Covid-19 Case Study

Alex

2025-06-10

Importing Library and Data

This markdown uses `tidyverse`, `forcats`, `ggplot2` and `usmap` libraries. Make sure to install!

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_deaths_global.csv")
urls <- str_c(url_in, file_names)

US_cases <- read_csv(urls[1])
global_cases <- read_csv(urls[2])
US_deaths <- read_csv(urls[3])
global_deaths <- read_csv(urls[4])
```

Cleaning Data

Although the tables are hard to read at the moment, I think I will keep them separate from each other and pivot/clean when needed for a graph.

Graphs

Pivoting Global Tables

```
global_deaths_long <- global_deaths %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
    names_to = "Date",
    values_to = "Deaths"
  ) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%y"))

global_cases_long <- global_cases %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
```

```
    names_to = "Date",
    values_to = "Cases"
  ) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%y"))
head(global_cases_long)
```

```
## # A tibble: 6 x 6
##   `Province/State` `Country/Region`   Lat  Long Date       Cases
##   <chr>            <chr>            <dbl> <dbl> <date>     <dbl>
## 1 <NA>             Afghanistan       33.9  67.7 2020-01-22     0
## 2 <NA>             Afghanistan       33.9  67.7 2020-01-23     0
## 3 <NA>             Afghanistan       33.9  67.7 2020-01-24     0
## 4 <NA>             Afghanistan       33.9  67.7 2020-01-25     0
## 5 <NA>             Afghanistan       33.9  67.7 2020-01-26     0
## 6 <NA>             Afghanistan       33.9  67.7 2020-01-27     0
```

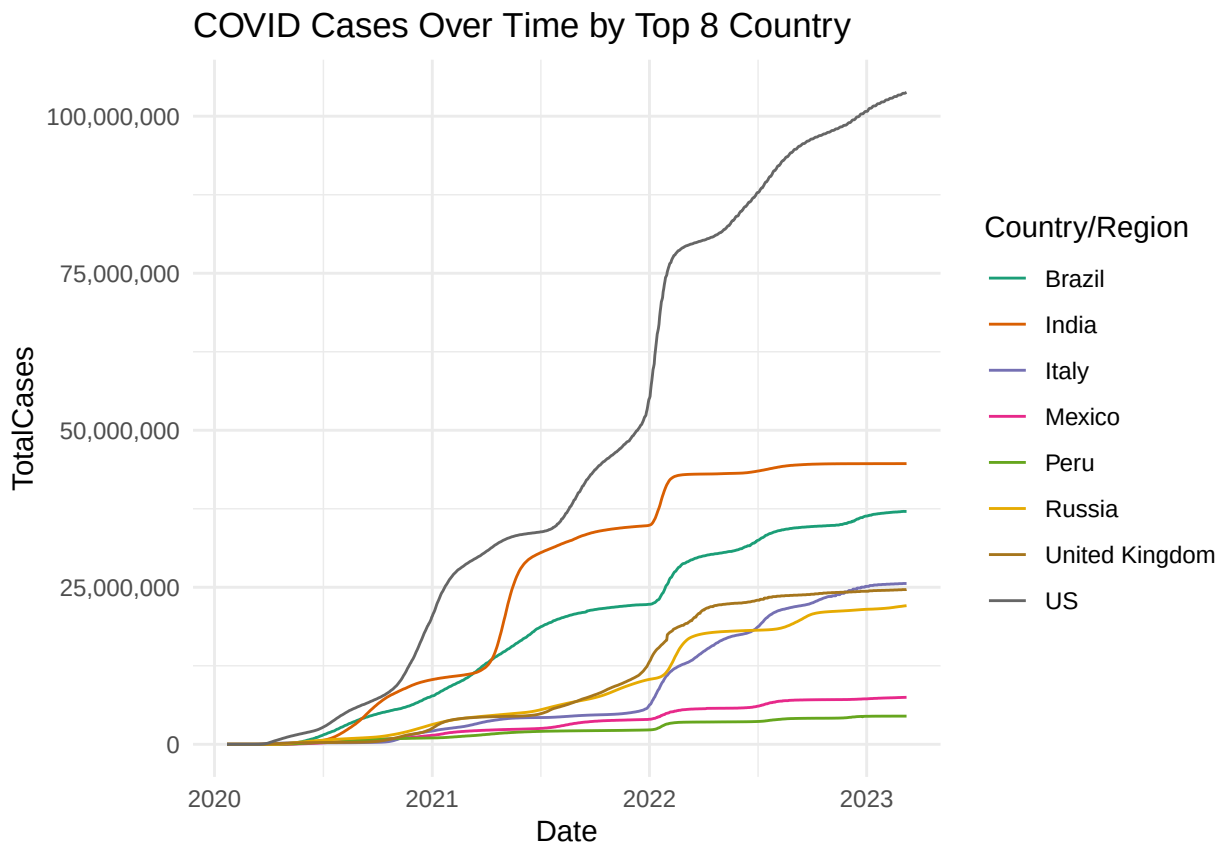
Pivoting the table like this allows us to see the daily cases per region

Grabbing Top 8 COuntries

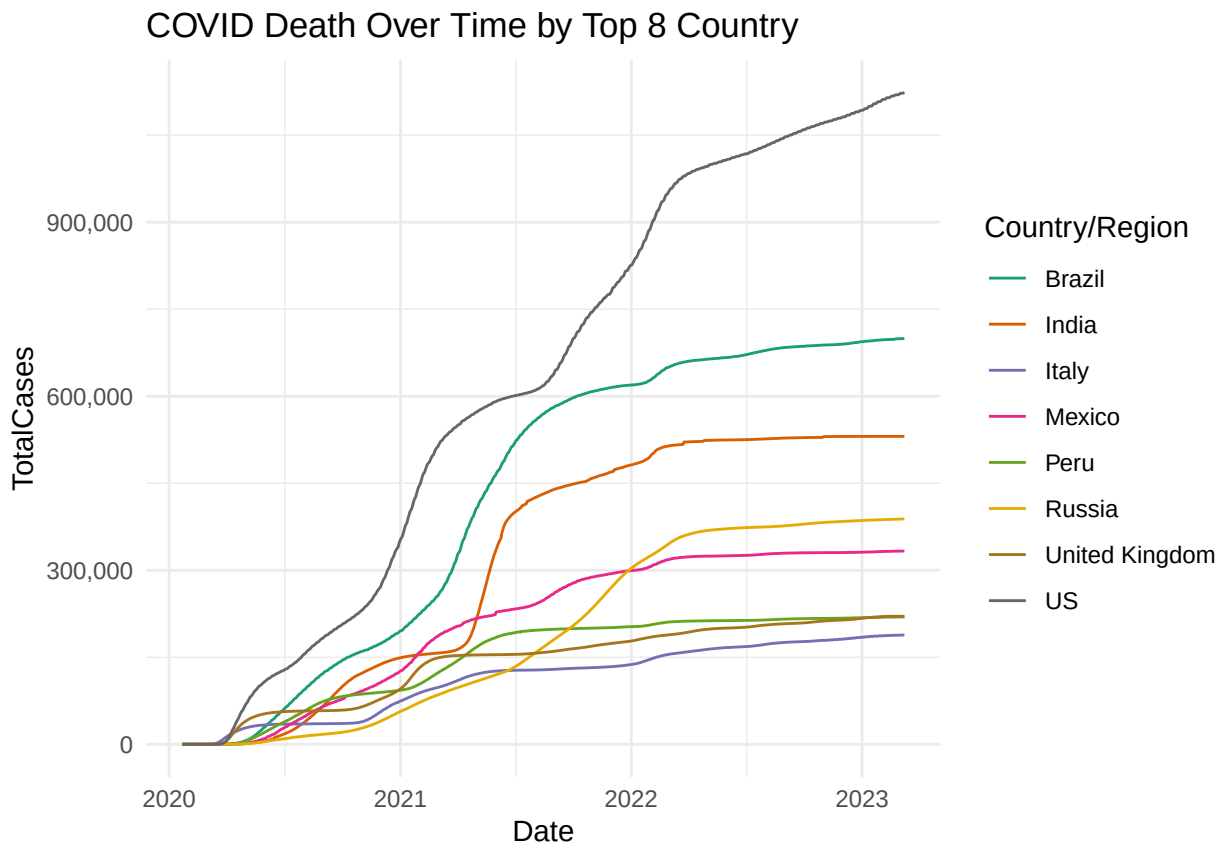
```
top_countries <- global_deaths_long %>%
  group_by(`Country/Region`) %>%
  summarise(Total = sum(Deaths, na.rm = TRUE)) %>%
  top_n(8, Total) %>%
  pull(`Country/Region`)
```

The 8 is arbitrary. It gives a wide variety of countries of different regions.

```
global_cases_long %>%
  filter(`Country/Region` %in% top_countries) %>%
  group_by(`Country/Region`, Date) %>%
  summarise(TotalCases = sum(Cases, na.rm = TRUE)) %>%
  ggplot(aes(x = Date, y = TotalCases, color = `Country/Region`)) +
  scale_color_brewer(palette = "Dark2") +
  geom_line() +
  scale_y_continuous(labels = scales::label_comma()) +
  labs(title = "COVID Cases Over Time by Top 8 Country") +
  theme_minimal()
```




```r
global_deaths_long %>%
  filter(`Country/Region` %in% top_countries) %>%
  group_by(`Country/Region`, Date) %>%
  summarise(TotalCases = sum(Deaths, na.rm = TRUE)) %>%
  ggplot(aes(x = Date, y = TotalCases, color = `Country/Region`)) +
  scale_color_brewer(palette = "Dark2") +
  geom_line() +
  scale_y_continuous(labels = scales::label_comma()) +
  labs(title = "COVID Death Over Time by Top 8 Country") +
  theme_minimal()
```



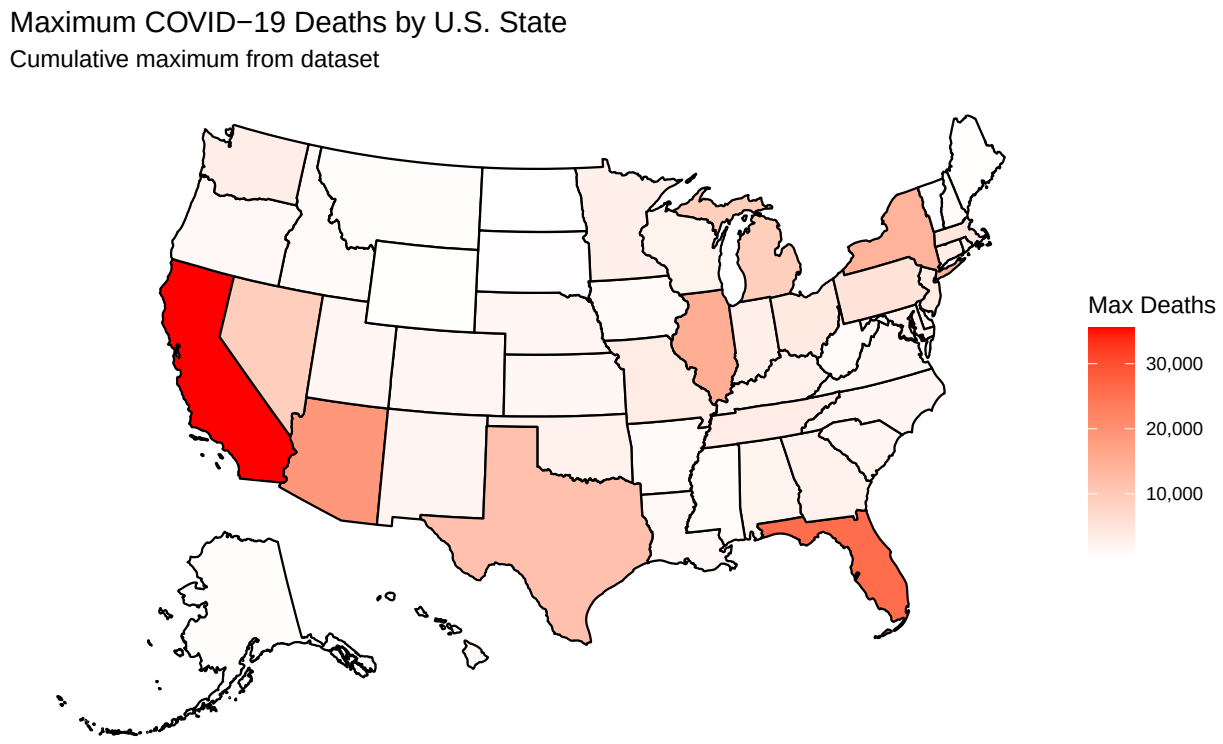
Most countries follow a similar trend. There is also an interesting spike in death cases at the beginning of 2022. This could be for multiple reasons like the holiday season or some major event. But considering that it happened across the globe, I assume that it was due to people loosening up restrictions.

Covid-19 in the US

```
US_deaths_by_state <- US_deaths %>%
  select(-UID, -iso2, -iso3, -code3, -FIPS, -Lat, -Long_, -Combined_Key, -Admin2) %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
    names_to = "Date",
    values_to = "Deaths"
  ) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%y")) %>%
  group_by(Province_State) %>%
  summarise(Max_Deaths = max(Deaths, na.rm = TRUE)) %>%
  mutate(state = Province_State)

plot_usmap(data=US_deaths_by_state, values="Max_Deaths", regions="states") +
  scale_fill_continuous(
    low = "white", high = "red",
    name = "Max Deaths",
    label = scales::comma
  ) +
```

```
  labs(
    title = "Maximum COVID-19 Deaths by U.S. State",
    subtitle = "Cumulative maximum from dataset"
  ) +
  theme(legend.position = "right")
```



From this graph, we can see that the California has the max deaths follow by Florida. But this graph does not show the population of each state.

Model

For my model, I want to do a basic linear regression to predict tomorrow's cases based on Today's cases. Since there is so much data, I will be filtering it to Los Angeles, California.

```
losangeles_data <- US_cases %>%
  select(-UID, -iso2, -iso3, -code3, -FIPS, -Lat, -Long_, -Combined_Key) %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
    names_to = "Date",
    values_to = "Cases"
  ) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%y")) %>%
  filter(Admin2 == "Los Angeles", Province_State=='California') %>%
```

```
  arrange(Date)
losangeles_data
```

```
## # A tibble: 1,143 x 5
##    Admin2      Province_State Country_Region Date       Cases
##    <chr>       <chr>          <chr>          <date>     <dbl>
##  1 Los Angeles California     US             2020-01-22     0
##  2 Los Angeles California     US             2020-01-23     0
##  3 Los Angeles California     US             2020-01-24     0
##  4 Los Angeles California     US             2020-01-25     0
##  5 Los Angeles California     US             2020-01-26     1
##  6 Los Angeles California     US             2020-01-27     1
##  7 Los Angeles California     US             2020-01-28     1
##  8 Los Angeles California     US             2020-01-29     1
##  9 Los Angeles California     US             2020-01-30     1
## 10 Los Angeles California     US             2020-01-31     1
## # i 1,133 more rows
```

Next I want to create a lag and model.

```
losangeles_data <- losangeles_data %>%
  mutate(
    daily_cases = Cases - lag(Cases),
    daily_yesterday = lag(daily_cases),
    daily_tomorrow = lead(daily_cases)
  ) %>%
  filter(!is.na(daily_cases), !is.na(daily_yesterday), !is.na(daily_tomorrow))
```

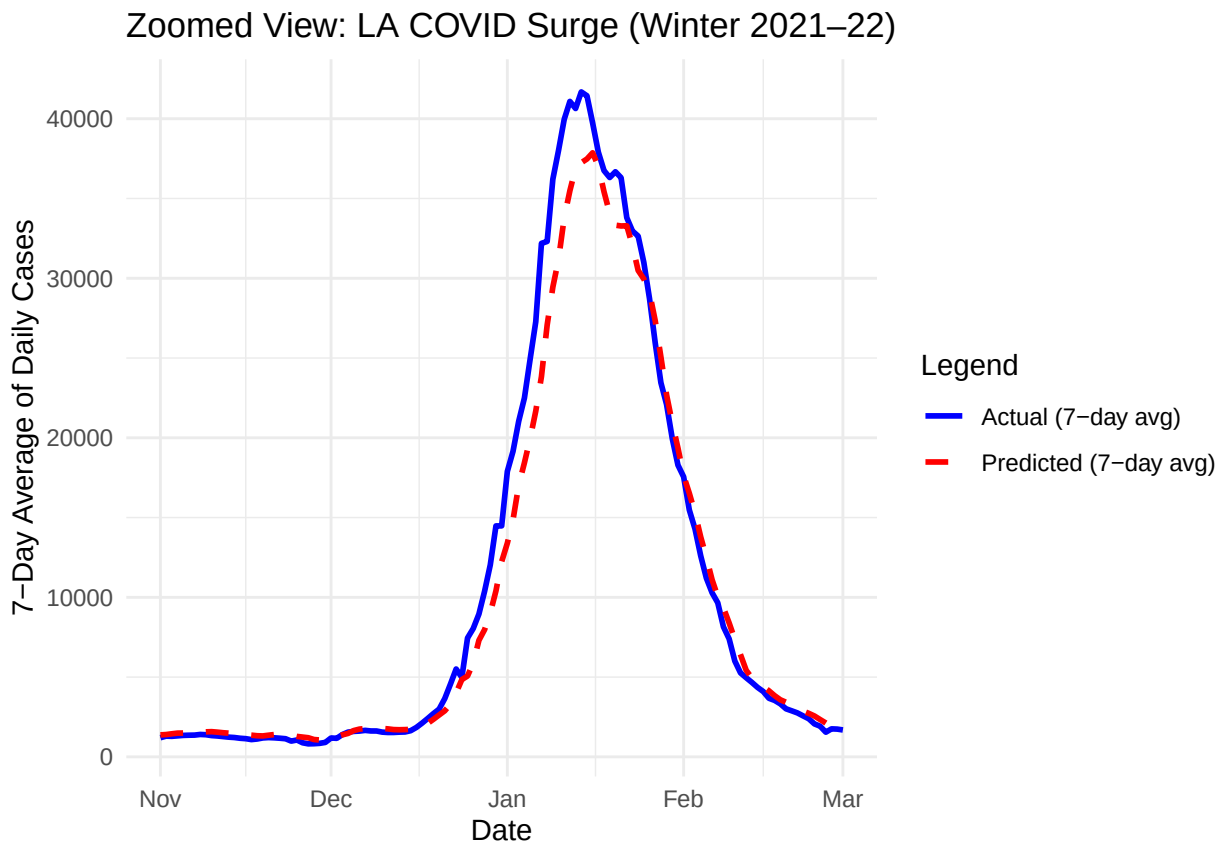
```
model <- lm(daily_tomorrow ~ daily_cases + daily_yesterday, data = losangeles_data)
summary(model)
```

```
##
## Call:
## lm(formula = daily_tomorrow ~ daily_cases + daily_yesterday,
##     data = losangeles_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21486.6   -786.2   -318.8    166.1  31575.9
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     318.79265  106.20691   3.002  0.00274 **
## daily_cases       0.42933    0.02614  16.426  < 2e-16 ***
## daily_yesterday   0.47399    0.02614  18.134  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3138 on 1137 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7412
## F-statistic:  1632 on 2 and 1137 DF,  p-value: < 2.2e-16
```

```r
losangeles_data <- losangeles_data %>%
  mutate(predicted = predict(model, newdata = losangeles_data))

losangeles_data <- losangeles_data %>%
  mutate(
    actual_smoothed = stats::filter(daily_tomorrow, rep(1/7, 7), sides = 1),
    predicted_smoothed = stats::filter(predicted, rep(1/7, 7), sides = 1)
  )

ggplot(losangeles_data %>%
  filter(Date >= as.Date("2021-11-01") & Date <= as.Date("2022-03-01")),
  aes(x = Date)) +
  geom_line(aes(y = actual_smoothed, color = "Actual (7-day avg)"), linewidth = 1) +
  geom_line(aes(y = predicted_smoothed, color="Predicted (7-day avg)"), linetype="dashed", linewidth=1)
  scale_color_manual(values = c("Actual (7-day avg)" = "blue", "Predicted (7-day avg)" = "red")) +
  labs(
    title = "Zoomed View: LA COVID Surge (Winter 2021-22)",
    y = "7-Day Average of Daily Cases", x = "Date", color = "Legend"
  ) +
  theme_minimal()
```



After a couple of different graphs, I decided to settle with doing a 7-day rolling average with the graph only showing one of the multiple spikes in Covid-19 cases. We can see here that the prediction does a pretty good job at predict the amount of cases.

Conclusion

Throughout this report, I identified the different tables that were imported and pivoted those tables to a more readable format. Using the new tables, I was able to create graphs to show Covid cases and deaths over time as well as a map of total Covid cases in the US. I also fitted a linear regression model

Bias

A big bias is reporting bias. Some countries may under report due to limited testing or political pressure. Some people might not even go out to test when they get Covid but its more likely that Covid death would be reported. Covid testing was also unreliable for a long time early on. There would be a lot of false negative.