Exploring COVID-19 Data I

Monica Buczynski

7/13/2020

```
library(tidyverse)
dat <- read_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv")
```

Question 1. The first date reported in the data is January 21, 2020. Find the latest available date reported in these data.

The last available date reported is July 13, 2020.

```
dat %>%
arrange(desc(date))
```

```
## # A tibble: 331,179 x 6
##    date       county   state   fips  cases deaths
##    <date>     <chr>    <chr>   <chr> <dbl>  <dbl>
##  1 2020-07-13 Autauga  Alabama 01001   728     16
##  2 2020-07-13 Baldwin  Alabama 01003  1359     12
##  3 2020-07-13 Barbour  Alabama 01005   413      2
##  4 2020-07-13 Bibb     Alabama 01007   231      1
##  5 2020-07-13 Blount   Alabama 01009   350      1
##  6 2020-07-13 Bullock  Alabama 01011   383     11
##  7 2020-07-13 Butler   Alabama 01013   660     29
##  8 2020-07-13 Calhoun  Alabama 01015   566      5
##  9 2020-07-13 Chambers Alabama 01017   702     30
## 10 2020-07-13 Cherokee Alabama 01019   136      7
## # ... with 331,169 more rows
```

Question 2. Find the cumulative number of deaths reported in the U.S. to date.

The cumulative number of deaths reported in the U.S to date (July 13, 2020) is 8,907,412.

```
dat %>%
summarize(total_deaths = sum(deaths))
```

```
## # A tibble: 1 x 1
##   total_deaths
##          <dbl>
## 1      8907412
```

Question 3. Find the cumulative number of cases reported in the U.S. to date.

The cumulative number of cases reported in the U.S. to date (July 13, 2020) is 170,824,985.

```
dat %>%
summarize(total_cases = sum(cases))
```

```
## # A tibble: 1 x 1
##   total_cases
##         <dbl>
## 1   170824985
```

Question 4. Which state reported the most total cases on the most recent date available?

New York reported the most cumulative cases as of July 13, 2020, followed by California and Florida.

```
dat %>%
group_by(state,date) %>%
summarize(total_cases = sum(cases)) %>%
filter(date =="2020-07-13") %>%
arrange(desc(total_cases))
```

```
## # A tibble: 54 x 3
## # Groups:   state [54]
##    state         date       total_cases
##    <chr>         <date>           <dbl>
##  1 New York      2020-07-13      406962
##  2 California    2020-07-13      336104
##  3 Florida       2020-07-13      282427
##  4 Texas         2020-07-13      273221
##  5 New Jersey    2020-07-13      177469
##  6 Illinois      2020-07-13      156288
##  7 Arizona       2020-07-13      123849
##  8 Georgia       2020-07-13      111937
##  9 Massachusetts 2020-07-13      111827
## 10 Pennsylvania  2020-07-13      100378
## # ... with 44 more rows
```

Question 5. Which county(ies) in the U.S. has/have the fewest cumulative confirmed cases to date?

As of July 13, 2020, the counties of Fallon, followed by Perkins and Roger Mills have the fewest cumulative confirmed cases.

```
dat %>%
group_by(county) %>%
summarize(total_cases = sum(cases)) %>%
arrange(total_cases)
```

```
## # A tibble: 1,895 x 2
##    county       total_cases
##    <chr>              <dbl>
##   1 Fallon                 2
##   2 Perkins                3
##   3 Roger Mills            3
##   4 Foard                  4
##   5 Hickory                6
##   6 Haakon                 8
##   7 Mora                   9
##   8 Ontonagon             10
##   9 Gilliam               11
##  10 Throckmorton          11
## # ... with 1,885 more rows
```

Question 6. Which county in Pennsylvania has the most total cases reported, to date? How many cases have they identified?

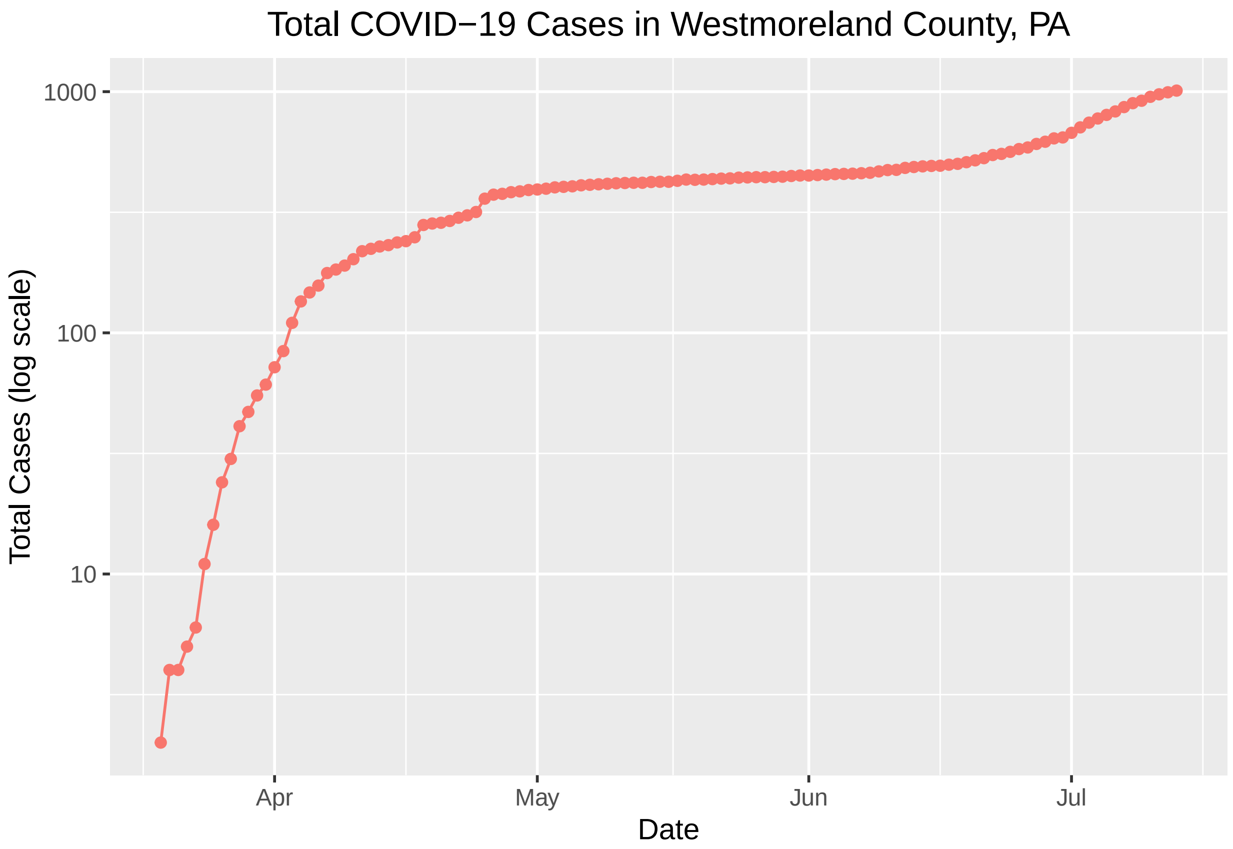
As of July 13, 2020, Philadelphia has the most total cases (1924840) reported in Pennsylvania.

```
dat %>%
group_by(county,state) %>%
summarize(total_cases = sum(cases)) %>%
filter(state =="Pennsylvania") %>%
arrange(desc(total_cases))
```

```
## # A tibble: 68 x 3
## # Groups:   county [68]
##    county       state        total_cases
##    <chr>        <chr>              <dbl>
##  1 Philadelphia Pennsylvania     1924840
##  2 Montgomery   Pennsylvania      606041
##  3 Delaware     Pennsylvania      535397
##  4 Bucks        Pennsylvania      422522
##  5 Berks        Pennsylvania      341464
##  6 Lehigh       Pennsylvania      336838
##  7 Lancaster    Pennsylvania      281085
##  8 Northampton  Pennsylvania      262778
##  9 Luzerne      Pennsylvania      245244
## 10 Chester      Pennsylvania      239072
## # ... with 58 more rows
```

Question 7. Make a plot of the number of cases over time in Westmoreland County—where St. Vincent College is located.

```
dat %>%
    group_by(county, state, date) %>%
summarize(total_cases = sum(cases)) %>%
  filter(county == "Westmoreland", state == "Pennsylvania") %>%
ggplot(aes(x = date, y = total_cases, col = county)) +
geom_line() +
geom_point() +
  scale_y_log10() +
labs(y = "Total Cases (log scale)",
x = "Date",
title = "Total COVID-19 Cases in Westmoreland County, PA") +
guides(color = FALSE) +
theme(plot.title = element_text(hjust = 0.5)) # centers title
```


Total COVID−19 Cases in Westmoreland County, PA
Total Cases (log scale)
Apr
May
Jun
Jul
Date