single column. For clarity and organization, I kept data from all states and filtered out provinces from the column `Province_State` followed by removing the column. Removing the column does not remove or lose data because the column is independent of the other columns. For tracking purposes, this was stored as a new data set name.

```
covid_df_all_states <- covid_df %>%
                       filter(Province_State == "All States") %>%
                       select(-Province_State)
```

Looking more closely at how the data was collected, there was both daily observations and cumulative observations. If the daily data gets compared to cumulative data, there would be a bias and wrong conclusions can be made. To avoid this bias, the project will focus on daily observations.

```
covid_df_all_states_daily <- covid_df_all_states %>%
                             select(Date, Country_Region, active, hospitalizedCurr, daily_tested, daily_
```

Extracting the Top Ten Countries with Most Covid-19 Cases

In this exercise, I had to group the data by country and then sum up the results for each column (tested, positive, active, hospitalized), followed by ordering the results by descending order to get the highest numbers at the top. To get the top 10, I used head() and selected the first ten and stored the result into its own data frame.

```
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(-tested)
#Display the top 10 rows
covid_top_10 <- head(covid_df_all_states_daily_sum, 10)
print(covid_top_10)
```

```
## # A tibble: 10 x 5
##    Country_Region   tested positive  active hospitalized
##    <chr>             <int>    <int>   <int>        <int>
##  1 United States  17282363  1877179       0            0
##  2 Russia         10542266   406368 6924890            0
##  3 Italy           4091291   251710 6202214      1699003
```

```
##  4 India           3692851    60959       0            0
##  5 Turkey          2031192   163941 2980960            0
##  6 Canada          1654779    90873   56454            0
##  7 United Kingdom  1473672   166909       0            0
##  8 Australia       1252900     7200  134586         6655
##  9 Peru             976790    59497       0            0
## 10 Poland           928256    23987  538203            0
```

Identifying the Highest Positive Against Tested Cases

The top 3 countries that had the highest number of positive cases against the number of tests was:
1. United Kingdom (proportion = 0.113)
2. United States (proportion = 0.108)
3. Turkey (proportion = 0.081)

```
#Vectors containing tested cases and positive cases made to get proportions by country

tested_cases <- covid_top_10$tested
positive_cases <- covid_top_10$positive

names(tested_cases) <- countries
names(positive_cases) <- countries

positive_tested_top_3 <- positive_cases / tested_cases
print(positive_tested_top_3)
```

```
##  United States         Russia          Italy          India         Turkey
##    0.108618191    0.038546552    0.061523368    0.016507300    0.080711720
##         Canada United Kingdom      Australia           Peru         Poland
##    0.054915490    0.113260617    0.005746668    0.060910738    0.025840932
```

Keeping Relevant Information

In the last excercise, the top 3 countries with the most positive cases against tested cases were identified. To not lose the rest of their data from `covid_top_10`, vectors for those 3 countries were made and binded into a matrix.

```
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)

covid_mat <- rbind(united_kingdom, united_states, turkey)
colnames(covid_mat) <- c("Ratio", "tested", "positive", "active", "hospitalized")
print(covid_mat)
```

```
##                Ratio   tested positive  active hospitalized
## united_kingdom  0.11  1473672   166909       0            0
## united_states   0.10 17282363  1877179       0            0
## turkey          0.08  2031192   163941 2980960            0
```

Putting it all Together

To be able to see all of the answers from the previous exercises, the function `list()` was used. This is possible because lists let us combine different types of data objects (vectors, matrix, data frames). Storing the information into a list will let us see the information in a global view with a single variable.

```
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)

data_frame_list <- list("Original Data" = covid_df,
                        "All States" = covid_df_all_states,
                        "Daily Data" = covid_df_all_states_daily,
                        "Top 10 Countries" = covid_top_10)

matrix_list <- list("Top 3 Countries" = covid_mat)

vectors_list <- list("column names" = vector_cols,
                     "countries" = countries)

data_structure_list <- list("Data Frames" = data_frame_list,
                            "Matrices" = matrix_list,
                            "Vectors" = vectors_list)
covid_analysis_list <- list("Question" = question,
                            "Answer" = answer,
                            "Data" = data_structure_list)
covid_analysis_list[2]
```

```
## $Answer
##  Positive tested cases.United States         Positive tested cases.Russia
##                            0.108618191                            0.038546552
##          Positive tested cases.Italy          Positive tested cases.India
##                            0.061523368                            0.016507300
##         Positive tested cases.Turkey         Positive tested cases.Canada
##                            0.080711720                            0.054915490
## Positive tested cases.United Kingdom      Positive tested cases.Australia
##                            0.113260617                            0.005746668
##           Positive tested cases.Peru         Positive tested cases.Poland
##                            0.060910738                            0.025840932
```

Conclusion

To reiterate the goal of this analysis conducted was to answer the following question: **Which countries have reported the highest number of positive cases in relation to the number of tests conducted?** Without knowing the number of tests given, the number of positive cases can be misleading about how COVID virus is spreading in each country. The top 3 countries reported were the UK, US, and Turkey. This data can be useful to confirm the efforts put in place to contain the virus by each country.