BST 270: Final Project

Daniel Xu

Due on January 25th, 2021

Our goal is to reproduce the following figures from the NYT article, while critiquing the ease of reproducibility:

1. **Figure 1:** New cases as a function of time with a rolling average plot
2. **Table 1:** Table of cases, hospitalizations and deaths
3. **Table 2:** Table of cases by state

First, let us import relevant packages for analysis.

```r
library(ggplot2) # data visualization
library(dplyr) # data wrangling
library(zoo) # time series analysis
library(scales) # complementary package to ggplot
library(knitr) # knitting formatting tools
```

Second, we can proceed to import the relevant data sets: - `us-states.csv` and `us.csv`: for cases and death data on the state and national level - `all-states-history.csv`: for hospitalization data

```r
# For cases and death data
us_states_data <- read.csv("us-states.csv", header = T)
us_data <- read.csv("us.csv", header = T)

# For hospitalization data
all_states_history_data <- read.csv("all-states-history.csv", header = T)
```

Figure 1

We see that we can largely recreate the figure that was used in the NYT article. That being said, it was a little bit unclear how exactly the rolling mean was calculated in terms of the alignment - while this is a more trivial concern, it could have been helpful to ensure that the exact same numbers were calculated. Otherwise, outliers look largely similar too.

```r
# Calculate new cases nationally
us_data$newcases = us_data$cases - lag(us_data$cases)

# Calculate 7 day rolling average
us_data$roll7dayavg = rollmean(us_data$newcases, k = 7, fill = NA, align = "right")

# Print graph
us_data$date <- as.Date(us_data$date)
ggplot(us_data, aes(x = date, y = newcases)) +
  geom_bar(stat = "identity", fill = "red", alpha = .3) +
  geom_line(aes(x = date, y = roll7dayavg, group = 1), color = "red") +
  scale_y_continuous(labels = scales::comma) +
  scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
  theme(axis.text.x=element_text(angle=60, hjust=1))
```

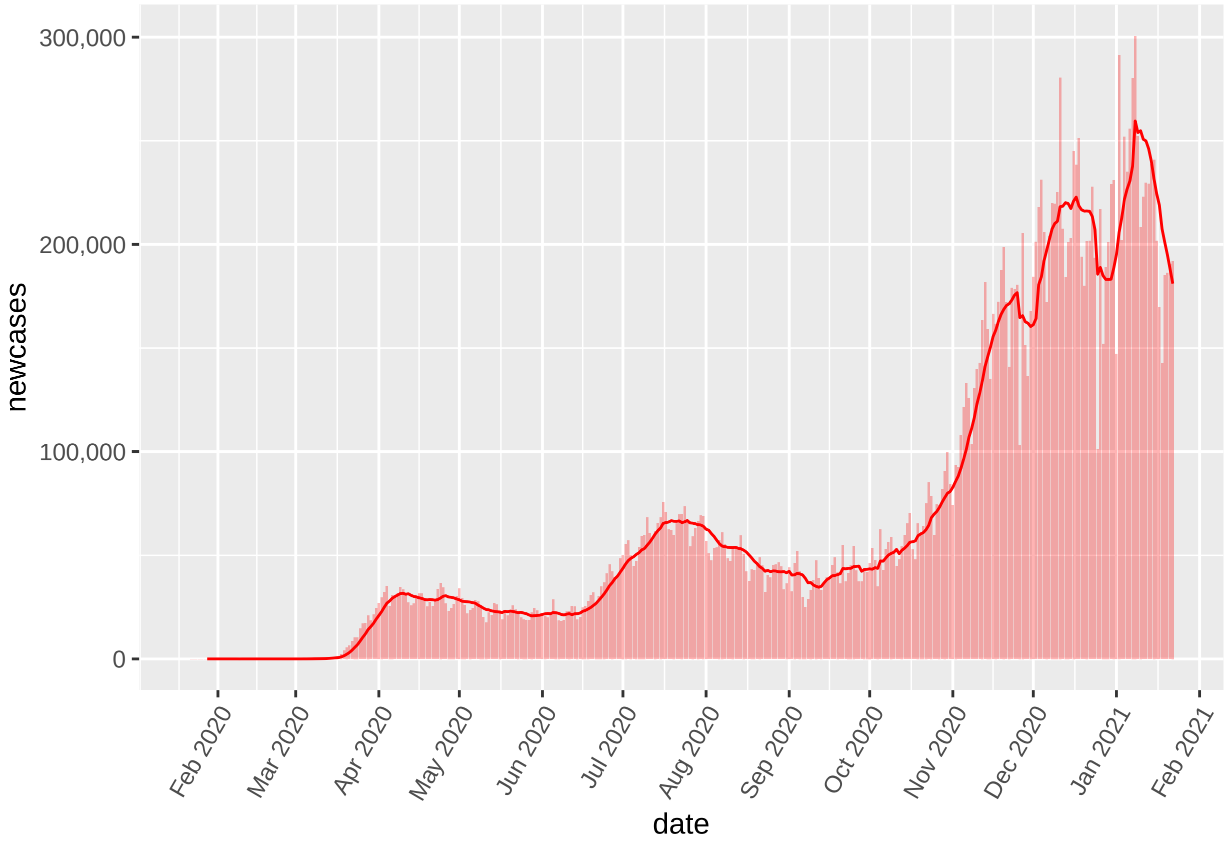

Table 1

Once again, we are able to exactly create the table that is printed in the NYT article, which is a good sign of the reproducibility in the article. Creating this table did however require using both datasets, as opposed to just one, though this is more of an observation than a critique. I do however have two small critiques:

- As mentioned earlier, it is slightly unclear how the rolling average is calculated (right-aligned, center-aligned, or left-aligned), which is used as an input for the 14-day change.
- It is slightly unclear that the 14-day change is using the rolling average, as this is mentioned in small text below the table itself.

```r
# Calculate total reported cases as of Jan 17, new cases on Jan 17, and the
# 14 day change trend using 7-day rolling averages
cases_totalreported <- us_data$cases[us_data$date == "2021-01-17"]
cases_onJan17 <- us_data$newcases[us_data$date == "2021-01-17"]
cases_daychange14 <- us_data$roll7dayavg[us_data$date == "2021-01-17"] /
  us_data$roll7dayavg[us_data$date == "2021-01-03"] - 1

# Calculate new deaths nationally and 7-day rolling average for deaths
us_data$newdeaths <- us_data$deaths - lag(us_data$deaths)
us_data$roll7dayavgdeaths <- rollmean(us_data$newdeaths, k = 7, fill = NA, align = "right")

# Calculate total reported deaths as of Jan 17, new deaths on Jan 17, and the
# 14 day change trend using 7-day rolling averages
deaths_totalreported <- us_data$deaths[us_data$date == "2021-01-17"]
deaths_onJan17 <- us_data$newdeaths[us_data$date == "2021-01-17"]
deaths_daychange14 <- us_data$roll7dayavgdeaths[us_data$date == "2021-01-17"] /
```

```
  us_data$roll7dayavgdeaths[us_data$date == "2021-01-03"] - 1

# Calculate new hospitalizations and 7 day rolling average for hospitalizations
hosp_data <- all_states_history_data %>%
  group_by(date) %>%
  summarize(newhosp = sum(hospitalizedCurrently, na.rm = TRUE))
hosp_data$roll7dayavghosp <- rollmean(hosp_data$newhosp, k = 7, fill = NA, align = "right")

# Calculate new hospitalizations on Jan 17 and 14-day change trend using 7-day
# rolling averages
hosp_onJan17 <- hosp_data$newhosp[hosp_data$date == "2021-01-17"]
hosp_daychange14 <- hosp_data$roll7dayavghosp[hosp_data$date == "2021-01-17"] /
  hosp_data$roll7dayavghosp[hosp_data$date == "2021-01-03"] - 1

# Print table
tableCases <-
  data.frame("Total Reported" = c(cases_totalreported, deaths_totalreported, ""),
              "On Jan 17" = c(cases_onJan17, deaths_onJan17, hosp_onJan17),
              "14-Day Change" = c(cases_daychange14, deaths_daychange14, hosp_daychange14))
rownames(tableCases) <- c("Cases", "Deaths", "Hospitalized")

kable(tableCases, booktabs = TRUE, digits = 2, format.args = list(big.mark = ","),
      col.names = c("Total Reported", "On Jan 17", "14-Day Change"))
```

	Total Reported	On Jan 17	14-Day Change
Cases	23983607	169,641	0.03
Deaths	397612	1,730	0.26
Hospitalized		124,387	0.03

Table 2

Similar to the other figures, we are also able to exactly recreate Table 2 from the NYT article. I would argue that this figure was even easier to reproduce, as nothing was ambiguous and it was entirely created from one dataset.

```
# These are the states that are chosen in the screenshot of the NYT article
# figure. Comment the following line out to display ALL states, or modify it
# with the states of interest.
statesForTable <- c("Arizona", "California", "South Carolina", "Rhode Island",
                    "Oklahoma", "Georgia", "Utah", "Texas", "New York", "Massachusetts")
us_states_data <- us_states_data %>%
  filter(state %in% statesForTable)


# Calculate total cases among all relevant US states (first column in table)
us_states_totalcases <- us_states_data[us_states_data$date == "2021-01-17",]

# Calculate new daily cases on a per-state basis
us_states_data <- us_states_data %>%
  group_by(state) %>%
  arrange(date) %>%
  mutate(newcases = cases - lag(cases))
```

```r
# Calculate the 7 day average in new daily cases on a per-state basis (second
# column in table)
us_states_mean <- us_states_data %>%
  filter(date >= "2021-01-11" & date <= "2021-01-17") %>%
  group_by(state) %>%
  summarize(meancases = mean(newcases))

# Print table
tableCasesByStates <- data.frame("Total" = us_states_totalcases$cases,
                                 "Avg" = us_states_mean$meancases)

rownames(tableCasesByStates) <- us_states_mean$state

kable(tableCasesByStates, booktabs = TRUE, digits = 0,
      format.args = list(big.mark = ","),
      col.names = c("Total Cases", "Daily Average In Last 7 Days"))
```

	Total Cases	Daily Average In Last 7 Days
Arizona	673,882	7,905
California	3,006,583	39,580
Georgia	791,322	8,457
Massachusetts	470,140	5,336
New York	1,242,818	15,281
Oklahoma	354,979	3,374
Rhode Island	104,443	976
South Carolina	388,184	4,808
Texas	2,127,334	22,782
Utah	323,837	2,548

Conclusion

As a brief conclusion, I wanted to commend the NYT on the reproducibility of the figures in their article. While there were a couple of points of confusion (e.g., how to calculate the rolling average), these points were very minor, and I would not see them altering any of the conclusions drawn in the article. Though specific methods used to create tables were not documented by the NYT, they were largely straightforward, and our attempts to exactly reproduce many (if not all of the figures) were successful.