BST 270 Individual Project

Tony Chen

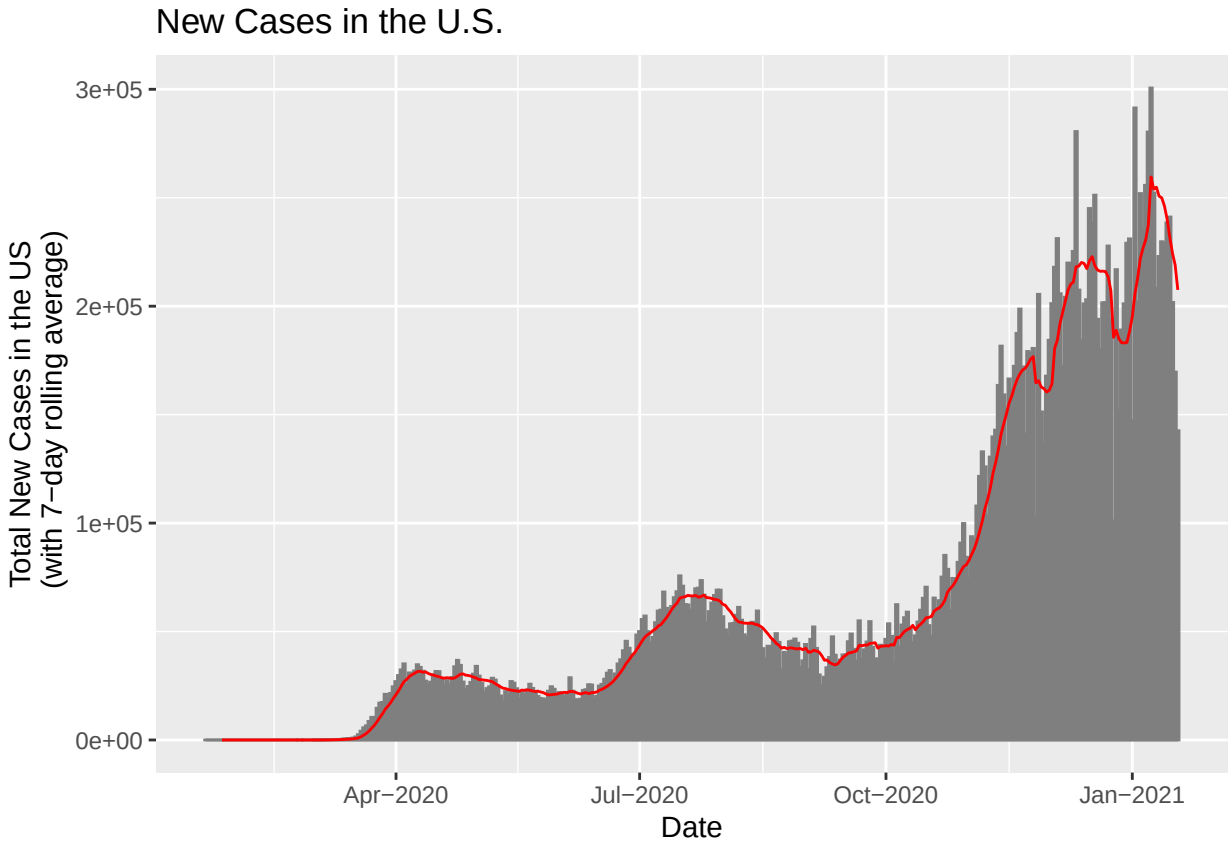
0) Data

COVID-19 data was downloaded from the New York Times GitHub repository, using data updated on Jan 19, 2021, around 9:30am ET. This includes both state- and county-level datasets on reported cases and deaths, collected from Jan 21, 2020 through Jan 18, 2021. The state-level data has a total of 17,724 observations across 55 states and territories. The county-level data has a total of 943,233 observations across 3,219 counties, as identified by their FIPS code. For each dataset, additional columns were created corresponding to the new cases and deaths for each day. Interestingly, there are several observations that have negative values in those columns. Small negative values might be due to adjusting for incorrect reporting from the previous day, but larger negative values may be the result of some systematic issues that need to be further inspected. The NAs in the new case and death columns correspond to the first day recorded for each state or county, for which there was no "previous" day.

Additional data on hospitalizations was downloaded from The COVID Tracking Project. The data includes 18,138 observations across 56 states and territories, collected from Jan 13, 2020 through Jan 19, 2021. For this analysis, only the data on how many people are "Currently Hospitalized" is needed.

1) New cases as a function of time

The state-level data was grouped by date, and the number of cases were summed across all available states with data from that date. NAs were ignored in these sums, and since these correspond to dates at the onset of the pandemic, these values would all be extremely small and not make a substantial difference in national totals. The plot below shows the daily new cases nationwide; the gray bars represent the daily count and the red line represents the 7-day average (i.e. starting from 6 days before).



2) Table of cases, hospitalizations, and deaths

The first two rows of the table below shows the total and newly reported number of cases and deaths as of Jan 18, 2021, as well as the 14-day changes based on the 7-day averages. The third row shows the number of hospitalizations on Jan 18, 2021, and the 14-day change in the hospitalizations calculated using the 7-day averages on Jan 4 and 18, 2021

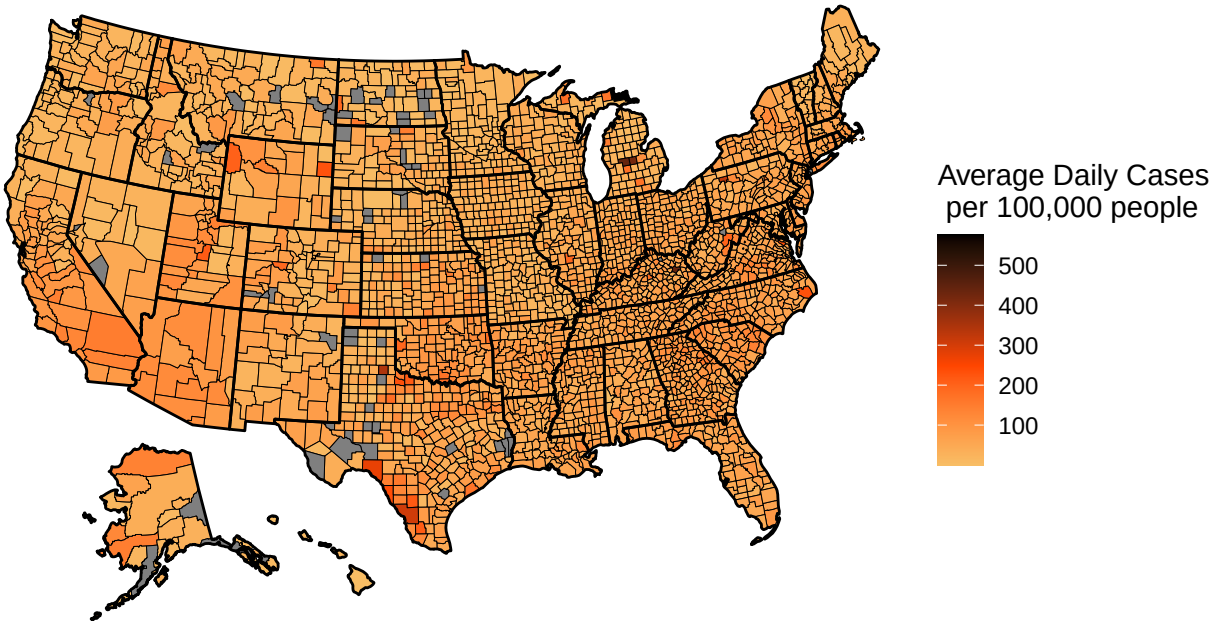
	Total Reported	On Jan. 17	14-Day % Change
Cases	23,983,607	169,641	3
Deaths	397,612	1,730	26
Hospitalizations	NA	124,387	3

3) County-level map for previous week

The map below shows the 7-day average of daily new cases per 100,000 people for each county. The `usmaps` package in R contains data on 2015 county population and tools for map plotting. While there are 3219 FIPS codes in the NYT county data, there are only 3142 FIPS codes in the `usmaps` data. Thus when joining the two datasets, there are only 3142 rows since county population is a necessary parameter for this figure.

As noted before, there are several instances of large negative case counts, and it is not clear why there are so many negative values or how they were dealt with for the New York Times figure. All counties with less than 1 average daily case per 100,000 people (including negative values) were set to NA. The darker shaded counties are considered "Hot Spots" where there are high concentrations of new cases. The gray shaded counties correspond to the hardset NAs for "few or no cases", although some of them had negative values, which needs to be looked into.

Hot Spot Counties



4) Table of cases by state

The table below shows the total number of cases and the 7 day average of new cases for each state as of Jan 17, 2021. The daily averages were rounded to the nearest whole count (that is, 0.5 and above rounds to 1).

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

Code

```
############
## SETUP ##
############

## install packages (if necessary)
# install.packages(c('knitr', 'kableExtra', 'tidyverse', 'zoo'))
# install.packages('https://cran.r-project.org/src/contrib/Archive/usmap/usmap_0.5.1.tar.gz',
#                  repos=NULL, type="source")

## knitr options
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE,
                      tidy.opts=list(width.cutoff=80))

## libraries
library(kableExtra)
library(tidyverse)
library(usmap)
library(zoo)


####################
## DATA WRANGLING ##
####################

## NEW YORK TIMES DATA
# raw state and county data
states0 = read.csv('us-states.csv')
counties0 = read.csv('us-counties.csv')

# add column for new cases and deaths
states = states0 %>%
  mutate(date = as.Date(date)) %>%
  arrange(state, date) %>%
  group_by(state) %>%
  mutate(new_cases = cases - lag(cases)) %>%
  mutate(new_deaths = deaths - lag(deaths)) %>%
  ungroup()

counties = counties0 %>%
  mutate(date = as.Date(date)) %>%
  arrange(fips, date) %>%
  group_by(fips) %>%
  mutate(new_cases = cases - lag(cases)) %>%
  mutate(new_deaths = deaths - lag(deaths)) %>%
  ungroup()

## COVID TRACKING PROJECT DATA
# raw data
hosp0 = read.csv('all-states-history.csv')

# subset data
```

```r
hosp = hosp0 %>%
  select(state, date, hospitalizedCurrently) %>%
  mutate(date = as.Date(date))

## SUMMARY OF DATA
# summary(states)
# summary(counties)
# length(unique(counties$fips, na.rm=T))
# summary(hosp)
# length(unique(hosp$state))


#####################################
## NEW CASES AS A FUNCTION OF TIME ##
#####################################

# omits missing values where new cases and 7 day average could not be calculated
us = states %>%
  group_by(date) %>%
  summarize(total_new_cases = sum(new_cases, na.rm=T)) %>%
  arrange(date) %>%
  mutate(avg_new_cases = rollmean(total_new_cases, k = 7, fill = NA, align='right'))

# barplot and rolling average line
ggplot(us, aes(x=date, y=total_new_cases)) +
  geom_bar(stat='identity', color='gray50') +
  geom_line(aes(x=date, y=avg_new_cases), color='red') +
  scale_x_date(date_labels = "%b-%Y") +
  ylab('Total New Cases in the US\n (with 7-day rolling average)') +
  xlab('Date') +
  ggtitle('New Cases in the U.S.')

################################################
## TABLE OF CASES, HOSPITALIZATIONS, AND DEATHS ##
################################################

## TOTAL COUNTS
case_death_total = states %>%
  group_by(state) %>%
  filter(date=='2021-01-17') %>%
  ungroup() %>%
  summarize(total_cases = sum(cases), # total reported
            total_deaths = sum(deaths),
            totalnew_cases = sum(new_cases), # new cases
            totalnew_deaths = sum(new_deaths))

hosp_total = hosp %>%
  group_by(date) %>%
  summarize(totalnew_hosps = sum(hospitalizedCurrently, na.rm=T)) %>%
  filter(date=='2021-01-17') %>%
  mutate(total_hosps=NA) %>%
  select(-date)
```

```r
## 14-DAY CHANGES
case_death_change = states %>%
  group_by(date) %>%
  arrange(date) %>%
  summarize(total_new_cases = sum(new_cases, na.rm=T),
            total_new_deaths = sum(new_deaths, na.rm=T)) %>%
  mutate(avg_new_cases = rollmean(total_new_cases, k = 7, fill = NA, align='right'),
         avg_new_deaths = rollmean(total_new_deaths, k = 7, fill = NA, align='right')) %>%
  filter(date=='2021-01-03' | date=='2021-01-17') %>%
  summarize(pct_cases = (avg_new_cases/lag(avg_new_cases) - 1) * 100,
            pct_deaths = (avg_new_deaths/lag(avg_new_deaths) - 1) * 100) %>%
  round(0) %>%
  na.omit()

hosp_change = hosp %>%
  group_by(date) %>%
  summarize(total_hosps = sum(hospitalizedCurrently, na.rm=T)) %>%
  arrange(date) %>%
  mutate(avg_new_hosps = rollmean(total_hosps, k=7, fill=NA, align='right')) %>%
  filter(date=='2021-01-03' | date=='2021-01-17') %>%
  summarize(pct_hosps = (avg_new_hosps/lag(avg_new_hosps) - 1)*100) %>%
  round(0) %>%
  na.omit()

### COMBINE RESULTS INTO TABLE
table2 = cbind(case_death_total, case_death_change, hosp_total, hosp_change) %>%
  gather(key=Type, value=Value) %>%
  separate(Type, c('Stat', 'Metric'), '_') %>%
  spread(key=Stat, value=Value) %>%
  select(`Total Reported`=total, `On Jan. 17`=totalnew, `14-Day % Change`=pct)
rownames(table2) = c('Cases', 'Deaths', 'Hospitalizations')

kable(table2, 'latex', format.args = list(big.mark = ","))


########################################
## COUNTY-LEVEL MAP FOR PREVIOUS WEEK ##
########################################

# 7 day average per county
counties1 = counties %>%
  group_by(fips) %>%
  arrange(date) %>%
  mutate(avg_new_cases = rollmean(new_cases, k=7, fill=NA, align='right')) %>%
  ungroup() %>%
  filter(date=='2021-01-17')

# merge with countypop data (from usmaps package)
counties2 = countypop %>%
  mutate(fips=as.integer(fips)) %>%
  left_join(counties1, by='fips') %>%
  mutate(avg_new_cases_pop = avg_new_cases / pop_2015 * 100000) %>% # cases / 100,000 people
  mutate(fips = as.integer(fips)) %>%
```

```
  mutate(avg_new_cases_pop = ifelse(avg_new_cases_pop < 1, NA, avg_new_cases_pop)) %>%
  na.omit()

# plot map
state_plot = plot_usmap("states")
county_plot <- plot_usmap(data = counties2, value='avg_new_cases_pop')

ggplot() +
  # county borders and case counts
  geom_polygon(data=county_plot[[1]],
               aes(x=x,
                   y=y,
                   group=group,
                   fill = county_plot[[1]]$avg_new_cases_pop),
               color = "black",
               size = 0.1) +
  # state borders
  geom_polygon(data=state_plot[[1]],
               aes(x=x,
                   y=y,
                   group=group),
               color = "black",
               fill = alpha(0.01)) +
  scale_fill_gradient2(low='lightgoldenrod', mid='orangered', high='black',
                       midpoint=250,
                       name='Average Daily Cases\n per 100,000 people') +
  ggtitle('Hot Spot Counties') +
  coord_equal() +
  theme_void()


##############################
## TABLE OF CASES BY STATE ##
##############################

# calculate 7-day averages for each state
table4 = states %>%
  group_by(state) %>%
  arrange(date) %>%
  mutate(avg_new_cases = round(rollmean(new_cases, k = 7, fill = NA, align='right'))) %>%
  filter(date=='2021-01-17') %>%
  select(`State`=state, `Total Cases`=cases,
         `Daily Avg New Cases in Last 7 Days`=avg_new_cases)

# subset of full table corresponding to NYT snapshot
state_subset = data.frame(State=c('Arizona', 'California', 'South Carolina',
                                  'Rhode Island', 'Oklahoma', 'Georgia', 'Utah',
                                  'Texas', 'New York', 'Massachusetts'))
table4_small = left_join(state_subset, table4)

kable(table4_small, 'latex', format.args = list(big.mark = ","))

``
```