tl_covid_exercise

Katherine Liu

12/13/2020

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

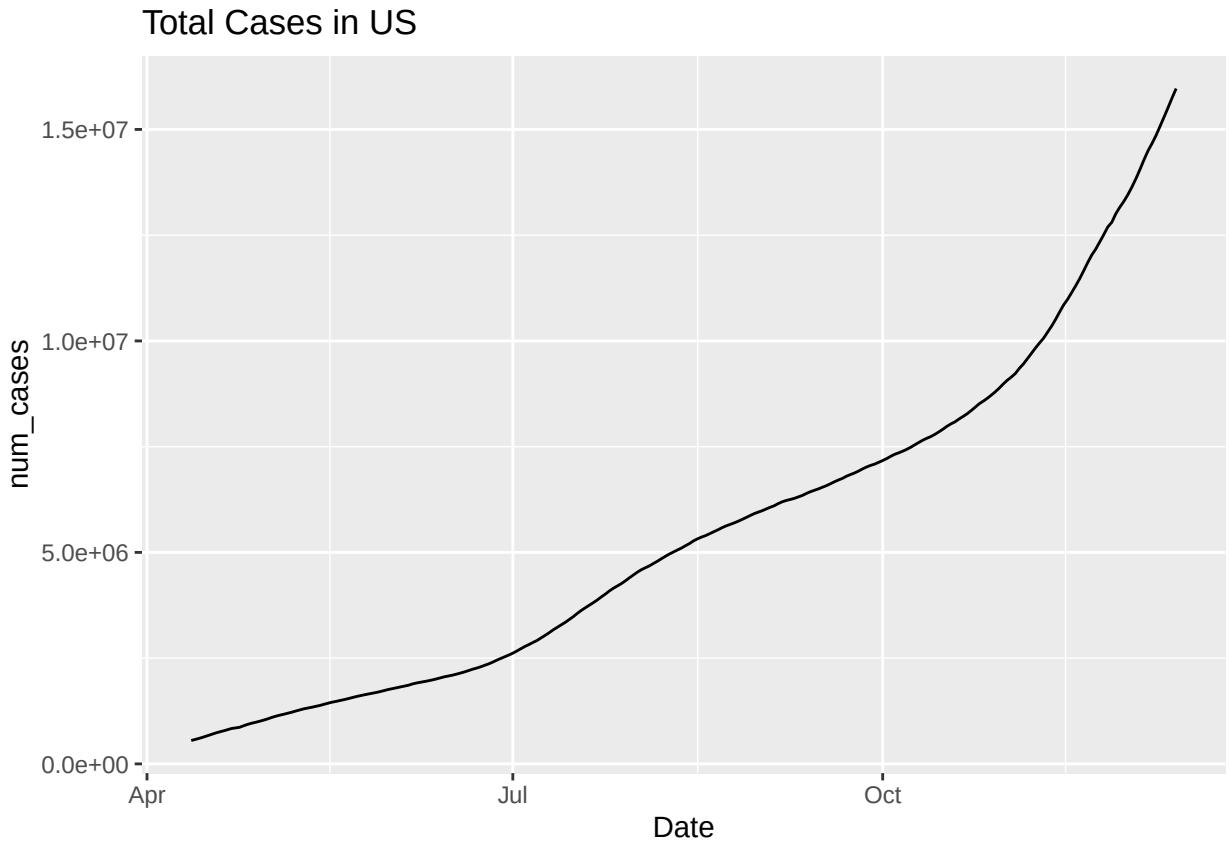
```
## Warning: package 'purrr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

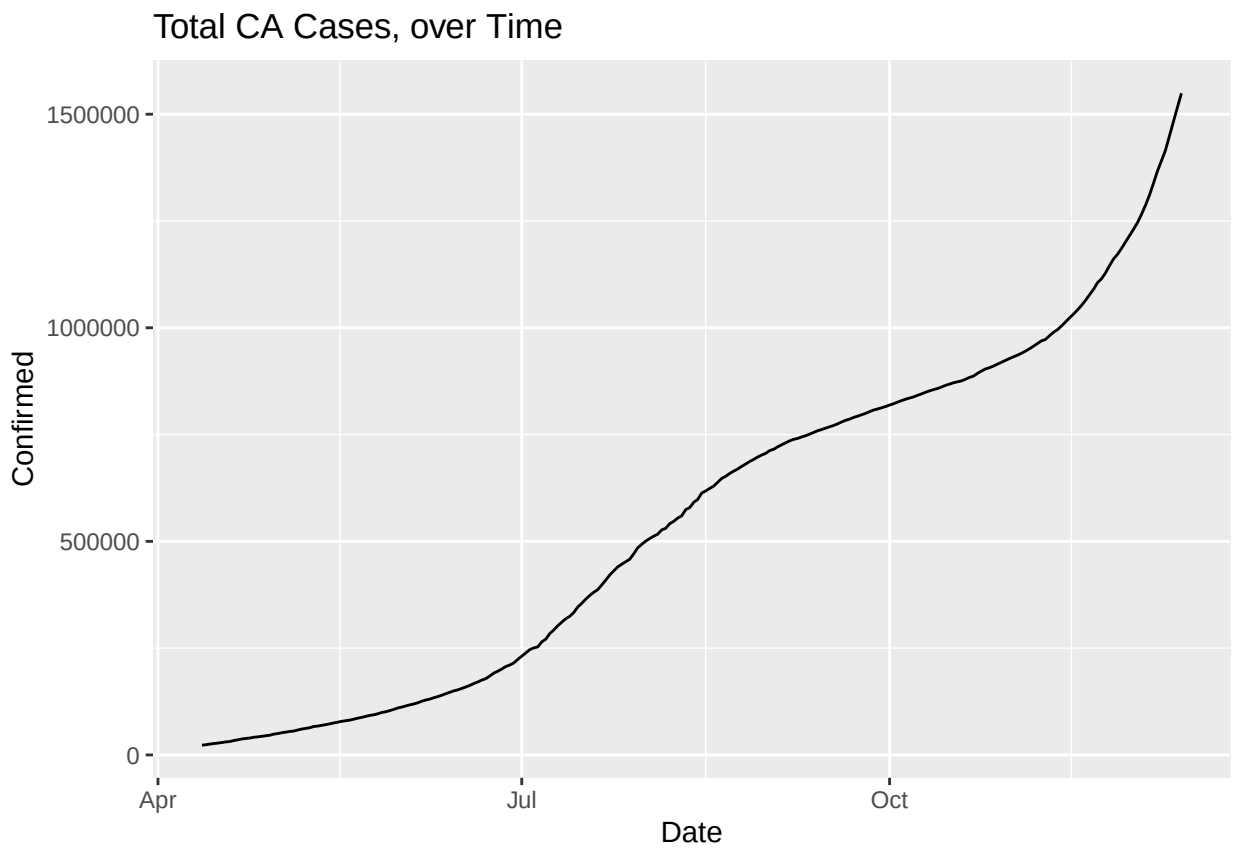
```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
covid_data <-
  list.files(pattern = "*.csv") %>%
  map_df(~fread(.))

remove = c("Diamond Princess", "District of Columbia", "Grand Princess", "Guam",
           "Puerto Rico", "American Samoa", "Northern Mariana Islands",
           "Recovered", "Virgin Islands")
state_data = covid_data[!covid_data$Province_State %in% remove]
state_data = select(state_data, -c(Country_Region))

colnames(state_data)[1] = "State"
colnames(state_data)[2] = "Date"
colnames(state_data)[4] = "Long"

state_data$Date = as.Date(state_data$Date, "%Y-%m-%d %H:%M:%S", tz = "GMT")
state_data[59,2] = state_data[58,2]

case_totals = state_data %>%
  group_by(Date) %>%
  mutate(num_cases = sum(Confirmed))%>%
  distinct(Date, num_cases)

ggplot(data = case_totals) +
  geom_line(mapping = aes(x = Date, y = num_cases)) +
  ggtitle("Total Cases in US")
```

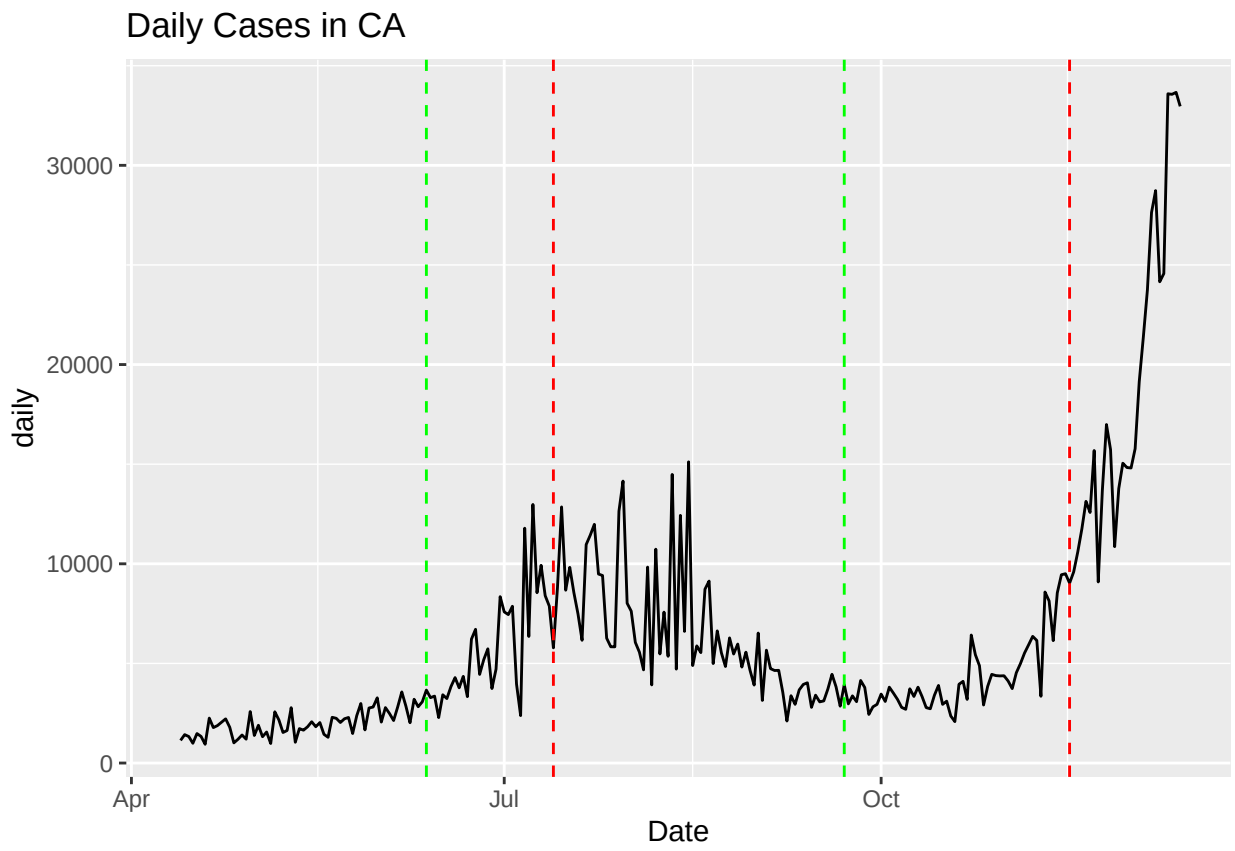


```
state_data %>%
  filter(State == "California") %>%
  ggplot(mapping = aes(x = Date, y = Confirmed)) +
  geom_line() +
  ggtitle("Total CA Cases, over Time")
```




```
state_data %>%
  filter(State == "California") %>%
  mutate(daily = Confirmed - lag(Confirmed)) %>%
  ggplot(mapping = aes(x = Date, y = daily)) +
  geom_line() +
  geom_vline(xintercept = as.numeric(as.Date("2020-06-12")), linetype=2, col = "green") +
  geom_vline(xintercept = as.numeric(as.Date("2020-07-13")), linetype=2, col = "red") +
  geom_vline(xintercept = as.numeric(as.Date("2020-11-16")), linetype=2, col = "red") +
  geom_vline(xintercept = as.numeric(as.Date("2020-09-22")), linetype=2, col = "green") +
  ggtitle("Daily Cases in CA")
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

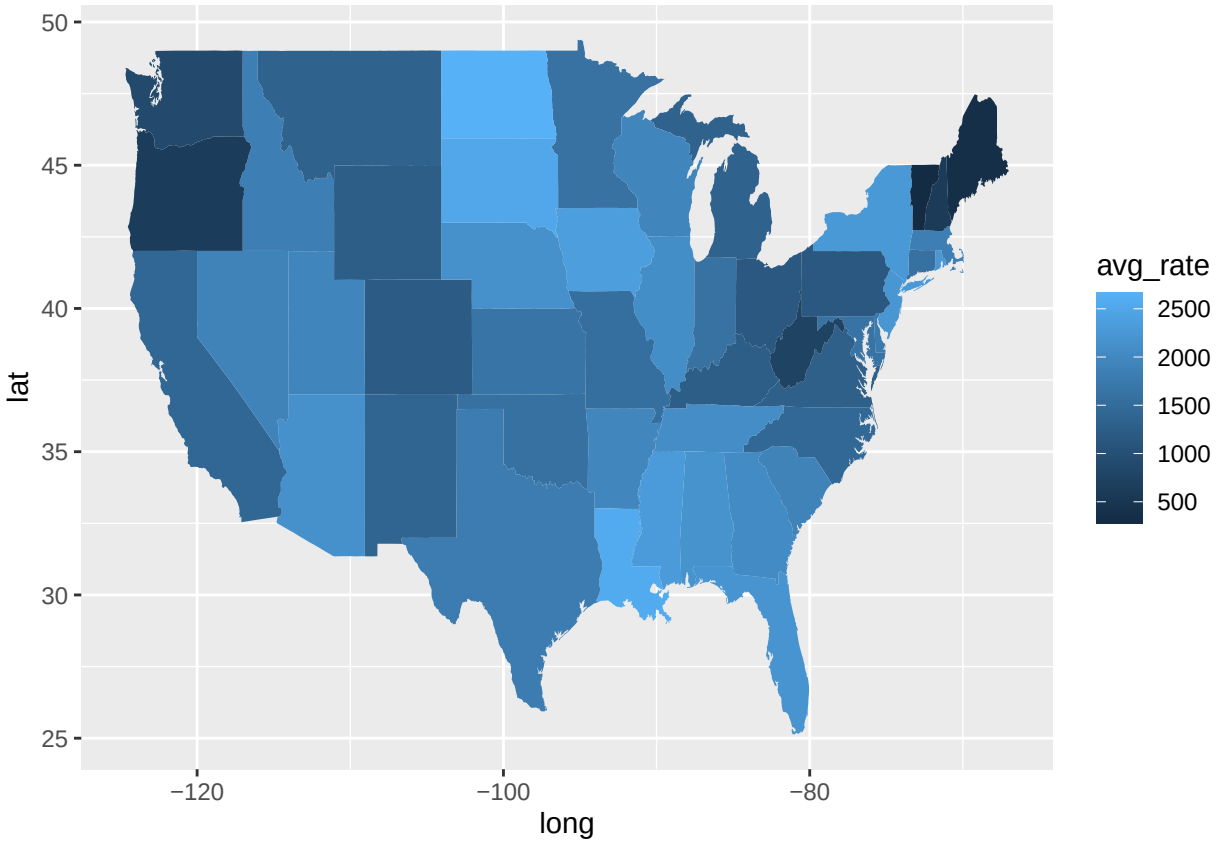



```
(state_avg = state_data %>%
  group_by(State) %>%
  mutate(avg_rate = mean(Incident_Rate)) %>%
  distinct(State, avg_rate) %>%
  arrange(avg_rate)
 )
```

```
## # A tibble: 50 x 2
## # Groups:   State [50]
##    State         avg_rate
##    <chr>            <dbl>
##  1 Vermont           281.
##  2 Maine             355.
##  3 Hawaii            500.
##  4 New Hampshire     601.
##  5 Oregon            614.
##  6 West Virginia     745.
##  7 Washington        892.
##  8 Alaska           1079.
##  9 Ohio             1146.
## 10 Pennsylvania     1155.
## # ... with 40 more rows
```

```
state_avg$State = tolower(state_avg$State)
state_df = map_data("state")
```

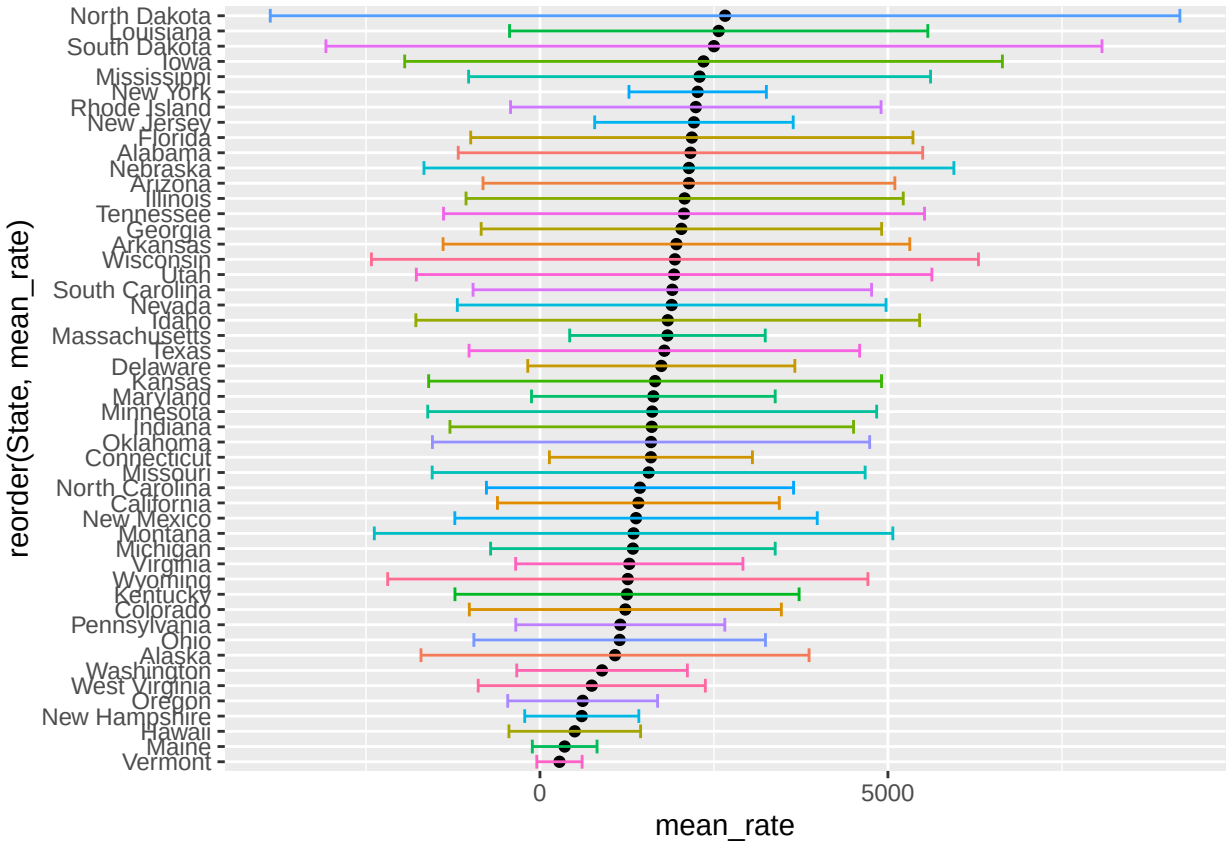
```
colnames(state_df)[5] = "State"


merge_state <- inner_join(state_df, state_avg, by = "State")
ggplot() +
  geom_polygon( data=merge_state,
          aes(x=long, y=lat, group=group, fill = avg_rate))
```




```
ci = state_data %>%
  group_by(State) %>%
  dplyr::summarize(mean_rate = mean(Incident_Rate, na.rm = TRUE),
            sd_rate = sd(Incident_Rate, na.rm = TRUE),
            n_rate = n()) %>%
  mutate(se_rate = sd_rate / sqrt(n_rate),
         lower_ci_rate = mean_rate - qt(1 - (0.05 / 1225), n_rate - 1) * se_rate,
         upper_ci_rate = mean_rate + qt(1 - (0.05 / 1225), n_rate - 1) * se_rate)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(ci, mapping = aes(x = reorder(State, mean_rate), y = mean_rate)) +
  geom_point() +
  geom_errorbar(data = ci, aes(ymin = mean_rate - 2*sd_rate, ymax = mean_rate + 2 * sd_rate, color = Sta
  theme(legend.position = "none") +
  coord_flip()
```


North Dakota
Louisiana
South Dakota
Iowa
Mississippi
New York
Rhode Island
New Jersey
Florida
Alabama
Nebraska
Arizona
Illinois
Tennessee
Georgia
Arkansas
Wisconsin
Utah
South Carolina
Nevada
Idaho
Massachusetts
Texas
Delaware
Kansas
Maryland
Minnesota
Indiana
Oklahoma
Connecticut
Missouri
North Carolina
California
New Mexico
Montana
Michigan
Virginia
Wyoming
Kentucky
Colorado
Pennsylvania
Ohio
Alaska
Washington
West Virginia
Oregon
New Hampshire
Hawaii
Maine
Vermont
reorder(State, mean_rate)
0
5000
mean_rate