SURV675: Assignment 3

Riki Engling

2025-05-08

Loading libraries

```
library(sparklyr)
library(tidyverse)
library(lubridate)
library(haven)
library(DBI)
library(dbplot)
library(corrr)
```

Load, Save, and Prep the Data

```
#Read in CSV files
uid_table <- read.csv("C:\\Users\\Owner\\Downloads\\UID_ISO_FIPS_LookUp_Table.csv")
time_series <- read_csv("C:\\Users\\Owner\\Downloads\\time_series_covid19_confirmed_global.csv")

#Create a Working Copy of the Data and Pivot it
time_long <- time_series %>%
  pivot_longer(
    cols = !(1:4),
    names_to = "Date",
    values_to = "Confirmed_COVID_Cases"
  )

#Making R Read the "Date" column as actual dates
time_long <- time_long %>%
  mutate(
    Date = lubridate::mdy(time_long$Date))

#Creating the "number of days since start" variable
time_long$Days <- as.numeric(time_long$Date - min(time_long$Date)) + 1
```

Connect to local Spark server

```r
#Specify configuration settings due to my computer struggling
config <- spark_config()

#Increasing memory of driver and executor
config$spark.driver.memory <- "14G"
config$spark.executor.memory <- "8G"

#Connect to server
sc <- spark_connect(master = "local", version = "3.5", config = config)
```

Move data to Spark

```r
#Send to server
uid_spark <- copy_to(sc, uid_table)
time_spark <- copy_to(sc, time_long)

#filter time series data to only include the data for specified countries
filtered_spark <- time_spark %>%
  filter(CountryRegion %in% c("United Kingdom", "Japan", "US", "Brazil", "Mexico")) %>%
    compute("filtered_spark")

#renaming uid columns so they match time series
uid_named <- uid_spark %>%
  rename("Long" = "Long_",
         "ProvinceState" = "Province_State",
         "CountryRegion" = "Country_Region") %>%
    compute("uid_named")

#Recoding blank cells to have NA
uid_clean <- uid_named %>%
  mutate(`ProvinceState` = ifelse(`ProvinceState` == "", NA, `ProvinceState`),
         `Admin2` = ifelse(`Admin2` == "", NA, `Admin2`)) %>%
    compute("uid_clean")

#Filter specified countries in uid dataset
uid_pop <- uid_clean %>%
    filter(CountryRegion %in% c("Brazil", "Japan", "Mexico", "US", "United Kingdom")) %>%
    compute("uid_pop")

#Create population sums
pop_sum <- uid_pop %>%
  group_by(CountryRegion) %>%
    summarise(Population = sum(Population, na.rm = TRUE)) %>%
      compute("pop_sum")

#Join datasets
full_pop <- filtered_spark %>%
  inner_join(pop_sum, by = "CountryRegion") %>%
  compute("full_pop")

#Calculate number of cases and rate of cases by country and day
```

```r
cases_spark <- full_pop %>%
    group_by(CountryRegion, Population, Date, Days) %>%
    summarize(Total_Cases = sum(Confirmed_COVID_Cases)) %>%
    mutate(Rate_of_Cases = Total_Cases/Population) %>%
      compute("cases_spark")
```

Data Visualization

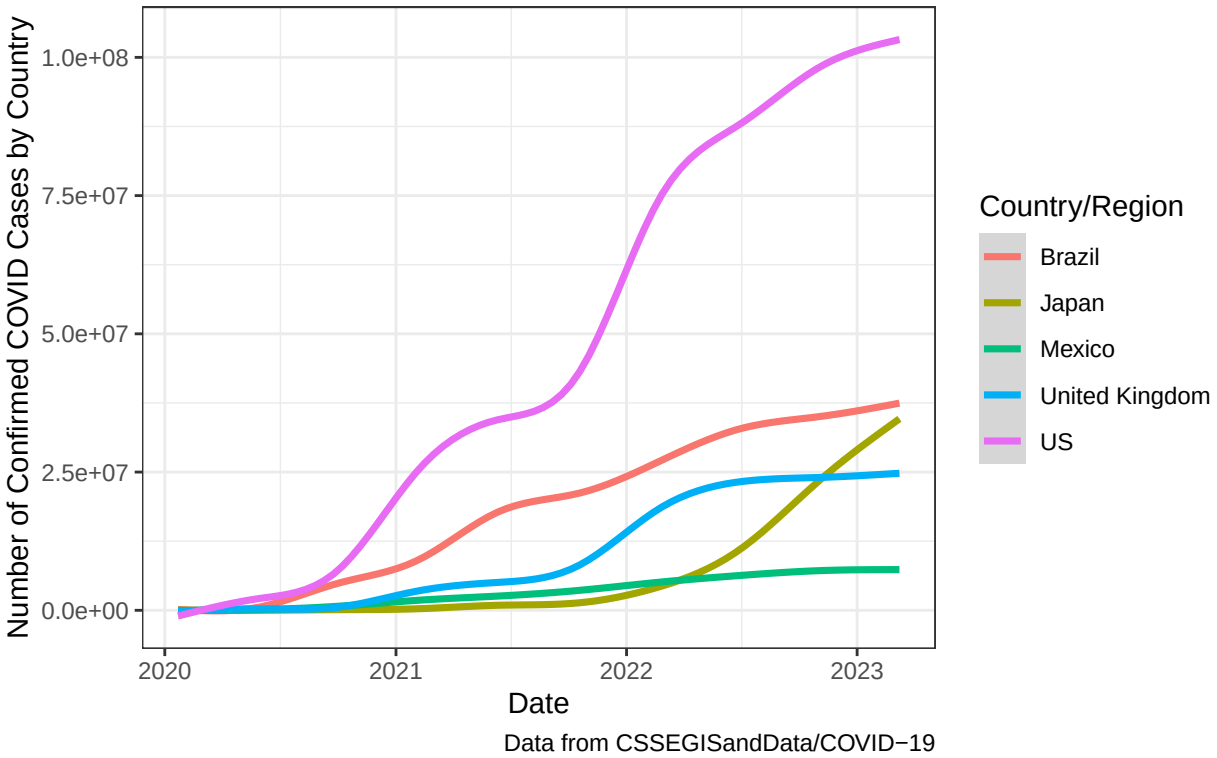
```r
#Graph Count by Country
Covid_Over_Time <- cases_spark %>%
  filter(!is.na(Total_Cases)) %>%
  ggplot(aes(x = Date, y = Total_Cases, color = `CountryRegion`)) +
    geom_smooth(linetype = 1,
      linewidth = 1.25) +
  theme_bw() +
    labs(x = "Date", y = "Number of Confirmed COVID Cases by Country",
        title = "Figure 1.
        Confirmed COVID-19 Cases Over Time",
        caption = "Data from CSSEGISandData/COVID-19",
        color = "Country/Region")

#Graph Rate by Country
Rate_Over_Time <- cases_spark %>%
  ggplot(aes(x = Date, y = Rate_of_Cases, color = `CountryRegion`)) +
    geom_smooth(linetype = 1,
      linewidth = 1.25) +
  theme_bw() +
    labs(x = "Date", y = "Rate of Infection",
        title = "Figure 2.
        Rate of Infection Over Time by Country",
        caption = "Data from CSSEGISandData/COVID-19",
        color = "Country/Region")
```

Interpretation

```r
print(Covid_Over_Time)
```

Figure 1.

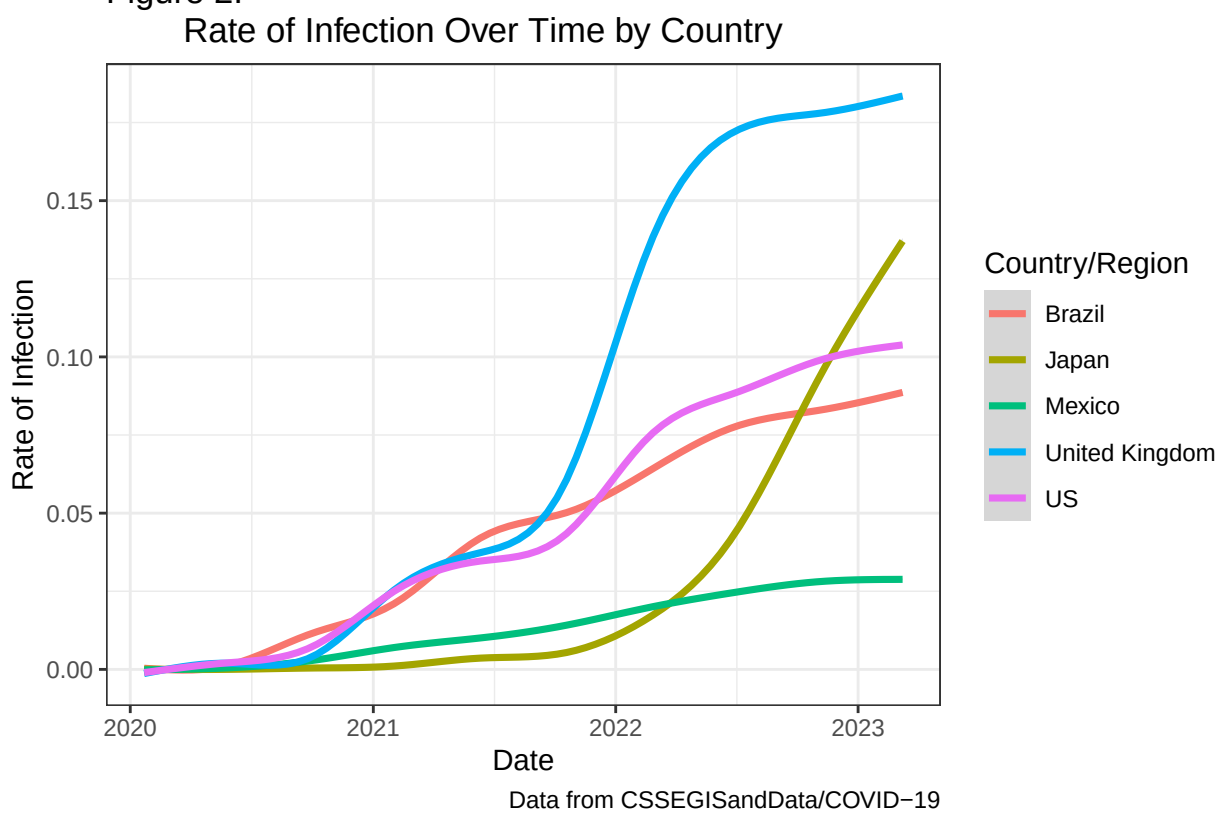


The graph of confirmed COVID cases over time by country reveals distinct patterns. For the first two years, Japan had the fewest cases, followed by Mexico and the UK. By mid-2022, Japan surpassed Mexico, and by early 2023, it exceeded the UK as well. Mexico, Brazil, and the UK saw a plateau between mid-2022 and 2023. Despite initially having fewer cases than Mexico, Japan's numbers eventually increased, while Mexico maintained a steady, slight incline.

Among countries with the highest case counts across multiple years, the US, Brazil, and the UK displayed similar trends of spikes and plateaus at comparable times, but with varying intensities. The US reported the highest numbers, with sharp winter spikes. Brazil also saw increased cases during winter, but less dramatically than the US. The UK had a modest rise in winter 2020-2021, plateaued throughout 2021, and experienced a significant spike during winter 2021-2022, nearly doubling by summer 2022, before stabilizing.

```
print(Rate_Over_Time)
```

Figure 2.



The graph of infection rates by country shows notable differences from the case count graph. Instead of the US leading, the UK reported the highest infection rate, with a sharp spike during the 2021-2022 winter and a gradual rise afterward. By late 2022, Japan had the second-highest rate, surpassing Brazil and the US. From 2020 to mid-2021, the US, UK, and Brazil vied for the highest rate, but by early 2022, the US had surpassed Brazil, while the UK long surpassed both. Unlike the other four countries, Mexico's infection rate closely mirrors its number of confirmed cases.

Modeling the Data

```
#Calculate log number of cases
log_spark <- cases_spark %>%
  mutate(Log_Cases = log1p(Total_Cases)) %>%
  ungroup() %>%
  compute("log_spark")

#Linear Regression
model <- log_spark %>%
  ml_linear_regression(Log_Cases ~ CountryRegion + Population + Days)
```

Interpretation

```
summary(model)
```

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4204  -0.5156   0.4499   1.2779   2.4403
##
## Coefficients:
##          (Intercept) CountryRegion_Brazil  CountryRegion_Japan
##         1.029341e+01         3.636107e-01        -1.487538e+00
## CountryRegion_Mexico     CountryRegion_US           Population
##        -1.120582e+00         8.736504e-01         9.367004e-10
##                 Days
##         7.629417e-03
##
## R-Squared: 0.6006
## Root Mean Squared Error: 2.243
```

The linear regression of the log number of cases provides some interesting information. The $beta_0$ value of 11.44 indicates that, when all predictor variables are equal to zero, the log number of cases is expected to be about 11. When examining the individual countries, the United States is expected to have a log of total cases about 8 cases higher than that of the UK ($beta_{US}$ = 8.15). Brazil also has a positive expected value compared to the UK ($beta_{Brazil}$ = 2.80). Mexico is expected to report similar log case totals as the UK, though slightly less ($beta_{Mexico}$ = -0.10). Japan is expected to report about 0.49 cases less than the United Kingdom ($beta_{Japan}$ = -0.49). Interestingly, the effect of population size is nearly negligable ($beta_{Pop}$ = -0.000000008). The number of days since the start of data collection did, however, have an impact on the log number of cases ($beta_{Days}$ = 0.008). These values should be interpreted with caution, though, as the model could fit the data better as it is currently off by about 2 units from the actual values ($RMSE$ = 2.24).