COVID-19 Fatality Rates in 10 Georgia Counties from January to November 2020

Zachary Gruca

2025-07-24

ECON 6011: R Coding and Analysis I

Final Project (Summer 2025)

Georgia Institute of Technology

Abstract

The research objective for this project is to compare the fatality rates of ten Georgia counties with which I am personally familiar to the fatality rate of all U.S. counties nationally. The familiarity I have with respect to the Georgia counties is the primary motivation for choosing them, but they also represent widely different demographics and geographies. The ten reference counties are as follows: Bartow, Clayton, Cobb, Coffee, Forsyth, Fulton, Gwinnett, Lowndes, Macon, and Paulding. The data for this project were obtained from **openICPSR** and were originally cataloged by **The New York Times** to track the cases and deaths of COVID-19 from the months of January through late November 2020. The data file contains a comprehensive chronological list of reported cases of infection and deaths inclusive of all U.S. counties for the dates in question.

Thanks for reading.

This project is also available on my Github profile, **zgruca** under the **ECON6011-final** repository.

Background

The primary educational motivation for this project is to advance my knowledge of and proficiency with R in particular and my understanding of data science more generally. To this end, I chose a data set that could demonstrate my limited working knowledge of R syntax and data cleaning techniques while challenging myself with some more advanced visualization, in particular using the R package **ggplot2**. Furthermore, I am interested in how COVID-19 affected local communities in the State of Georgia, especially early in the 2020 global pandemic, so this data set helped me to determine how such communities (at the county level) were affected in relation to national and State averages.

The Data

As mentioned in the abstract, the data set used for this project was obtained from **openICPSR**. The data are saved as a .txt file named us-counties.txt. As the data set is comma delimited, it is necessary to import the data as a .csv file. The data will be assigned to the object titled 'covid'. Below are the summary statistics for the us-counties.txt file.

```
summary(covid)
```

```
##      date              county             state               fips
##  Length:758243      Length:758243      Length:758243      Min.   : 1001
##  Class :character   Class :character   Class :character   1st Qu.:18177
##  Mode  :character   Mode  :character   Mode  :character   Median :29207
##                                                           Mean   :31225
##                                                           3rd Qu.:46095
##                                                           Max.   :78030
##                                                           NA's   :7218
##      cases              deaths
##  Min.   :     0   Min.   :    0.00
##  1st Qu.:    26   1st Qu.:    0.00
##  Median :   152   Median :    2.00
##  Mean   :  1447   Mean   :   45.74
##  3rd Qu.:   683   3rd Qu.:   15.00
##  Max.   :364613   Max.   :24206.00
##
```

The variables contained in the data set are self-explanatory save one, 'fips', which is listed as an integer but is better classified as a string. 'Fips' refers to a federal identification number determined by area, similar to a zip code. However, since this information is redundant given the inclusion of State and county information, it will be eliminated from the data set. Furthermore, the 'date' variable must be redefined as a Date instead of a character, which will also be included in the cleaned data set.

```
covid <- covid[,-4]
covid$date <- as.Date(covid$date, format="%Y-%m-%d")
```

After these two modifications,we are left with the following summary statistics:

```
summary(covid)
```

```
##       date               county             state               cases
##  Min.   :2020-01-21   Length:758243      Length:758243      Min.   :     0
##  1st Qu.:2020-05-29   Class :character   Class :character   1st Qu.:    26
##  Median :2020-07-28   Mode  :character   Mode  :character   Median :   152
##  Mean   :2020-07-26                                        Mean   :  1447
##  3rd Qu.:2020-09-25                                        3rd Qu.:   683
##  Max.   :2020-11-22                                        Max.   :364613
##      deaths
##  Min.   :    0.00
##  1st Qu.:    0.00
##  Median :    2.00
##  Mean   :   45.74
##  3rd Qu.:   15.00
##  Max.   :24206.00
```

A problem with the data set is that, while it is conveniently organized chronologically by day from January to November, the variables 'cases' and 'deaths' are cumulative, meaning that one cannot easily retrieve the total cases and deaths without manipulating the data. Thus, it is necessary to extract data for all counties for the last day of reporting, which is November 22, 2020 as noted in the summary statistics. This can be done by using the 'max()' function in R and assigning a new variable for the extracted dates:

```
last_date <- max(covid$date, na.rm = TRUE)
covid_date <- covid[covid$date == last_date, ]
```

It is then possible to calculate the fatality rate for all U.S. counties in the data set by dividing the sum of all deaths by the sum of all cases, multiplying the quotient by 100, and assigning to a new object variable, 'fatality_rate'. I can add this new information to the data set by assigning the variable directly into the extant data set, 'covid_date'.

```
covid_date$fatality_rate <- ifelse(covid_date$cases == 0, NA, covid_date$deaths / covid_date$cases) * 1
```

```
summary(covid_date)
```

```
##       date               county             state               cases
##  Min.   :2020-11-22   Length:3247        Length:3247        Min.   :     0.0
##  1st Qu.:2020-11-22   Class :character   Class :character   1st Qu.:   390.5
##  Median :2020-11-22   Mode  :character   Mode  :character   Median :   949.0
##  Mean   :2020-11-22                                        Mean   :  3792.3
##  3rd Qu.:2020-11-22                                        3rd Qu.:  2495.5
##  Max.   :2020-11-22                                        Max.   :364613.0
##
##      deaths         fatality_rate
##  Min.   :    0.00   Min.   :  0.0000
##  1st Qu.:    4.00   1st Qu.:  0.7692
##  Median :   14.00   Median :  1.4286
##  Mean   :   78.71   Mean   :  1.8132
##  3rd Qu.:   42.00   3rd Qu.:  2.3923
##  Max.   :24206.00   Max.   :100.0000
##                     NA's   :2
```

Data for Georgia and 10 Reference Counties

Now, I simply have to repeat this process for the State of Georgia and the ten reference counties of my choosing. I assign the variables 'ga_date' for the State of Georgia and 'county_date' for the ten Georgia counties. For the ten counties, I will also assign the variable 'all_counties' to extract the counties I need. Summary statistics for both the State of Georgia and the ten reference counties are found below.

```
ga_date <- covid_date[covid_date$state == "Georgia",]
summary(ga_date)
```

```
##       date               county             state               cases
##  Min.   :2020-11-22   Length:160         Length:160         Min.   :   36
##  1st Qu.:2020-11-22   Class :character   Class :character   1st Qu.:  522
##  Median :2020-11-22   Mode  :character   Mode  :character   Median : 1027
##  Mean   :2020-11-22                                        Mean   : 2706
##  3rd Qu.:2020-11-22                                        3rd Qu.: 2526
##  Max.   :2020-11-22                                        Max.   :38870
##      deaths       fatality_rate
##  Min.   :  0.00   Min.   :0.000
##  1st Qu.: 15.00   1st Qu.:1.700
##  Median : 29.50   Median :2.486
##  Mean   : 55.97   Mean   :2.860
##  3rd Qu.: 61.00   3rd Qu.:3.695
##  Max.   :684.00   Max.   :9.636
```

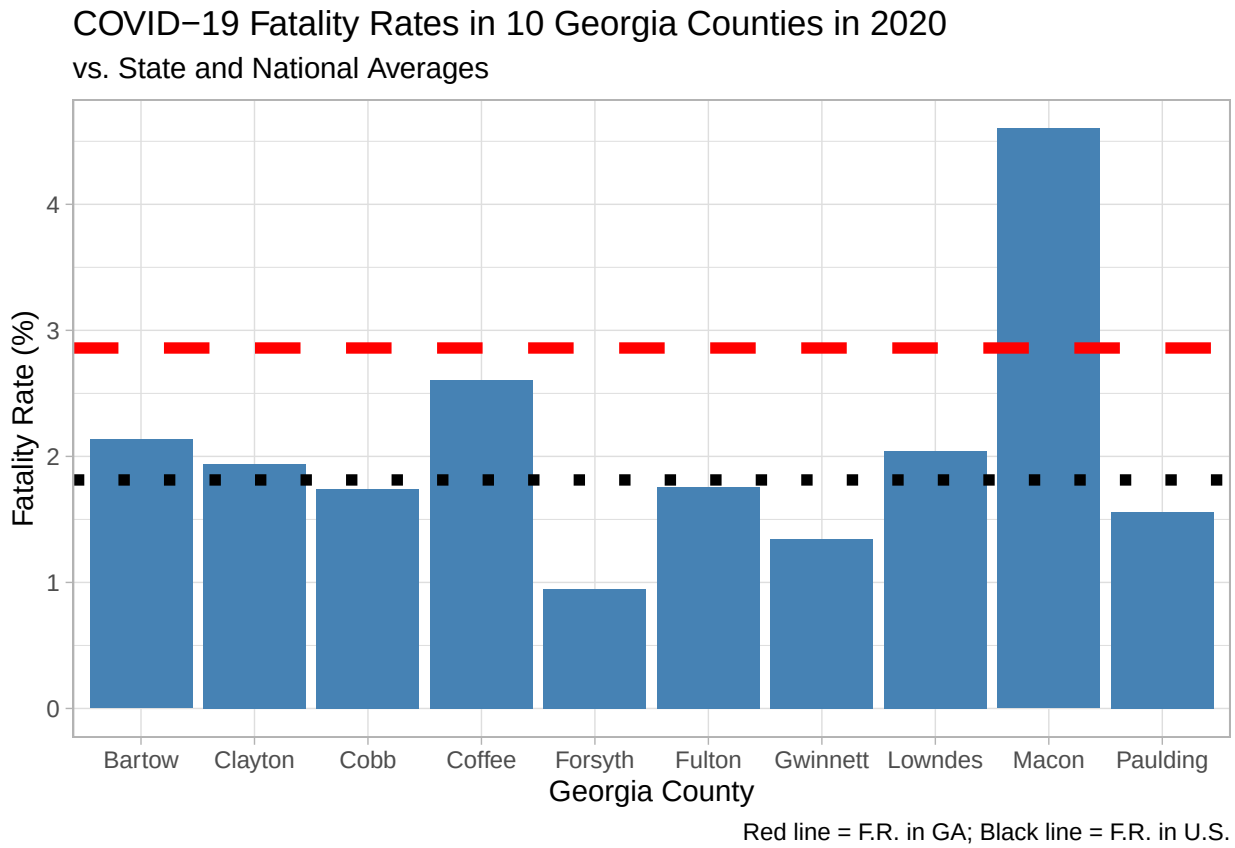
```
all_counties <- c("Bartow", "Clayton", "Cobb", "Coffee", "Dekalb", "Forsyth", "Fulton", "Gwinnett", "Lou
county_date <- ga_date[ga_date$county %in% all_counties, ]
summary(county_date)
```

```
##       date               county             state               cases
##  Min.   :2020-11-22   Length:10          Length:10          Min.   :  326
##  1st Qu.:2020-11-22   Class :character   Class :character   1st Qu.: 4846
##  Median :2020-11-22   Mode  :character   Mode  :character   Median : 6256
##  Mean   :2020-11-22                                        Mean   :14116
##  3rd Qu.:2020-11-22                                        3rd Qu.:24098
##  Max.   :2020-11-22                                        Max.   :38870
##      deaths       fatality_rate
##  Min.   : 15.00   Min.   :0.9471
##  1st Qu.: 74.25   1st Qu.:1.6016
##  Median :114.50   Median :1.8512
##  Mean   :234.50   Mean   :2.0668
##  3rd Qu.:424.25   3rd Qu.:2.1111
##  Max.   :684.00   Max.   :4.6012
```

Data Visualization

```
barplot <- ggplot(aes(x = county, y = fatality_rate), data = county_date) +
  geom_bar(stat="identity", fill="steelblue") +
  xlab("Georgia County") +
  ylab("Fatality Rate (%)") +
  labs(title="COVID-19 Fatality Rates in 10 Georgia Counties in 2020") +
  labs(subtitle="vs. State and National Averages") +
  labs(caption="Red line = F.R. in GA; Black line = F.R. in U.S.") +
  geom_hline(aes(yintercept = 1.8132), linewidth = 2, color = "black", linetype = 3) +
  geom_hline(aes(yintercept = 2.860), linewidth = 2, color = "red", linetype = 2) +
  theme_light()

print(barplot)
```



Results and Next Steps

In terms of what can be gleaned from the cleaned and visualized data set, the results indicate that the fatality rate for the State of Georgia, represented by the red line horizontal, is over one percent higher (2.860%) than the national average (1.8132%), which is represented by the black horizontal. Exactly half (5) of the ten reference counties are above the national fatality rate average, with one county (Macon) notably higher than the State and national averages at a surprising ~4.7% based on the visualization.

Although not a direct part of my research question, it is noteworthy that Macon county is the smallest of the reference counties surveyed, which means that medical and other resources could have played a significant role in the increased fatality rate. By way of induction, a brief examination of the bar graph yields what seems to be a generally inverse correlation between increasing population size per county versus decreasing fatality rates, which would be an interesting next step in the analysis of this data set. One immediately noticeable exception is Forsyth county, which has the lowest fatality rate of the ten reference counties, which could indicate another variable that must needs be accounted for, namely, the average household income per county. Although Forsyth is smaller than several of the other counties with higher fatality rates, affluence could play a significant role in the resources available to a county, and perhaps even more so than population size. While there are many wealthy households in Fulton, which includes Atlanta, for instance, there are also many disadvantaged and poor communities in Fulton to influence the data.

Overall, the objectives set out in the midterm leading up to this final project were satisfied. I was able to manipulate and clean the data as intended, and the data visualization in ggplot2 rendered even more beautifully than anticipated. I did, however, encounter some complications that needed to be resolved during final calculations, etc.

In the midterm write-up, I did not account for zeroes in the divisor, which in this case is the variable 'cases'. This resulted in a Mean and Maximum of Infinity in the 3rd iteration of summary statistics. This produced inaccurate fatality rates for the cleaned data sets. I fixed this problem by including the 'ifelse' argument when adding the fatality rate to the data set, which turned a couple of otherwise problematic zeroes in the 'cases' column to NAs. Another problem I faced was an error message when producing the intended histogram. While I'm unsure whether this applies to histograms in general, when programming in ggplot2, it does not seem possible to define both x and y variables to a geom_histogram(), which is what I intended to create for my visualization of the data. Because I needed to define the x variable as county and the y variable as the fatality rate, however, it was necessary to use the geom_bar() function with the 'stat="identity"' modifier instead. With that said, the **ECON 6011: R Coding and Analysis I** course was sufficient for the programming I needed to accomplish for this particular project, as I did not resort to any additional packages, libraries, or functions that had not been covered in the course.