Project 1

Alphoncina (Allie) Lyamuya

8/27/2021

Contents

Introduction

Since it was first reported in Wuhan, China in December 2019, Covid-19 has dramatically changed the world as we know it. From shutdowns across the globe, most of which began earlier in March 2020, Covid-19 has not only devastated the global economy but also claimed so many lives across continents. In efforts to curb down the devasting effects of this "unprecedented global crisis", countries, international organizations, and institutions across the world have adopted varied measures to mitigate the effects of Covid-19. It is important to recognize that different countries were hit differently by Covid-19. While some had many cases of Covid-19 infections and record high in Covid-related deaths, some had relatively less positive cases and mortality rates. Even efforts to "flatten the curve" were not the same across countries. In this analysis, I'm interested to answer the question: Which countries have had the highest number of positive cases against the number of tests? I will be using the dataset from Kaggle which include data collected between the 20th of January and the 1st of June 2020.

Data Importation and Exploration

```
# Load libraries
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(ggplot2)
```

```
# Import the data
covid_df <- read_csv("covid19.csv")
```

```
## Rows: 10903 Columns: 14
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Continent_Name, Two_Letter_Country_Code, Country_Region, Province_...
## dbl  (9): positive, hospitalized, recovered, death, total_tested, active, ho...
## date (1): Date
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
covid_df1 <- covid_df

# Explore the dataset
dim(covid_df)
```

```
## [1] 10903    14
```

```
vector_cols <- colnames(covid_df)
vector_cols
```

```
##  [1] "Date"                   "Continent_Name"
##  [3] "Two_Letter_Country_Code" "Country_Region"
##  [5] "Province_State"          "positive"
##  [7] "hospitalized"           "recovered"
##  [9] "death"                  "total_tested"
## [11] "active"                 "hospitalizedCurr"
## [13] "daily_tested"           "daily_positive"
```

```
head(covid_df)
```

```
## # A tibble: 6 x 14
##   Date       Continent_Name Two_Letter_Country_Co~ Country_Region Province_State
##   <date>     <chr>          <chr>                  <chr>          <chr>
## 1 2020-01-20 Asia           KR                     South Korea    All States
## 2 2020-01-22 North America  US                     United States  All States
## 3 2020-01-22 North America  US                     United States  Washington
## 4 2020-01-23 North America  US                     United States  All States
## 5 2020-01-23 North America  US                     United States  Washington
## 6 2020-01-24 Asia           KR                     South Korea    All States
## # ... with 9 more variables: positive <dbl>, hospitalized <dbl>,
## #   recovered <dbl>, death <dbl>, total_tested <dbl>, active <dbl>,
## #   hospitalizedCurr <dbl>, daily_tested <dbl>, daily_positive <dbl>
```

```
glimpse(covid_df)
```

```
## Rows: 10,903
## Columns: 14
## $ Date                   <date> 2020-01-20, 2020-01-22, 2020-01-22, 2020-01-2~
## $ 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               <dbl> 1, 1, 1, 1, 1, 2, 1, 1, 4, 0, 3, 0, 0, 0, 0, 1~
## $ hospitalized           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ recovered              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ death                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ total_tested           <dbl> 4, 1, 1, 1, 1, 27, 1, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ active                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ hospitalizedCurr       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ daily_tested           <dbl> 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ daily_positive         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Data Cleaning and Transformation

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

```
# Selecting relevant variables
covid_df_all_states_daily <- covid_df_all_states %>%
  select(Date, Country_Region, active, hospitalizedCurr, daily_tested,
         daily_positive)
```

```r
# Transform the dataset
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)

covid_df_all_states_daily_sum
```

```
## # A tibble: 108 x 5
##    Country_Region   tested positive  active hospitalized
##    <chr>             <dbl>    <dbl>   <dbl>        <dbl>
##  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
## # ... with 98 more rows
```

```r
# Extract the top 10 rowa from the dataset
covid_top_10 <- head(covid_df_all_states_daily_sum, 10)
covid_top_10
```

```
## # A tibble: 10 x 5
##    Country_Region   tested positive  active hospitalized
##    <chr>             <dbl>    <dbl>   <dbl>        <dbl>
##  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
```

Analyze the Dataset

```r
# Creating vectors
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

# Name the vectors with countries' name vector
names(tested_cases) <- countries
names(positive_cases) <- countries
names(active_cases) <- countries
names(hospitalized_cases) <- countries

# Identify top 3 positive against tested cases
positive_cases / tested_cases
```

```
##  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
```

```r
positive_tested_top_3 <- c(0.113260617, 0.108618191, 0.080711720)
names(positive_tested_top_3) <- c("United Kingdom", "United States", "Turkey")

positive_tested_top_3
```

```
## United Kingdom  United States         Turkey
##     0.11326062     0.10861819     0.08071172
```

```r
# Create 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)

# Create matrix
covid_mat <- rbind(united_kingdom, united_states, turkey)

# Rename the cols of the matrix
colnames(covid_mat) <- c("Ratio", "tested", "positive", "active",
                         "hospitalized")

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
```

```r
# Create a character 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)

# Create list
dataframes_list <- list(covid_df, covid_df_all_states,
                        covid_df_all_states_daily, covid_top_10)
matrices_list <- list(covid_mat)
vectors_list <- list(vector_cols, countries)

data_structure_list <- list(dataframes_list, matrices_list, vectors_list)

# Create another list
covid_analysis_list <- list(question, answer, data_structure_list)
covid_analysis_list
```

```
## [[1]]
## [1] "Which countries have had the highest number of positive cases \nagainst the number of tests?"
##
## [[2]]
## Positive tested cases.United Kingdom  Positive tested cases.United States
##                           0.11326062                           0.10861819
##         Positive tested cases.Turkey
##                           0.08071172
##
## [[3]]
## [[3]][[1]]
## [[3]][[1]][[1]]
## # A tibble: 10,903 x 14
##    Date       Continent_Name Two_Letter_Countr~ Country_Region Province_State
##    <date>     <chr>          <chr>              <chr>          <chr>
##  1 2020-01-20 Asia           KR                 South Korea    All States
##  2 2020-01-22 North America  US                 United States  All States
##  3 2020-01-22 North America  US                 United States  Washington
##  4 2020-01-23 North America  US                 United States  All States
##  5 2020-01-23 North America  US                 United States  Washington
##  6 2020-01-24 Asia           KR                 South Korea    All States
##  7 2020-01-24 North America  US                 United States  All States
##  8 2020-01-24 North America  US                 United States  Washington
##  9 2020-01-25 Oceania        AU                 Australia      All States
## 10 2020-01-25 Oceania        AU                 Australia      Australian Capit~
## # ... with 10,893 more rows, and 9 more variables: positive <dbl>,
## #   hospitalized <dbl>, recovered <dbl>, death <dbl>, total_tested <dbl>,
## #   active <dbl>, hospitalizedCurr <dbl>, daily_tested <dbl>,
## #   daily_positive <dbl>
##
## [[3]][[1]][[2]]
## # A tibble: 3,781 x 13
##    Date       Continent_Name Two_Letter_Country_Code Country_Region positive
##    <date>     <chr>          <chr>                   <chr>             <dbl>
##  1 2020-01-20 Asia           KR                      South Korea           1
```

```
##  2 2020-01-22 North America  US                       United States         1
##  3 2020-01-23 North America  US                       United States         1
##  4 2020-01-24 Asia           KR                       South Korea           2
##  5 2020-01-24 North America  US                       United States         1
##  6 2020-01-25 Oceania        AU                       Australia             4
##  7 2020-01-25 Europe         GB                       United Kingdom        1
##  8 2020-01-25 North America  US                       United States         1
##  9 2020-01-26 Oceania        AU                       Australia             4
## 10 2020-01-26 Asia           IL                       Israel                0
## # ... with 3,771 more rows, and 8 more variables: hospitalized <dbl>,
## #   recovered <dbl>, death <dbl>, total_tested <dbl>, active <dbl>,
## #   hospitalizedCurr <dbl>, daily_tested <dbl>, daily_positive <dbl>
##
## [[3]][[1]][[3]]
## # A tibble: 3,781 x 6
##    Date       Country_Region active hospitalizedCurr daily_tested daily_positive
##    <date>     <chr>           <dbl>            <dbl>        <dbl>          <dbl>
##  1 2020-01-20 South Korea         0                0            0              0
##  2 2020-01-22 United States       0                0            0              0
##  3 2020-01-23 United States       0                0            0              0
##  4 2020-01-24 South Korea         0                0            5              0
##  5 2020-01-24 United States       0                0            0              0
##  6 2020-01-25 Australia           0                0            0              0
##  7 2020-01-25 United Kingdom      0                0            0              0
##  8 2020-01-25 United States       0                0            0              0
##  9 2020-01-26 Australia           0                0            0              0
## 10 2020-01-26 Israel              0                0            0              0
## # ... with 3,771 more rows
##
## [[3]][[1]][[4]]
## # A tibble: 10 x 5
##    Country_Region   tested positive  active hospitalized
##    <chr>             <dbl>    <dbl>   <dbl>        <dbl>
##  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
##
##
## [[3]][[2]]
## [[3]][[2]][[1]]
##                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
##
##
## [[3]][[3]]
```

```
## [[3]][[3]][[1]]
##  [1] "Date"                    "Continent_Name"
##  [3] "Two_Letter_Country_Code" "Country_Region"
##  [5] "Province_State"          "positive"
##  [7] "hospitalized"            "recovered"
##  [9] "death"                   "total_tested"
## [11] "active"                  "hospitalizedCurr"
## [13] "daily_tested"            "daily_positive"
##
## [[3]][[3]][[2]]
##  [1] "United States"  "Russia"         "Italy"          "India"
##  [5] "Turkey"         "Canada"         "United Kingdom" "Australia"
##  [9] "Peru"           "Poland"
```

Discussion and Conclusion

In this analysis, I was interested to answer the question: Which countries have had the highest number of positive cases against the number of tests? To answer this question, I first filtered the dataset by "All States" to explore the values based on Covid-19 data by individual countries. I then selected variables–columns related to daily measures–needed to answer the aforementioned question. The dataset was then grouped by country names and summarized by summing the values of each measure in the dataset. Since the question is seeking the highest number of positive cases against the number of tests, the top 10 countries were extracted from the dataset for further analysis.

From the top 10 data, the top 3 countries were identified by calculating the ratio of positive cases over the number of tests. The results of the top three countries and their ratios were stored in a vector named "positive_tested_top_3". To make the results of this analysis more accessible and digestible, several lists and lists of lists were created. All these lists were stored in the list "covid_analysis_list" which can be used to explore all the results leading up to the answer to the question explored in the analysis.