Dataquest: COVID-19 Trends

Cindy Zhang

Contents

Introduction

This is my solution to Dataquest's COVID-19 Guided Project from Course 2 (Data Structures in R). It answers the question: **Which countries have had the highest number of positive cases against the number of tests?**

Loading the Data

```
covid_df <- data.frame(read.csv("covid19.csv"))
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"
```

```
typeof(vector_cols)
```

```
## [1] "character"
```

```
head(covid_df)
```

```
##         Date Continent_Name Two_Letter_Country_Code Country_Region
## 1 2020-01-20           Asia                      KR    South Korea
## 2 2020-01-22  North America                      US  United States
## 3 2020-01-22  North America                      US  United States
## 4 2020-01-23  North America                      US  United States
## 5 2020-01-23  North America                      US  United States
## 6 2020-01-24           Asia                      KR    South Korea
##   Province_State positive hospitalized recovered death total_tested active
## 1     All States        1            0         0     0            4      0
## 2     All States        1            0         0     0            1      0
## 3     Washington        1            0         0     0            1      0
## 4     All States        1            0         0     0            1      0
## 5     Washington        1            0         0     0            1      0
## 6     All States        2            0         0     0           27      0
##   hospitalizedCurr daily_tested daily_positive
## 1                0            0              0
## 2                0            0              0
## 3                0            0              0
## 4                0            0              0
## 5                0            0              0
## 6                0            5              0
```

```
glimpse(covid_df)
```

```
## Observations: 10,903
## Variables: 14
## $ Date                   <fct> 2020-01-20, 2020-01-22, 2020-01-22, 2020-01...
## $ Continent_Name         <fct> Asia, North America, North America, North A...
## $ Two_Letter_Country_Code <fct> KR, US, US, US, US, KR, US, US, AU, AU, AU,...
## $ Country_Region         <fct> South Korea, United States, United States, ...
## $ Province_State         <fct> All States, All States, Washington, All Sta...
## $ positive               <int> 1, 1, 1, 1, 1, 2, 1, 1, 4, 0, 3, 0, 0, 0, 0...
## $ hospitalized           <int> 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...
## $ death                  <int> 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, ...
## $ active                 <int> 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...
## $ daily_tested           <int> 0, 0, 0, 0, 0, 5, 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...
```

The glimpse() function is useful when exploring a new dataset because it makes it possible to see every column in a data frame and shows as much as it can so that you can visualize what the dataset looks like.

Isolating Rows We Need

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

We can remove `Province_State` without losing information because it does not provide valuable analysis and it doesn't affect other variables in the set.

Isolating Columns We Need

```
covid_df_all_states_daily <- subset(covid_df_all_states, select = c(Date, Country_Region, active, hospi
```

Extracting 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), hospitali
  arrange(desc(tested))
covid_df_all_states_daily_sum
```

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

```
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
##    <fct>             <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

```
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
```

```
names(tested_cases) <- countries
names(positive_cases) <- countries
names(active_cases) <- countries
names(hospitalized_cases) <- countries
```

```
positive_tested_ratio <- sort(positive_cases/tested_cases, decreasing=TRUE)
positive_tested_top_3 <- positive_tested_ratio[1:3]
```

Keeping Relevant Information

```
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")
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 All Together

```
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)
dataframes <- c(covid_df, covid_df_all_states, covid_df_all_states_daily, covid_df_all_states_daily_sum
matrices <- covid_mat
vectors <- c(active_cases, countries, hospitalized_cases, positive_cases, positive_tested_ratio, positi
data_structure_list <- c(dataframes, matrices, vectors)
covid_analysis_list <- c(question, answer, data_structure_list)
covid_analysis_list[2]
```

```
## $`Positive tested cases.United Kingdom`
## [1] 0.113
```