# COVID-19 Data Analysis with R

Sandro Mikautadze

Last compiled on 25/04/2022

# Contents

# 1 Introduction

This is exactly the TLDR you were looking for!

Before starting to read, I would like to specify that the type of analysis is purely **descriptive** and does not aim at understanding the correlations and causation of the results.

All code chunks in this report use variables defined in the scripts.

**Enjoy!**

## 1.1 What is the project about?

The following project is an individual assignment taken from a course on Dataquest. The aim is to analyse a dataset collected from January 20th, 2020 to June 1st, 2020 taken from Kaggle.

The research question is the following: **which countries have had the highest number of positive cases against the number of tests?**

# 2 Data

## 2.1 Raw data Overview

We first analyse the dimensions, the column names and the information provided by each column in the dataset:

```
glimpse(covid19_raw)
```

```
## 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~
```

Looking at the dataset repository from Kaggle we have further information on the meaning of each column:

- `Date`: date of the data collection
- `Country_Region`: country names
- `Province_State`: state/province names; value "All States" is put when state/provincial level data is NA

- `positive`: cumulative number of positive cases reported
- `active`: number of actively cases on that day
- `hospitalized`: cumulative number of hospitalized cases reported
- `hospitalizedCurr`: number of actively hospitalized cases on that day
- `recovered`: cumulative number of recovered cases reported
- `death`: cumulative number of deaths reported
- `total tested`: cumulative number of tests conducted

**Note**: Not all columns in our version of the data frame were present in the data description on Kaggle at the time of download. This indicates that the downloaded file was not updated to a later version of the data set. Yet, Dataquest gives the information provided by the other columns:

- `Continent_Name`: continent name
- `Two_Letter_Country_Code`: country codes
- `Country_Region`: country names
- `daily_tested`: number of tests conducted on the day; if daily data is unavailable, `daily_tested` is averaged across number of days in between
- `daily_positive`: number of positive cases reported on the day; if daily data is unavailable, `daily_positive` is averaged across number of days in between

## 2.2 Raw Data Cleaning

The first clean that we do is remove the `Province_State` column. Indeed, it might create some unwanted bias, as it also gives information about the specific province. So, to make the data "nationalized", we avoid looking at precise regions, and consider only those rows with "All States" value:

```
glimpse(covid19_allstates)
```

```
## Rows: 3,781
## Columns: 14
## $ Date                    <chr> "2020-01-20", "2020-01-22", "2020-01-23", "202~
## $ Continent_Name          <chr> "Asia", "North America", "North America", "Asi~
## $ Two_Letter_Country_Code <chr> "KR", "US", "US", "KR", "US", "AU", "GB", "US"~
## $ Country_Region          <chr> "South Korea", "United States", "United States~
## $ Province_State          <chr> "All States", "All States", "All States", "All~
## $ positive                <int> 1, 1, 1, 2, 1, 4, 1, 1, 4, 0, 3, 1, 1, 5, 0, 0~
## $ 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, 27, 1, 0, 31, 1, 0, 3, 51, 52, 1, 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, 5, 0, 0, 0, 0, 0, 0, 12, 21, 0, 0, 0,~
## $ daily_positive          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
```

As we can see we're left with just 30% of the data we previously had.

The second thing that we need to be aware of is the "unit of measurement" adopted by each column. For example, there are factors with cumulative information, others with daily information. To better address our research question, we only consider the columns providing daily data, that is `Date`, `Country_Region`, `active`, `hospitalizedCurr`, `daily_tested`, `daily_positive`. We get the following refactored data frame:

```
glimpse(covid19_allstates_daily)
```

```
## Rows: 3,781
## Columns: 6
## $ Date             <chr> "2020-01-20", "2020-01-22", "2020-01-23", "2020-01-24~
## $ Country_Region   <chr> "South Korea", "United States", "United States", "Sou~
## $ active           <int> 0, 0, 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, 0, 0,~
## $ daily_tested     <int> 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 12, 21, 0, 0, 0, 1, 10,~
## $ daily_positive   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1,~
```

Now our data is clean.

# 3 Analysis

Based on our cleaned dataset, we notice that our data is collected on a daily basis. Therefore, we can find the ratio of the overall number of positive cases over the total number of tests performed in each country each day. What We want to do is create a data set that groups `Country_Region` and aggregates data on all tests made, positive, active and hospitalized cases in the period of time that the dataset covers (i.e. from January 20th to June 1st).

Displaying the countries based on those that have done more tests, we get:

```
## # A tibble: 15 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
## 11 South Korea      916276    11493  302633            0
## 12 Israel           546626    16647       0        30653
## 13 Germany          518647    29943       0            0
## 14 Belgium          511055    54209  220744            0
## 15 Czech Republic   446758     9321       0            0
```

Now, the result can be easily found by dividing the `positive` cases over the `tested` column and ranking the result for each country. We are going to do it using vectors, extracted from the data frame, as an exercise.

We first extract a vector of the 10 countries that have done more tests, over which we are going to conclude the rest of the analysis:

```
countries
```

```
##  [1] "United States"  "Russia"         "Italy"          "India"
##  [5] "Turkey"         "Canada"         "United Kingdom" "Australia"
##  [9] "Peru"           "Poland"
```

**Note**: if we extract the vector on the whole cleaned data set we get that the length of the vector `countries` is 108, which means that we cover more than half of the world with our data. Yet we decide to restrict the study case to just the top 10 because we have more data on the tests made, and hence a better chance of having the correct data (i.e. a more accurate estimate). As a matter of fact, there might be countries with higher ratios, but more non sufficient data, making our data exploration biased (or at least not correct enough).

Then, we extract our two final vectors to run the analysis and assign to each key the name of the country. Thus, we get:

```
tested_cases
```

```
##  United States         Russia          Italy          India         Turkey
##       17282363       10542266        4091291        3692851        2031192
##         Canada United Kingdom      Australia           Peru         Poland
##        1654779        1473672        1252900         976790         928256
```

```
positive_cases
```

```
##  United States         Russia          Italy          India         Turkey
##        1877179         406368         251710          60959         163941
##         Canada United Kingdom      Australia           Peru         Poland
##          90873         166909           7200          59497          23987
```

The conclusion lies in a simple division: `positive_cases / tested_cases`, which yields the following result:

```
positive_tested_ratio
```

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

# 4 Conclusion

Referring the research question, since there is no precise meaning in the word *highest*, we decide to choose the top 4 countries with the highest positive/tests ratio. By doing so, we get:

```
## United Kingdom  United States         Turkey          Italy
##     0.11326062     0.10861819     0.08071172     0.06152337
```

Therefore the **top 4 countries** were (in order):

1. **United Kingdom** with a ratio of `0.11326062`

2. **United States** with a ratio of 0.10861819
3. **Turkey** with a ratio of 0.08071172
4. **Italy** with a ratio of 0.06152337

Their corresponding data is displayed in the next matrix:

```
##         Positive/Tested %   Tested Positive
## UK              11.326062  1473672   166909
## US              10.861819 17282363  1877179
## Turkey           8.071172  2031192   163941
## Italy            6.152337  4091291   251710
```

# 5 Methods of Improvement and Comments

Even though the structure of this analysis was very clear and defined, the results obtained might not properly answer the question. As a matter of fact, the whole analysis was mainly focused on the countries that had more data available regarding the number of tests made. It follows that, on the one hand, we ended up having a pretty good estimate of the ratio for those countries with many swabs, on the other hand, we might have excluded other "candidate" countries, simply because they did not have a sufficient number of swabs made in this dataset. From a purely statistical point of view, I cannot assess whether the obtained result has a true value or not, as I still don't have sufficient knowledge on this matter. To improve the analysis, we could have tried to find a better assessment method for the countries to include in the analysis. Overall, I am satisfied with the analysis, being this my first project.