DTSA5301-FINAL-COVID19

N. Cataland

2/24/2025

Introduction

Though we can't quite yet assert COVID-19 is fully "behind us", we do now have the benefit of hindsight when trying to understand this disease and its vectors. One claim in particular that was frequently floated in the early days of the pandemic was that cases would increase during the winter months because more people would be gathering indoors. Now that we have a few years worth of data, I thought it would be interesting to look into if this panned out to be accurate or not. For our purposes, we will generalize out the hypothesis a bit to investigate if there is a statistically significant relationship between a geographical region's temperature and COVID-19 case count.

About the data

The analytic approach for this assignment was largely informed by the availability of quality (and free) data. Conveniently, the COVID-19 case data was provided for us, courtesy of Johns Hopkins Center for Systems Science and Engineering (CSSE).Though global case data is available, I've opted to limit the scope of my analysis to cases in the United States, as there may be customs or cultural practices in other countries that I am not familiar with that could influence the results.

This left our only remaining task to find a reliable source for historical U.S. weather stats. Fortunately for us, the National Oceanic and Atmospheric Administration (NOAA) has an entire branch dedicated to making environmental data widely available for organizations, researchers and the public alike–completely free of use! NOAA's National Centers for Environmental Information (NCEI) website includes many interactive visualizations to explore its wealth of climate data, and also links to download the data in a variety of formats for analysis in a program of your choice. (More on that later. . .)

Preliminaries

Before we can get to loading and analyzing our data, we need to first load the libraries that will be assisting us by providing some convenient utilities for data loading, processing, and visualization.

```
library(dplyr)
library(tidyr)
library(stringr)
library(readr)
library(lubridate)
library(knitr)
library(ggplot2)
```

Part 1 - Data Loading

Now that we have our libraries installed, let's get to work. Before we can get to analyzing, we need some data. We will first load and prepare the COVID-19 data, largely following the steps that our instructor,

Dr. Jane Wall, so kindly demonstrated for us.

```
us_confirmed_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_c
us_cases <- read_csv(us_confirmed_url, col_types = cols(.default = "c"))

us_cases <- us_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "Date",
               values_to = "Cases") %>%
  select(Admin2:Cases) %>%
  mutate(Date = mdy(Date)) %>%
  select(-c(Lat, Long_, Country_Region, Combined_Key))

names(us_cases)[1] <- "County"
names(us_cases)[2] <- "State"

head(us_cases)
```

```
## # A tibble: 6 x 4
##   County  State   Date       Cases
##   <chr>   <chr>   <date>     <chr>
## 1 Autauga Alabama 2020-01-22 0
## 2 Autauga Alabama 2020-01-23 0
## 3 Autauga Alabama 2020-01-24 0
## 4 Autauga Alabama 2020-01-25 0
## 5 Autauga Alabama 2020-01-26 0
## 6 Autauga Alabama 2020-01-27 0
```

```
us_deaths_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covi
us_deaths <- read_csv(us_deaths_url, col_types = cols(.default = "c"))

us_deaths <- us_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "Date",
               values_to = "Cases") %>%
  select(Admin2:Cases) %>%
  mutate(Date = mdy(Date)) %>%
  select(-c(Lat, Long_, Country_Region, Combined_Key))

names(us_deaths)[1] <- "County"
names(us_deaths)[2] <- "State"

head(us_deaths)
```

```
## # A tibble: 6 x 5
##   County  State   Population Date       Cases
##   <chr>   <chr>   <chr>      <date>     <chr>
## 1 Autauga Alabama 55869      2020-01-22 0
## 2 Autauga Alabama 55869      2020-01-23 0
## 3 Autauga Alabama 55869      2020-01-24 0
## 4 Autauga Alabama 55869      2020-01-25 0
## 5 Autauga Alabama 55869      2020-01-26 0
## 6 Autauga Alabama 55869      2020-01-27 0
```

More challenging however, was collecting our climate data. The NCEI website had exactly what we need: historical time series U.S. county temperature averages from a robust and reliable source. But alas, the dashboard only allows us to manually download one month of data at a time.

Good thing we data scientists are not so easily discouraged. When we inspect the download url, a rather simple pattern emerges! We are able to exploit this pattern to create a looping R function that will collect all of this data into a single dataframe.

NOTE: I've included the aforementioned function code below for those interested, but I updated this report to instead pull the downloaded climate data directly from my Github repository since the NCEI/NOAA sites have been experiencing frequent outages in recent weeks.

```
# # Initialize an empty dataset to hold our data
# climate_data <- data.frame()
#
# for (year in 2020:2023) {
#   for (month in 1:12) {
#     month_str <- sprintf("%02d", month)
#     url <- paste0("https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/110
#
#     tryCatch({
#       # The first 4 lines of each csv contains a description of the dataset that needs to be
#       temp_data <- read_csv(url, skip = 4)
#       temp_data <- temp_data[, 1:4]
#       colnames(temp_data) <- c("ID", "County", "State", "Avg_Temp")
#       temp_data$Date <- as.Date(paste(year, month_str, "01", sep = "-"))
#       climate_data <- bind_rows(climate_data, temp_data)
#
#     }, error = function(e) {
#       message(paste("Failed to download data:", url))
#     })
#   }
# }
#
# head(climate_data)
# write.csv(climate_data, "NOAA_Average_US_County_Temps.csv", row.names = FALSE)

climate_data_url <- "https://raw.githubusercontent.com/nicolecataland/DTSA5301-Reports/refs/heads/main/

climate_data <- read_csv(climate_data_url, col_types = cols(.default = "c"))

climate_data <- climate_data %>%
  mutate(Date = ymd(Date)) %>%
  select(-c(ID))

head(climate_data)
```

```
## # A tibble: 6 x 4
##   County         State   Avg_Temp Date
##   <chr>          <chr>   <chr>    <date>
## 1 Autauga County Alabama 50.7     2020-01-01
## 2 Baldwin County Alabama 54.8     2020-01-01
## 3 Barbour County Alabama 51.6     2020-01-01
## 4 Bibb County    Alabama 48.6     2020-01-01
## 5 Blount County  Alabama 47.1     2020-01-01
```

```
## 6 Bullock County Alabama 50.6     2020-01-01
```

Part 2 - Data Tidying

Now we that have all of our climate and case data loaded, we just need to do a bit more work to massage the datasets so that they play nicely together.

Since the John Hopkins CSSE data repository was archived on March 10th, 2023, we will filter our data to the beginning of 2020 through February 2023. My planned analysis is mostly concerned with confirmed COVID-19 cases, but I loaded the COVID-19 deaths data as well so we can use the "Population" field it contains. We will use that data shortly for computing "cases per capita" to normalize case counts across different county populations.

```
us_cases <- us_cases %>%
  filter(Date >= ymd("2020-01-01") & Date <= ymd("2023-02-28")) %>%
  mutate(Year = year(Date), Month = month(Date)) %>%
  mutate(Year = as.numeric(Year), Month = as.numeric(Month))

us_deaths <- us_deaths %>%
  filter(Date >= ymd("2020-01-01") & Date <= ymd("2023-02-28")) %>%
  mutate(Year = year(Date), Month = month(Date)) %>%
  mutate(Year = as.numeric(Year), Month = as.numeric(Month))

population_data <- us_deaths %>%
  select(State, County, Population) %>%
  distinct()

us_cases <- left_join(us_cases, population_data, by = c("State", "County"))
```

To be able to join our climate and case tables, we need to remove the word "County" from the county names column in my climate data. Also, since our climate data contains monthly averages, we need to aggregate our COVID-19 case data by month before we can combine the tables.

```
climate_data <- climate_data %>%
  filter(Date >= ymd("2020-01-01") & Date <= ymd("2023-02-28")) %>%
  mutate(Year = year(Date), Month = month(Date)) %>%
  mutate(Year = as.numeric(Year), Month = as.numeric(Month))

climate_data$County <- gsub(" County", "", climate_data$County)

us_cases$Cases <- as.numeric(us_cases$Cases)

monthly_cases <- us_cases %>%
  group_by(State, County, Population, Year, Month) %>%
  summarize(Monthly_Cases = sum(Cases), .groups="drop")

monthly_cases <- monthly_cases %>%
  mutate(Population = as.numeric(Population)) %>%
  mutate(Cases_Per_100k = (Monthly_Cases / Population) * 100000)

monthly_cases <- monthly_cases %>%
  mutate(Cases_Per_10k = (Monthly_Cases / Population) * 10000)
```

```
monthly_climate_cases <- inner_join(monthly_cases, climate_data, by = c("State", "County", "Year", "Mon

head(monthly_climate_cases)
```

```
## # A tibble: 6 x 10
##   State County Population  Year Month Monthly_Cases Cases_Per_100k Cases_Per_10k
##   <chr> <chr>       <dbl> <dbl> <dbl>         <dbl>          <dbl>         <dbl>
## 1 Alab~ Autau~      55869  2020     1             0              0             0
## 2 Alab~ Autau~      55869  2020     2             0              0             0
## 3 Alab~ Autau~      55869  2020     3            46           82.3          8.23
## 4 Alab~ Autau~      55869  2020     4           722         1292.          129.
## 5 Alab~ Autau~      55869  2020     5          3661         6553.          655.
## 6 Alab~ Autau~      55869  2020     6         11455        20503.         2050.
## # ... with 2 more variables: Avg_Temp <chr>, Date <date>
```

Part 3 - Data Analysis

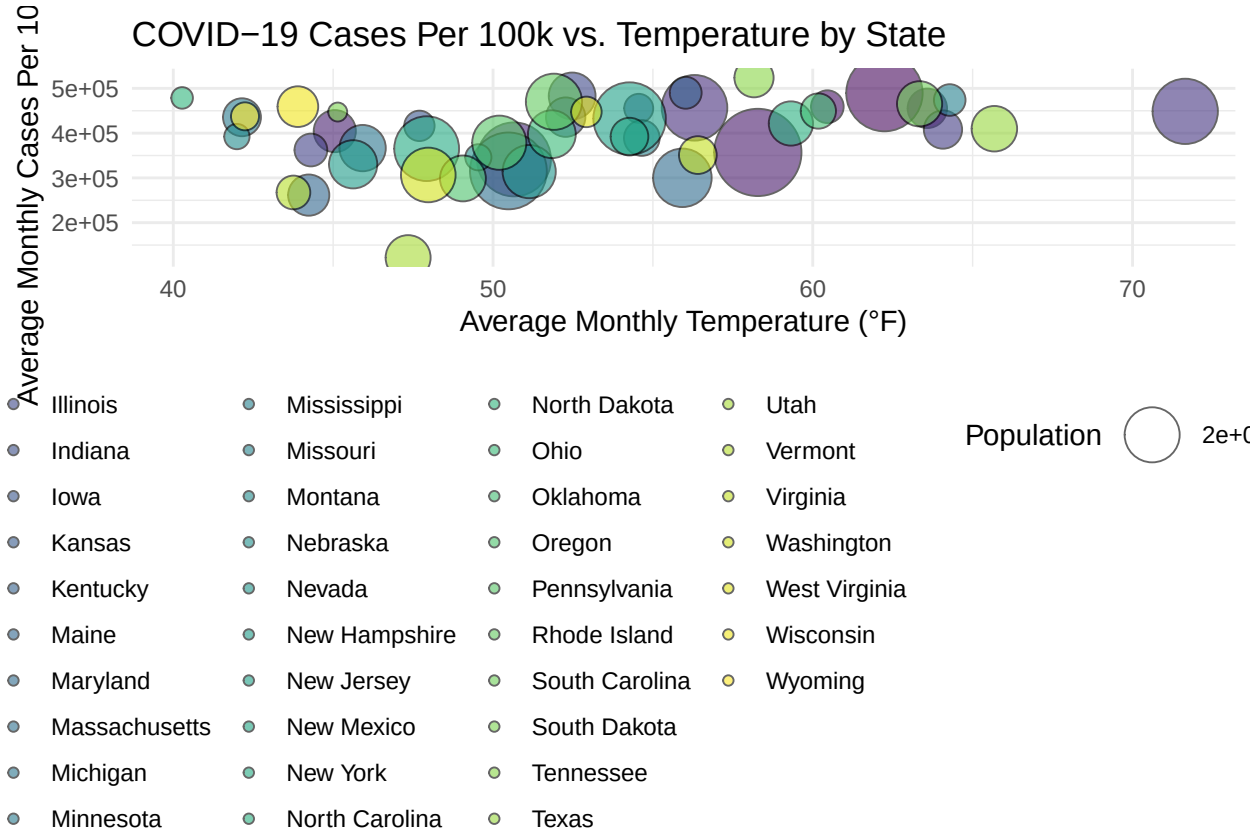
Before diving into the more fine-grained county level analysis, let's aggregate our data by state to see if we can spot any broader trends. After aggregating the data, we will plot it using a bubble graph to see if there are any obvious tendencies between average temps and case counts.

(If you have Plotly installed, I recommend uncommenting the next below chunk to view the interactive version of the graph instead!)

```
monthly_climate_cases$Avg_Temp <- as.numeric(monthly_climate_cases$Avg_Temp)

bubble_data <- monthly_climate_cases %>%
  filter(!is.na(Cases_Per_100k) & !is.na(Avg_Temp) & !is.na(Population)) %>%
  group_by(State) %>%
  summarise(
    Avg_Cases_Per_100k = mean(Cases_Per_100k, na.rm = TRUE),
    Avg_Temp = mean(Avg_Temp, na.rm = TRUE),
    Population = mean(Population, na.rm = TRUE),
    .groups = "drop"
  )

ggplot(bubble_data, aes(x = Avg_Temp, y = Avg_Cases_Per_100k, size = Population, fill = State)) +
  geom_point(alpha = 0.6, shape = 21, color = "black") +
  scale_size(range = c(3, 15)) +
  scale_fill_viridis_d() +
  labs(title = "COVID-19 Cases Per 100k vs. Temperature by State",
       x = "Average Monthly Temperature (°F)",
       y = "Average Monthly Cases Per 100,000",
       size = "Population",
       fill = "State") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



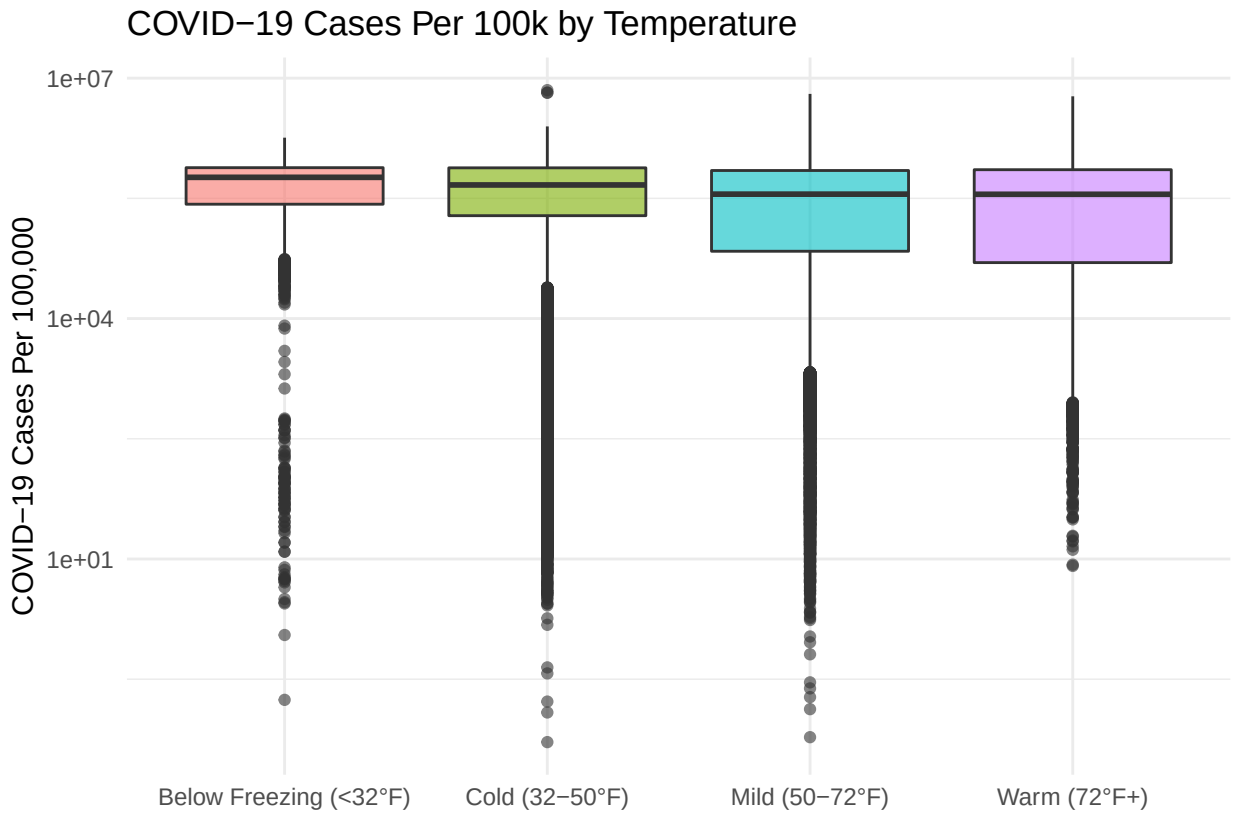

```
# Uncomment and run the below code to view an interactive version of the bubble graph.
# Installation of the plotly library is required!
#
# library(plotly)
# p <- ggplot(bubble_data, aes(x = Avg_Temp, y = Avg_Cases_Per_100k, size = Population, fill = State)) +
#   geom_point(alpha = 0.6, shape = 21, color = "black") +
#   scale_size(range = c(3, 15)) +
#   scale_fill_viridis_d() +
#   labs(title = "COVID-19 Cases Per 100k vs. Temperature by State",
#        x = "Average Monthly Temperature (°F)",
#        y = "Average Monthly Cases Per 100k",
#        size = "Population",
#        fill = "State") +
#   theme_minimal() +
#   theme(legend.position = "bottom")
# ggplotly(p, tooltip = "text")
```

The bubble graph might look neat, but unfortunately for our purposes, it wasn't very insightful.

To help see past the clutter, let's instead organize our monthly case counts into bins according to the monthly average temperature. We will visualize this data using a box plot so we can get an idea of the distribution of cases for each temperature category at a glance.

```
ggplot(monthly_climate_cases, aes(x = cut(Avg_Temp, breaks = c(-Inf, 32, 50, 72, Inf),
                                  labels = c("Below Freezing (<32°F)", "Cold (32-50°F)", "Mild (50-72°F)"
                                  y = Cases_Per_100k,
```

```
                              fill = cut(Avg_Temp, breaks = c(-Inf, 32, 50, 72, Inf),
                              labels = c("Below Freezing", "Cold", "Mild (50-72)", "Warm (72+)")))) +
  geom_boxplot(alpha = 0.6) +
  scale_y_log10() +
  labs(title = "COVID-19 Cases Per 100k by Temperature",
       x = "",
       y = "COVID-19 Cases Per 100,000") +
  theme_minimal() +
  theme(legend.position = "none")
```



Though again there does not appear to be any obvious trends in the data, there's a lot more information we can draw from this graph than our previous! The distribution of cases appears relatively consistent across temperature variations, with similar medians and interquartile ranges. That being said, the graph suggests that in below freezing conditions, case counts were generally more consistent and had less variability within that central range. However we must be conscious of potential bias at play: it's also possible that regions with extremely cold temperatures had fewer data points or a more uniform response to COVID-19, leading to a more compressed distribution.

Time to bring out the statistical big guns!

Part 4 - Data Modeling

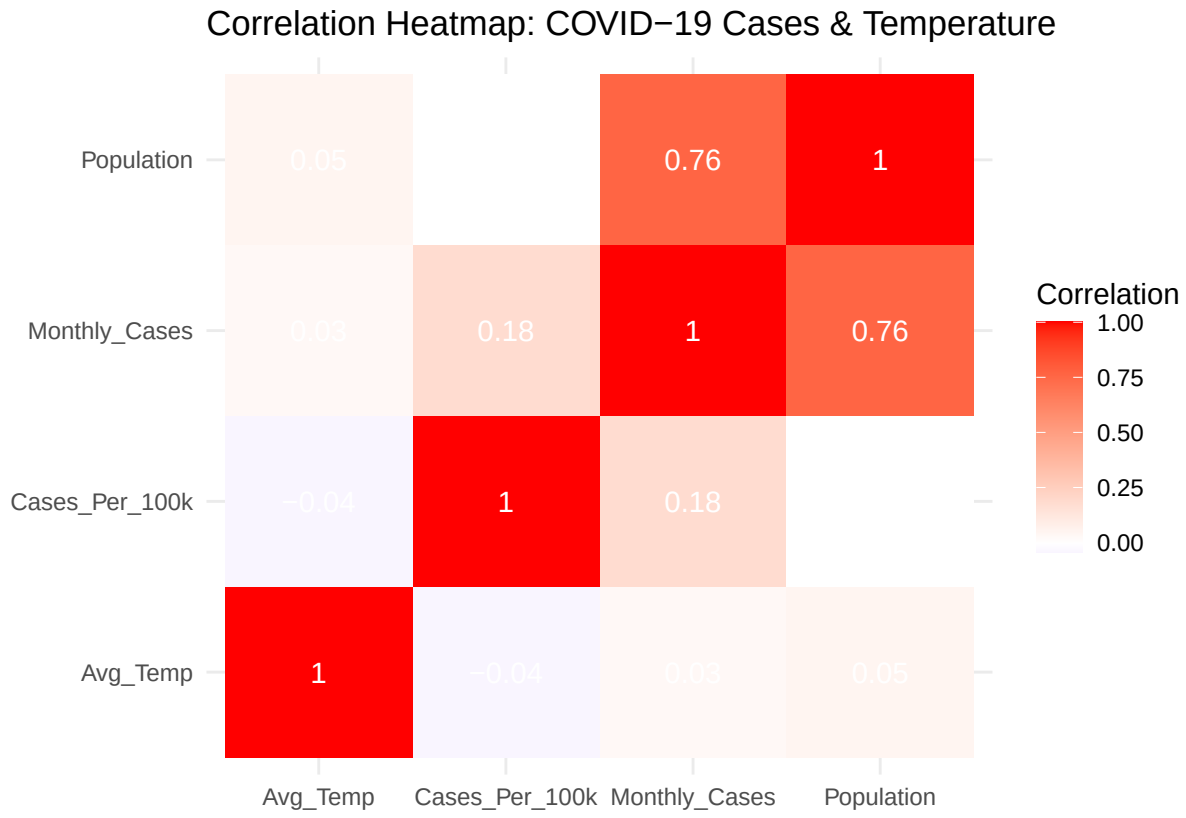
To add a little statistical rigor to our analysis, we will do two things. First, we will use R's built-in cor() function to compute the covariance of our case rates and plot it on a heat map. Second, we will compute a linear regression model to see if there is any predictive power of temperature on the COVID-19 case rate.

Between these two statistical techniques, any relationship between these two factors should reveal itself, if it does in fact exist.

```
cor_data <- monthly_climate_cases %>%
  select(Monthly_Cases, Cases_Per_100k, Avg_Temp, Population)

cor_matrix <- cor(cor_data, use = "complete.obs")
cor_melted <- as.data.frame(cor_matrix) %>%
  mutate(Variable1 = rownames(.)) %>%
  pivot_longer(-Variable1, names_to = "Variable2", values_to = "Correlation")

ggplot(cor_melted, aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile() +
  geom_text(aes(label = round(Correlation, 2)), color = "white", size = 4) +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  labs(title = "Correlation Heatmap: COVID-19 Cases & Temperature",
       x = "", y = "") +
  theme_minimal()
```




```
model <- lm(Cases_Per_100k ~ Avg_Temp, data = monthly_climate_cases)
summary(model)
```

```
##
## Call:
## lm(formula = Cases_Per_100k ~ Avg_Temp, data = monthly_climate_cases)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -453420 -346548  -55206  291335 6659377
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 460420.98    3357.45  137.13   <2e-16 ***
## Avg_Temp      -833.50      58.56  -14.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 351300 on 114188 degrees of freedom
## Multiple R-squared:  0.001771,   Adjusted R-squared:  0.001763
## F-statistic: 202.6 on 1 and 114188 DF,  p-value: < 2.2e-16
```

The correlation heatmap and regression model appear to tell us the same thing: temperature can not explain COVID-19 case variability. Even though the p-value of our model is small which suggests there does exist a statistical relationship, due to the high standard error (351,300!) the model is of no value to us for any practical purpose. However, the slope of our regression (-833.5) suggests that higher temperatures are weakly associated with fewer cases, but more analysis would be needed to confirm such suspicions.

Part 5 - Conclusion and Bias Statement

Though the idea that COVID-19 cases would increase in colder weather intuitively makes sense on its face, our findings did not support this claim. The results of our analysis showed that higher temperatures were linked to slightly fewer cases, but the effect was very small. This means that while temperature might play a role, it's not a strong predictor of COVID-19 spread on its own. Other factors like population density and public health policies likely had a much bigger impact, highlighting just how complex COVID-19 transmission really is.

All that being said, there are some important limitations and sources of bias to consider. Case reporting varied between states, weather data wasn't always consistent, and key factors like mask mandates and vaccination rates weren't included in this analysis. Plus the model assumed a simple relationship between temperature and cases, when in reality the connection could be much more complicated. Moving forward, a better approach would include more data on government policies and other environmental factors to get a clearer picture of what really drives COVID-19 trends.