Investigating COVID-19 Virus Trends

Eda AYDIN

21 02 2021

Contents

Introduction

A pneumonia of unknown cause detected in Wuhan, China was first internationally reported from China on 31 December 2019. Today we know this virus as Coronavirus. COVID-19 which stands for COronaVIrus Disease is the disease caused by this virus. Since then, the world has been engaged in the fight against this pandemic. Several measures have therefore been taken to "flatten the curve". We have consequently experienced social distancing and many people have passed away as well.

In the solidarity to face this unprecedented global crisis, several organizations did not hesitate to share several datasets allowing the conduction of several kinds of analysis in order to understand this pandemic.

It is natural for us to analyze these datasets by ourselves to answer questions since we cannot always rely on the news, and we are data scientists.

In this project, we use a dataset, from Kaggle. This dataset was collected between the 20th of January and the 1st of June 2020. The purpose of this Project is to build our skills and understanding of the data analysis workflow by evaluating the COVID-19 situation through this dataset.

Understanding Data

```
# importing the csv file
library(readr)
covid_df <- read_csv("covid19.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Date = col_date(format = ""),
##   Continent_Name = col_character(),
##   Two_Letter_Country_Code = col_character(),
##   Country_Region = col_character(),
##   Province_State = col_character(),
##   positive = col_double(),
##   hospitalized = col_double(),
##   recovered = col_double(),
##   death = col_double(),
##   total_tested = col_double(),
##   active = col_double(),
##   hospitalizedCurr = col_double(),
##   daily_tested = col_double(),
##   daily_positive = col_double()
## )
```

```
# finding dimension of data frame
dimension_ofdataset <- dim(covid_df)
print(dimension_ofdataset)
```

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

```
# finding the number of columns in this data frame
ncol <- ncol(covid_df)
print(ncol)
```

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

```
# finding the number of rows in this data frame
nrow <- nrow(covid_df)
print(nrow)
```

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

```
# finding the name of columns in this data frame
vector_cols <- colnames(covid_df)
print(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"
```

```r
# displaying the first six rows in this data frame
head_rows <- head(covid_df)
print(head_rows)
```

```
## # 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>
```

```r
# displaying the last six rows in this data frame
tail_rows <- tail(covid_df)
print(tail_rows)
```

```
## # A tibble: 6 x 14
##   Date       Continent_Name Two_Letter_Country_Co~ Country_Region Province_State
##   <date>     <chr>          <chr>                  <chr>          <chr>
## 1 2020-06-01 Asia           IN                     India          All States
## 2 2020-06-01 Asia           ID                     Indonesia      All States
## 3 2020-06-01 Europe         PL                     Poland         All States
## 4 2020-06-01 Europe         RS                     Serbia         All States
## 5 2020-06-01 Asia           TW                     Taiwan         All States
## 6 2020-06-01 Asia           VN                     Vietnam        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>
```

```r
# learning about a tibbles columns, types and dimensions
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 4.0.4
```

```r
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~
```

Isolating the Rows We Need

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# filtering the rows related to "All States" from the Province_State
covid_df_all_states <- filter(covid_df, Province_State == "All States")
print(head(covid_df_all_states))
```

```
## # 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 2020-01-23 North America  US                     United States  All States
## 4 2020-01-24 Asia           KR                     South Korea    All States
## 5 2020-01-24 North America  US                     United States  All States
## 6 2020-01-25 Oceania        AU                     Australia      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>
```

```
# removing the Province_State column from the data frame
covid_df <- select(covid_df, -Province_State)
print(head(covid_df))
```

```
## # A tibble: 6 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 2020-01-22 North America  US                    United States         1
## 4 2020-01-23 North America  US                    United States         1
## 5 2020-01-23 North America  US                    United States         1
## 6 2020-01-24 Asia           KR                    South Korea           2
## # ... with 8 more variables: hospitalized <dbl>, recovered <dbl>, death <dbl>,
## #   total_tested <dbl>, active <dbl>, hospitalizedCurr <dbl>,
## #   daily_tested <dbl>, daily_positive <dbl>
```

Isolating the Columns We Need

```
# selecting the columns related to daily measures from covid_df_all_states
covid_df_all_states_daily <- select(covid_df_all_states, Date, Country_Region, active,
                                    hospitalizedCurr, daily_tested, daily_positive)
print(head(covid_df_all_states_daily))
```

```
## # A tibble: 6 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
```

Extracting the Top Ten Tested Cases Countries

```
covid_df_all_states_daily_sum <- covid_df_all_states_daily %>%
  group_by(Country_Region) %>%
  summarize(
  tested = sum(daily_tested),
  positive = sum(daily_positive),
  active = sum(active),
  hospitalized = sum(hospitalizedCurr)) %>%
  arrange(-tested)

print(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
```

```
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>             <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
```

Identifying the Highest Positive Against Tested Cases

```
# Creating the following vector from the covid_top_10 dataframe
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
```

```
# writing code to name the previous vectors by using names() function
names(tested_cases) <- countries
names(positive_cases) <- countries
names(active_cases) <- countries
names(hospitalized_cases) <- countries
```

```
# identify the top three ratio
positive_tested_top_3 <- sort(positive_cases/tested_cases, decreasing = TRUE)
```

Keeping relevant information

```
# 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)
```

```r
# creating matrix combining these vectors
covid_mat <- rbind(united_kingdom, united_states, turkey)

# rename the columns of this matrix with the vector
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
```

Pulling all together

```r
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)
print(positive_tested_top_3[1:3])
```

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

```r
# creating list that contains the data structure
dataframes <- c(covid_df, covid_df_all_states, covid_df_all_states_daily,
                covid_df_all_states_daily_sum, covid_top_10)

matrices <- covid_mat

vectors <- c(active_cases, countries, hospitalized_cases,
             positive_cases, positive_tested_top_3)

data_structure_list <- c(dataframes, matrices, vectors)

covid_analysis_list <- c(question, answer, data_structure_list)
```