COVID-19 Testing Trends [Guided Project]

Jonathan

12/22/2020

Summary

Using a dataset provided by Kaggle, this project aims to build my skills and understanding of the data analysis workflow by evaluating the COVID-19 situation. Source: https://www.kaggle.com/lin0li/covid19testing?select=tested_worldwide.csv (https://www.kaggle.com/lin0li/covid19testing?select=tested_worldwide.csv) Note: All N/A values are replaced with 0 for this test. Due to the lack of observations and recordings, this test is deemed to be inconclusive.

Questions

Which country has the highest number of positive cases against the number of tests?

Code

```
# read and setup the dataframes
# record some general information
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ dplyr   1.0.2
## ✓ tibble  3.0.4     ✓ stringr 1.4.0
## ✓ tidyr   1.1.2     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
```

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

```
covid_df <- read_csv("tested_worldwide.csv")
```

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

```
covid_df[is.na(covid_df)] <- 0
dimension <- dim(covid_df)
vector_cols <- colnames(covid_df)

head(covid_df)
```

```
## # A tibble: 6 x 12
##   Date       Country_Region Province_State positive active hospitalized
##   <date>     <chr>          <chr>             <dbl>  <dbl>        <dbl>
## 1 2020-01-16 Iceland        All States            3      0            0
## 2 2020-01-17 Iceland        All States            4      0            0
## 3 2020-01-18 Iceland        All States            7      0            0
## 4 2020-01-20 South Korea    All States            1      0            0
## 5 2020-01-22 United States  All States            0      0            0
## 6 2020-01-22 United States  Massachusetts         0      0            0
## # … with 6 more variables: hospitalizedCurr <dbl>, recovered <dbl>,
## #   death <dbl>, total_tested <dbl>, daily_tested <dbl>, daily_positive <dbl>
```

```
glimpse(covid_df)
```

```
## Rows: 27,641
## Columns: 12
## $ Date             <date> 2020-01-16, 2020-01-17, 2020-01-18, 2020-01-20, 202…
## $ Country_Region   <chr> "Iceland", "Iceland", "Iceland", "South Korea", "Uni…
## $ Province_State   <chr> "All States", "All States", "All States", "All State…
## $ positive         <dbl> 3, 4, 7, 1, 0, 0, 0, 0, 0, 2, 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, 0…
## $ hospitalized     <dbl> 0, 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, 0…
## $ recovered        <dbl> 0, 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, 0…
## $ total_tested     <dbl> 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 27, 0, 0, 0, 0, 0, 0, …
## $ daily_tested     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0…
## $ daily_positive   <dbl> 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

```
# filter the dataframe as we do not consider state specific cases
covid_df_all_states <- covid_df %>% filter(Province_State == "All States") %>% select(-P
rovince_State)
covid_df <- covid_df %>% select(-Province_State)
```

```
# select important daily information
covid_df_all_states_daily <- covid_df_all_states %>%
  select(Date, Country_Region, active, hospitalizedCurr, daily_tested, daily_positive)
```

```
# find the sum of useful info and retrieve the top 10s
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)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
covid_top_10 <- head(covid_df_all_states_daily_sum, 10)
```

```
# Split the information into different vectors, find the ratio, select top 3
countries <- covid_top_10$Country_Region

tested_cases <- covid_top_10$tested
names(tested_cases) <- countries

positive_cases <- covid_top_10$positive
names(positive_cases) <- countries

active_cases <- covid_top_10$active
names(active_cases) <- countries

hospitalized_cases <- covid_top_10$hospitalized
names(hospitalized_cases) <- countries

positive_tested_top_10 <- positive_cases / tested_cases

positive_tested_top_3 <- sort(positive_tested_top_10, decreasing = T)[1:3]
positive_tested_top_3
```

```
## United States         Italy        Turkey
##    0.07193386    0.05382004    0.05089994
```

```
# Knowing top 3, create a presentable table
## Altered code so this works for any given data
country_names <- names(positive_tested_top_3)

first_pos <- positive_tested_top_3[country_names[1]]
first_stats <- covid_df_all_states_daily_sum %>% filter(Country_Region == country_names[
1]) %>% select(-Country_Region)
first <- c(first_pos, first_stats)

second_pos <- positive_tested_top_3[country_names[2]]
second_stats <- covid_df_all_states_daily_sum %>% filter(Country_Region == country_names
[2]) %>% select(-Country_Region)
second <- c(second_pos, second_stats)

third_pos <- positive_tested_top_3[country_names[3]]
third_stats <- covid_df_all_states_daily_sum %>% filter(Country_Region == country_names[
3]) %>% select(-Country_Region)
third <- c(third_pos, third_stats)

covid_mat <- rbind(first, second, third)
colnames(covid_mat) <- c("ratio", "tested", "positive", "active", "hospitalized")
rownames(covid_mat) <- country_names
covid_mat
```

```
##               ratio      tested    positive active   hospitalized
## United States 0.07193386 136937092 9850413  0        0
## Italy         0.05382004 17370389  934875   17176595 2401146
## Turkey        0.05089994 4351655   221499   4025622  0
```

```
# Test conclusions and answers to original question
question <- "Which countries have had the highest number of positive cases against the n
umber of tests?"

answer <- c("Positive tested cases: " = positive_tested_top_3)
result_list_dataframes <- list(covid_df, covid_df_all_states, covid_df_all_states_daily,
covid_df_all_states_daily_sum)
result_list_matrices <- list(covid_mat)
result_list_vectors <- list(vector_cols, countries)
data_structure_list <- list(dataframes = result_list_dataframes, matrices = result_list_
matrices, vectors = result_list_vectors)

covid_analysis_list <- list(Question = question, "Answers & Results" = answer, "Data Str
uctures" = data_structure_list)
covid_analysis_list[1:2]
```

```
## $Question
## [1] "Which countries have had the highest number of positive cases against the number
of tests?"
##
## $`Answers & Results`
## Positive tested cases: .United States         Positive tested cases: .Italy
##                             0.07193386                              0.05382004
##        Positive tested cases: .Turkey
##                             0.05089994
```

The Sample List of Data Structures created in this program

```
covid_analysis_list
```

```
## $Question
## [1] "Which countries have had the highest number of positive cases against the number
of tests?"
##
## $`Answers & Results`
## Positive tested cases: .United States         Positive tested cases: .Italy
##                            0.07193386                              0.05382004
##        Positive tested cases: .Turkey
##                            0.05089994
##
## $`Data Structures`
## $`Data Structures`$dataframes
## $`Data Structures`$dataframes[[1]]
## # A tibble: 27,641 x 11
##    Date       Country_Region positive active hospitalized hospitalizedCurr
##    <date>     <chr>              <dbl>  <dbl>        <dbl>             <dbl>
##  1 2020-01-16 Iceland                3      0            0                 0
##  2 2020-01-17 Iceland                4      0            0                 0
##  3 2020-01-18 Iceland                7      0            0                 0
##  4 2020-01-20 South Korea            1      0            0                 0
##  5 2020-01-22 United States          0      0            0                 0
##  6 2020-01-22 United States          0      0            0                 0
##  7 2020-01-22 United States          0      0            0                 0
##  8 2020-01-23 United States          0      0            0                 0
##  9 2020-01-23 United States          0      0            0                 0
## 10 2020-01-23 United States          0      0            0                 0
## # … with 27,631 more rows, and 5 more variables: recovered <dbl>, death <dbl>,
## #   total_tested <dbl>, daily_tested <dbl>, daily_positive <dbl>
##
## $`Data Structures`$dataframes[[2]]
## # A tibble: 7,881 x 11
##    Date       Country_Region positive active hospitalized hospitalizedCurr
##    <date>     <chr>              <dbl>  <dbl>        <dbl>             <dbl>
##  1 2020-01-16 Iceland                3      0            0                 0
##  2 2020-01-17 Iceland                4      0            0                 0
##  3 2020-01-18 Iceland                7      0            0                 0
##  4 2020-01-20 South Korea            1      0            0                 0
##  5 2020-01-22 United States          0      0            0                 0
##  6 2020-01-23 United States          0      0            0                 0
##  7 2020-01-24 South Korea            2      0            0                 0
##  8 2020-01-24 United States          0      0            0                 0
##  9 2020-01-25 Australia              0      0            0                 0
## 10 2020-01-25 United Kingdom         1      0            0                 0
## # … with 7,871 more rows, and 5 more variables: recovered <dbl>, death <dbl>,
## #   total_tested <dbl>, daily_tested <dbl>, daily_positive <dbl>
##
## $`Data Structures`$dataframes[[3]]
## # A tibble: 7,881 x 6
##    Date       Country_Region active hospitalizedCurr daily_tested daily_positive
##    <date>     <chr>           <dbl>             <dbl>        <dbl>          <dbl>
##  1 2020-01-16 Iceland             0                 0            0              0
##  2 2020-01-17 Iceland             0                 0            0              1
##  3 2020-01-18 Iceland             0                 0            0              3
```

```
##  4 2020-01-20 South Korea         0                0            0               0
##  5 2020-01-22 United States       0                0            0               0
##  6 2020-01-23 United States       0                0            0               0
##  7 2020-01-24 South Korea         0                0            5               0
##  8 2020-01-24 United States       0                0            0               0
##  9 2020-01-25 Australia           0                0            0               0
## 10 2020-01-25 United Kingdom      0                0            0               0
## # … with 7,871 more rows
##
## $`Data Structures`$dataframes[[4]]
## # A tibble: 146 x 5
##    Country_Region    tested positive   active hospitalized
##    <chr>              <dbl>    <dbl>    <dbl>        <dbl>
##  1 United States  136937092  9850413        0            0
##  2 India          106267322    60959        0            0
##  3 Italy           17370389   934875 17176595      2401146
##  4 Russia          11319603   432269  7621860            0
##  5 Canada           9873530   259992  1354390            0
##  6 Australia        8874298        0   394222        36384
##  7 Israel           4915043      402        0        22726
##  8 Turkey           4351655   221499  4025622            0
##  9 Peru             3578707    59497        0            0
## 10 Brazil           3474441    10321        0            0
## # … with 136 more rows
##
##
## $`Data Structures`$matrices
## $`Data Structures`$matrices[[1]]
##               ratio      tested    positive active   hospitalized
## United States 0.07193386 136937092 9850413  0        0
## Italy         0.05382004 17370389  934875   17176595 2401146
## Turkey        0.05089994 4351655   221499   4025622  0
##
##
## $`Data Structures`$vectors
## $`Data Structures`$vectors[[1]]
##  [1] "Date"             "Country_Region"   "Province_State"   "positive"
##  [5] "active"           "hospitalized"     "hospitalizedCurr" "recovered"
##  [9] "death"            "total_tested"     "daily_tested"     "daily_positive"
##
## $`Data Structures`$vectors[[2]]
##  [1] "United States" "India"         "Italy"         "Russia"
##  [5] "Canada"        "Australia"     "Israel"        "Turkey"
##  [9] "Peru"          "Brazil"
```