ECO 322 Research Project #1

March 2022

Stony Brook University

COVID-19 Analysis: The Effectiveness of Restrictions in the United States

By Jesse Freitag, Kailey Ali, Gavin Vergara, Charlie Clark, Qiushi Yin

Stony Brook University

Introduction

Jesse: Overall Manager, R-Coder

Charlie: Head Coder, Quality Control

Kailey: Note Taker, Quality Control

Gavin: Note Taker, Quality Control

Qiushi: R-Coder, Note Taker

Interest: It has been widely debated the effectiveness of lockdowns and restrictions on the number of COVID-19 cases throughout the United States. We seeked to discover this answer by exploring different characteristics of the pandemic via R's COVID-19 database.

Academic Studies on Lockdown Significance

The impact of lockdown timing on COVID-19 transmission across US counties, (Huang et al., *The Lancet*)

- Concludes that lockdowns are effective at controlling the spread of COVID-19 cases.

Lockdowns were not effective on COVID-19 spread (Richards J, Briggs W., and Axe D. 2020)

- To judge from the evidence, the answer is clear: Mandated lockdowns had little effect on the spread of the coronavirus. The charts below show the daily case curves for the United States as a whole and for thirteen U.S. states. As in almost every country, we consistently see a steep climb as the virus spreads, followed by a transition (marked by the gray circles) to a flatter curve. At some point, the curves always slope downward, though this wasn't obvious for all states until the summer.

Our Hypothesis: There is a statistically significant difference in cases during stricter lockdown periods and increased restriction levels. Stricter lockdown and restriction policies are associated with lower COVID-19 transmission.

COVID 19- Package in R

The package contains regional, day-by-day data on COVID cases throughout the world, on the country, state, and local levels, from March 2020 to the present day.

Includes 47 relevant variables such as cumulative cases, deaths, hospitalizations, number of people fully vaccinated, updated continuously as new data is found.

Policy measures are included in the data, which are specific factors of the stringency index. They are measured from zero, no measures taken, up to a maximum 5, the most stringent (see next slide).

Our major independent variable was the stringency index, which is the overall strictness of lockdown measures in a given area. Our major dependent variable was the number of cases.

Stringency Index

The stringency index is the arithmetic mean of 9 sub-indices calculated for each of the following indicator variables from the data:

- C1 – School closing $\in \{0,1,2,3,\mathit{Blank}\}$
- C2 – Workplace closing $\in \{0,1,2,3,\mathit{Blank}\}$
- C3 – Cancel Public Events $\in \{0,1,2,\mathit{Blank}\}$
- C4 – Gathering Restrictions $\in \{0,1,2,3,4,\mathit{Blank}\}$
- C5 – Public Transport Closed $\in \{0,1,2,\mathit{Blank}\}$
- C6 – Stay at Home Requirements $\in \{0,1,2,3,\mathit{Blank}\}$
- C7 – Internal Movement Restrictions $\in \{0,1,2,\mathit{Blank}\}$
- C8 – International Movement Restrictions $\in \{0,1,2,3,4,\mathit{Blank}\}$
- H1 – Public Information Campaigns $\in \{0,1,2,\mathit{Blank}\}$

R Libraries Used

library(COVID19) - provides access to the COVID-19 data package.

library(data.table) - creates data tables more conveniently than ‘data.frame’, allows for modifying columns of data without making copies of the data table.

library(ggplot2) - a system for 'declaratively' creating graphics, based on “The Grammar of Graphics”. You provide the data, tell 'ggplot2' how to map variables to aesthetics, and it takes care of the details.

library(sf) - ‘simple features’ helps with converting column-style data to spatial vector data.

library(tigris) - gives access to exact shapes of US states and cities according to the US Census Bureau. Used for applying spatial data to make map visuals.

library(mapview) - creates interactive map visuals of spatial data.

library(usmap) - allows for making maps of the entire United States, including Alaska and Hawaii.

Problems Encountered

1. Cities inherit restriction levels from state, so a city-by-city analysis is not optimal.
2. Needed to change cumulative cases and deaths to daily cases and deaths
3. Florida had an outlier of a daily case total of less than -40,000
4. County population inherited from state
5. Boolean flag description was ambiguous; we were unsure if they implied that local level or state level policies were implemented first.

Solutions

1. Conducted a state-by-state analysis, instead.
2. Shifted the data using the shift() function by taking the difference between the previous day and current day cumulative cases, same process with daily deaths
3. Replaced all negative daily cases and deaths values with a value of zero
4. Used online census data to obtain population for Suffolk, Nassau County and New York City.
5. Focused strictly on stringency index for aggregate level of restrictions.

Assumptions

When calculating the case concentrations, we assumed that people did not test positive twice, which is why we sampled 2021 data (unlikely to test positive twice within 12 months). All cases are unique.

We assumed that the population of a state did not change from 2020 to 2022, as the population listed in the database remains constant from 2020-2022.

We assumed that the stringency index at any given moment accurately reflected statewide restrictions, and did not lag after a restriction was lifted.

Exploratory Data Analysis

Stony Brook University

Methodology – Retrieving the Data

Loaded the US COVID data by State into data.frame object and then converted it into a data.table object (below)

load_data.r

```r
USA <- covid19( country = c("United States"), level = 2, verbose = FALSE)
setDT(USA)
```

Using the shift() function, we set a new variable named ‘DailyCases’ in our data.table allowing us to operate more efficiently (below)

create_daily_cases.r

```r
# for daily_deaths, replaced "confirmed" with "deaths"...
USA[, previous := shift(confirmed, n = 1, type = "lag", fill = NA_INTEGER_), by = list(state)]

# for daily_deaths, replaced "daily_cases" with "daily_deaths" and "confirmed" with "deaths"...
USA[, daily_cases := confirmed - previous]

USA[, previous := NULL] # removes the previous feature from the data
```

Methodology – Retrieving the Data

Using a for loop and the assign() function, we were able to create fifty different data.table objects, one for each state (top right)

create_state_dts.r

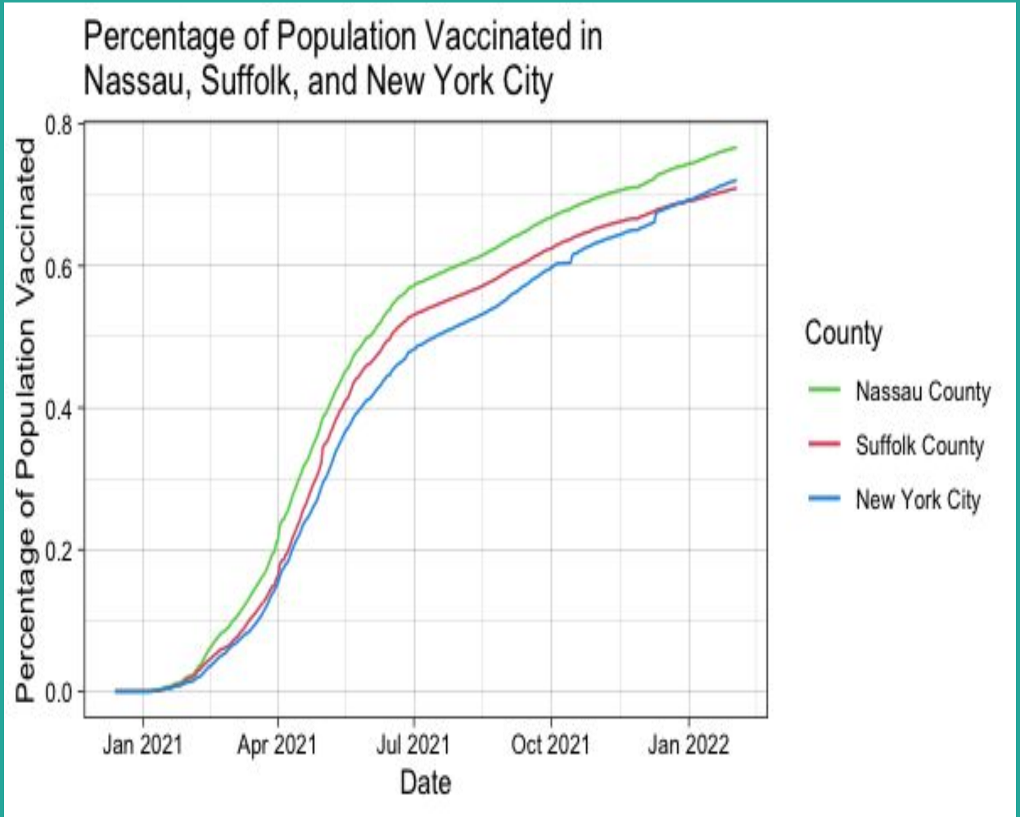
```r
x ← c(1:50)

# iterate through each state using a for loop...
for (i in x) {
  # use built-in state package...
  assign(state.abb[index], USA[state == state.name])
}
```

concentration_heatmap.r

```r
# create data.frame to be used in heat map...
df ← data.frame(
  states = state.name,
  fips = z,
  concentration = conc
)

# use usmap package's function plot_usmap to create heatmap...
# of the concentration of cases in each state...
# as a proportion of the state's population...
plot_usmap(
  data = df,
  values = "concentration",
  labels = TRUE,
  color = "dimgrey",
  size = 0.5
)
```

Utilizing the usmap library, we lined the case concentrations up with the FIPS code and obtained our GIS plot for each state (left)

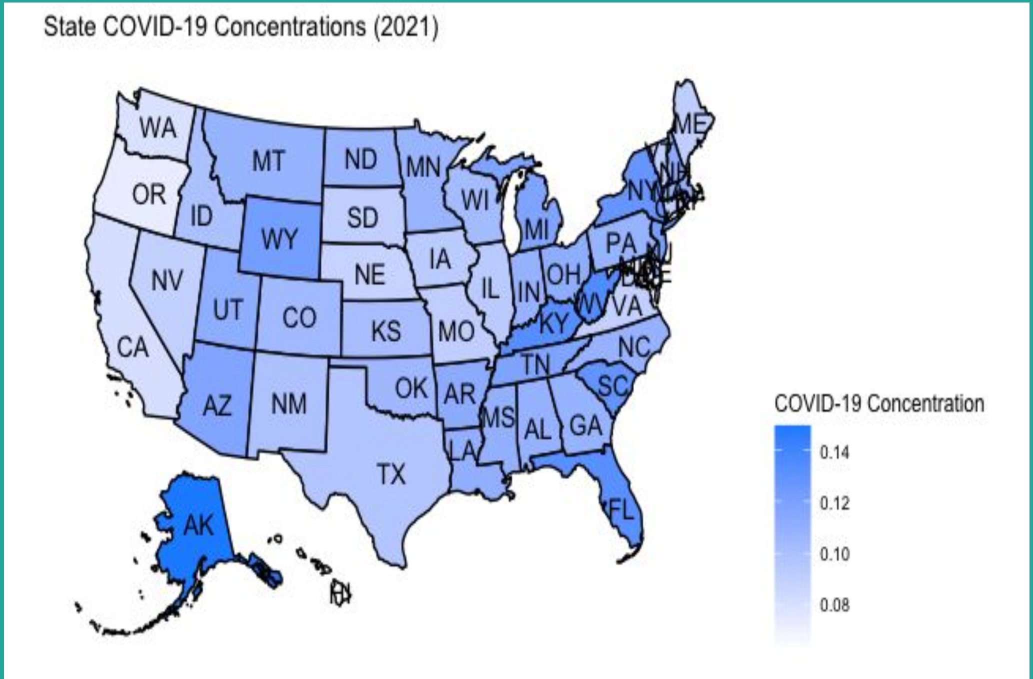
Percentage of Population Vaccinated in Suffolk County, Nassau County, and New York City throughout the pandemic starting from the release of the first COVID-19 Vaccine in December 2020

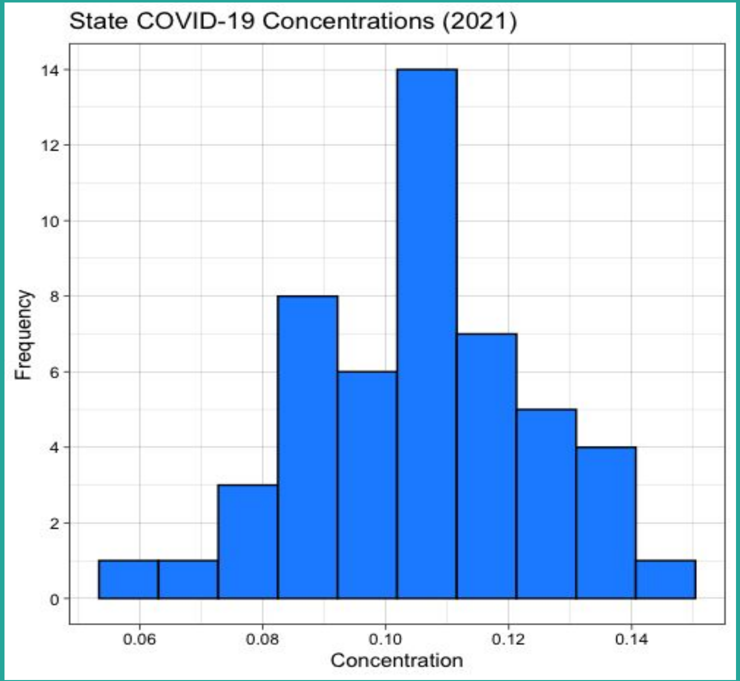


plot_local_vacc.r

```r
plot(Suffolk_perc_Vacc$date,
     Suffolk_perc_Vacc$Perc_Vacc,
     type = "l",
     col = 2,
     xlim = as.Date(c("2020-12-13", "2022-02-20")),
     xlab = "Date",
     ylab = "Percentage of Population Vaccinated",
     main = "Percentage of Population Vaccinated")

lines(Nassau_perc_Vacc$date,
      Nassau_perc_Vacc$Perc_Vacc,
      type = "l",
      col = 3)

lines(NYC_perc_Vacc$date,
      NYC_perc_Vacc$Perc_Vacc,
      type = "l",
      col = 4)

legend("topleft",
       c("Suffolk", "Nassau", "NYC"),
       lty = 1,
       col = 2:4)
```

Proportion of Population Affected by COVID-19 in each State from January 1st, 2021 to December 31, 2021.





The distribution of case concentrations follows a (surprisingly) approximate normal distribution with n = 50 samples.

Summary of Proportion of Population Affected by COVID by State in 2021

Top 5 States with the **most** COVID-19 Cases as Proportion of Population in 2021:

Alaska :	0.1495328
West Virginia :	0.1354956
Rhode Island :	0.1351258
Kentucky :	0.1333515
Florida :	0.1323736

Top 5 states with the **lowest** COVID-19 Cases as Proportion of Population in 2021:

Hawaii :	0.06223515
Maryland :	0.07012528
Oregon :	0.07286704
Washington :	0.07902409
Nebraska :	0.08118381

Mean : 0.1061

Median : 0.1086

Standard Deviation : 0.0180

Conclusion

State COVID Case Concentration as a Proportion of Population for 2021

Surprisingly, the distribution of the case concentrations followed an approximate normal distribution with $n = 50$ samples.

The average proportion of population affected by COVID for each state was around 10.6%, with the highest concentration in Alaska, with 14.95% of the population affected, and the lowest concentration in Hawaii, 6.22%.

Proportion of Population Vaccinated on Long Island and NYC

Currently, Nassau County has the highest proportion of residents that are vaccinated, followed by NYC, and lastly, Suffolk County.

Approximately 78% of residents in Nassau County are fully vaccinated and 72% of residents are fully vaccinated in Suffolk County and New York City as of February 20, 2022.

Main Experimentation

Methodology - Comparing Data from Periods of Strict and Lenient Regulations

Truncated the data to only include certain features from the year 2021 (top)

trim_data.r

```r
# only include date, state, daily_cases, daily_detaths, and stringency_index...
USA ← USA[, c("date", "state", "daily_cases", "daily_deaths", "stringency_index")

# trim the data to the year 2021: 2021-01-01 to 2021-12-31...
USA ← USA[date ⩾ "2021-01-01" & date ⩽ "2021-12-31", ]
```

Replaced negative daily cases values with a value of zero; followed same process for daily deaths, except “daily_cases” was replaced with “daily_deaths” in the code (bottom right)

replace_negatives.r

```r
# check how many records have negative daily cases...
length(USA[daily_cases < 0, ]$daily_cases)

# replace each such record with a 0 value...
USA[daily_cases < 0, daily_cases := 0]

# make sure all such records were corrected...
length(USA[daily_cases < 0, ]$daily_cases) # should print 0
```

Methodology - Comparing Data from Periods of Strict and Lenient Regulations

After assigning state abbreviations as variables (shown before), made sure none of the states had missing (NA or NULL) stringency_index, daily_cases, and daily_deaths records (top right)

Louisiana, Maryland, and Rhode Island all had NA/NULL values in stringency_index, so the mean stringency_index for each state (excluding the NA/NULL values) were calculated and used in imputation (Louisiana's imputation below)

check_loop.r

```
# loop through each state data.table variable...
# replace "stringency_index" with "daily_cases" and "daily_deaths"
# for other two check loops...
for (index in state_indices) {
  state_ptr ← get(state.abb[index])

  # prints FALSE if no NA or NULL values...
  print(paste(state.name[index],
              any(is.na(state_ptr$stringency_index)),
              sep = "⟶"))
}
```

impute.r

```
LA_na_indices ← which(is.na(LA$stringency_index),
                      arr.ind = TRUE)) # LA only had one

LA$stringency_index[LA_na_indices] =
  mean(LA$stringency_index[LA_na_indices - 15:LA_na_indices + 15],
       na.rm = TRUE)

any(is.na(LA$stringency_index)) # should print FALSE
```

None of the states had missing daily_cases or daily_deaths values, so no imputation was needed for them

Methodology - Comparing Data from Periods of Strict and Lenient Regulations

Defined a function to create a categorical feature in the passed data.table:

for each state

if a given record had stringency_index $\leq$ mean stringency_index,

then the value of stringency_category was set to “Less than or Equal to Average.”

otherwise it was set to “Greater than Average” (right)

create_stringency_category.r

```r
create_stringency_category <- function(state_dt) {
  stringency_category_vector <- vector()

  for index in c(1:length(state_dt$stringency_index)) {
    if (state_dt$stringency_index[index] <= mean(state_dt$stringency_index)) {
      stringency_category_vector[length(stringency_category_vector) + 1] <-
        "Less than or Equal to Average"
    }
    else {
      stringency_category_vector[length(stringency_category_vector) + 1] <-
        "Greater than Average"
    }
  }

  setDT(state_dt)
  state_dt[, stringency_category := stringency_category_vector]

  return (state_dt)
}
```

Methodology - Comparing Data from Periods of Strict and Lenient Regulations

Created lenient and strict data.tables for every existing state variable (top right)

Looped through each state and performed a 2-sided T-test on the average number of daily cases associated with the lenient and strict data.tables

The null hypothesis was that lenient mean daily cases equaled strict mean daily cases (bottom right)

create_lenient_and_strict.r

```
# note: state_indices previously defined as c(1:50)...
for (index in state_indices) {
  assign(state.abb[index], create_stringency_category(get(state.abb[index])))

  assign(paste(state.abb[index], "lenient",     = "_"),
         setDT(get(state.abb[index])[stringency_category == "Less than or Equal to Average"]))

  assign(paste(state.abb[index], "strict",     = "_"),
         setDT(get(state.abb[index])[stringency_category == "Greater than Average"]))
}
```

perform_ttests.r

```
# note: state_indices was previously defined as c(1:50)...
for (index in state_indices) {
  assign(paste(state.abb[index], "diff_ttest_obj", sep = "_"),
         t.test(
           get(paste(state.abb[index], "lenient", sep = "_"))$daily_cases,
           get(paste(state.abb[index], "strict", sep = "_"))$daily_cases
         ))
}
```

Methodology - Comparing Data from Periods of Strict and Lenient Regulations

evaluate_ttest.r

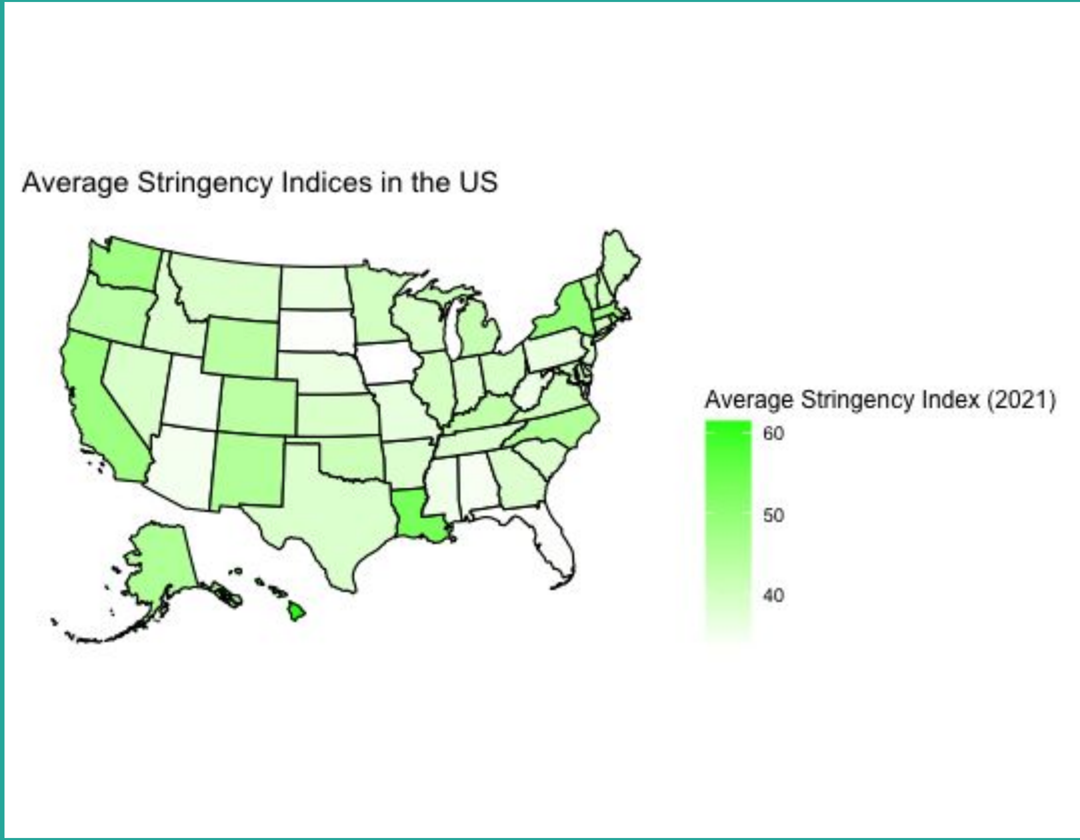
```r
for (index in state_indices) {
  if (get(paste(state.abb[index], "diff_ttest_obj", sep = "_"))$p.value < 0.01) {
    difference_states_large_alpha[length(difference_states_large_alpha) + 1] <-
      state.abb[index]
    difference_states_middle_alpha[length(difference_states_middle_alpha) + 1] <-
      state.abb[index]
    difference_states_small_alpha[length(difference_states_small_alpha) + 1] <-
      state.abb[index]
  }
  else if (get(paste(state.abb[index], "diff_ttest_obj", sep = "_"))$p.value < 0.05) {
    difference_states_large_alpha[length(difference_states_large_alpha) + 1] <-
      state.abb[index]
    difference_states_middle_alpha[length(difference_states_middle_alpha) + 1] <-
      state.abb[index]
  }
  else if (get(paste(state.abb[index], "diff_ttest_obj", sep = "_"))$p.value < 0.10) {
    difference_states_large_alpha[length(difference_states_large_alpha) + 1] <-
      state.abb[index]
  }
  else {
    no_difference_states[length(no_difference_states) + 1] <- state.abb[index]
  }
}
```

Iterated through each state's T-test object to determine if there was a statistically significant difference between strict and lenient mean daily cases at the 0.01, 0.05, and 0.10 levels of significance.

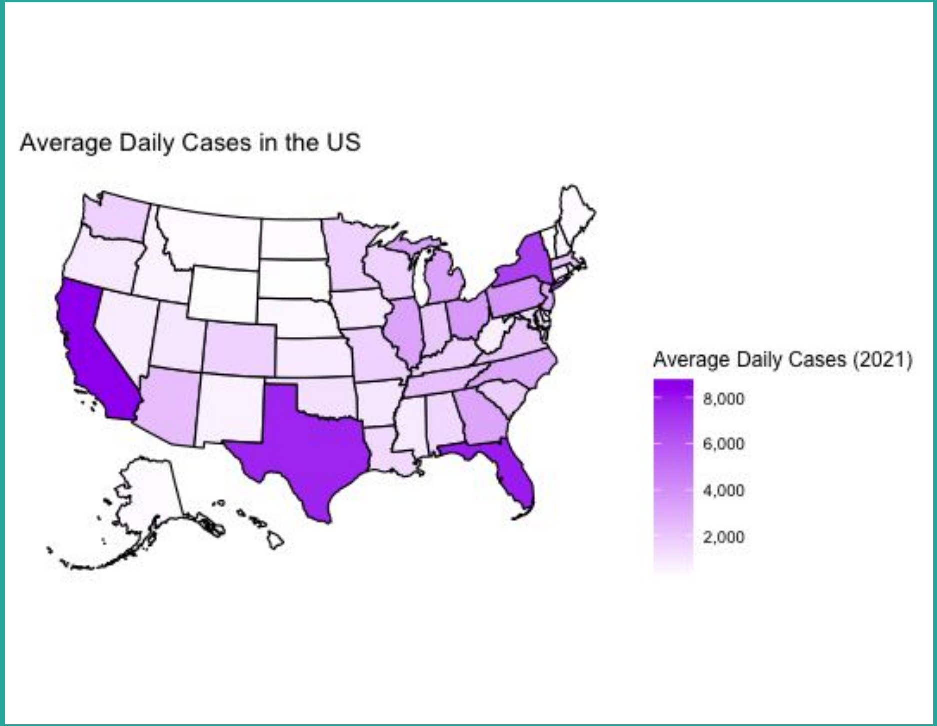
Performed similar processes to create and evaluate 1-sided T-tests for each state:

- one with an alternative hypothesis that mean daily cases during periods of lenient restrictions was statistically less than that of periods with strict restrictions
- the other with an alternative hypothesis that mean daily cases during periods of lenient restrictions was statistically greater than that of periods with strict regulations (left)

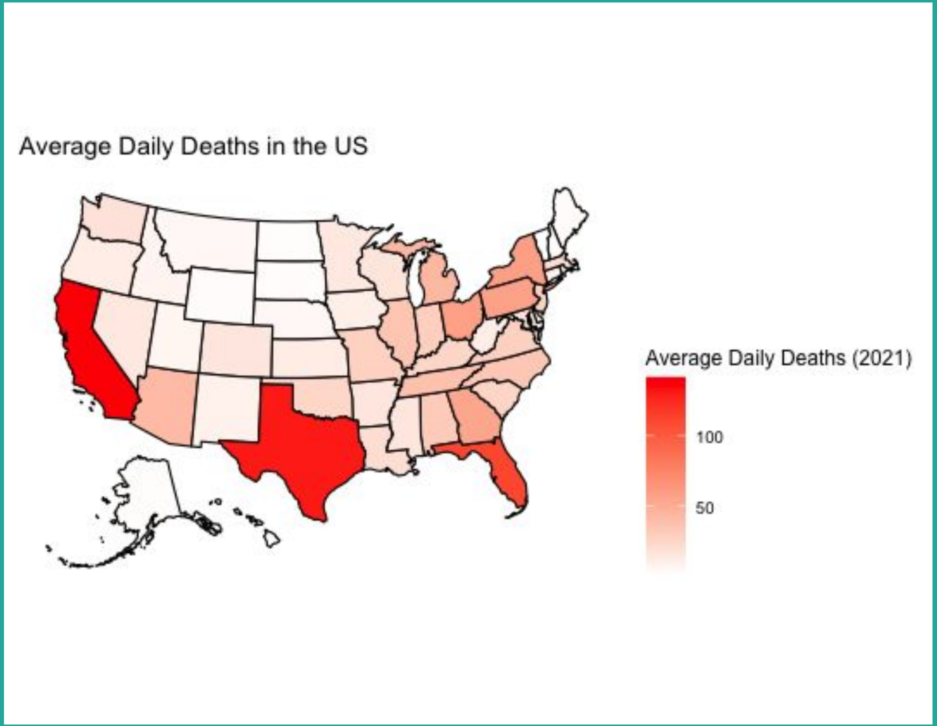
Average Stringency Index by State (2021)



Average Daily Cases and Deaths by State (2021)

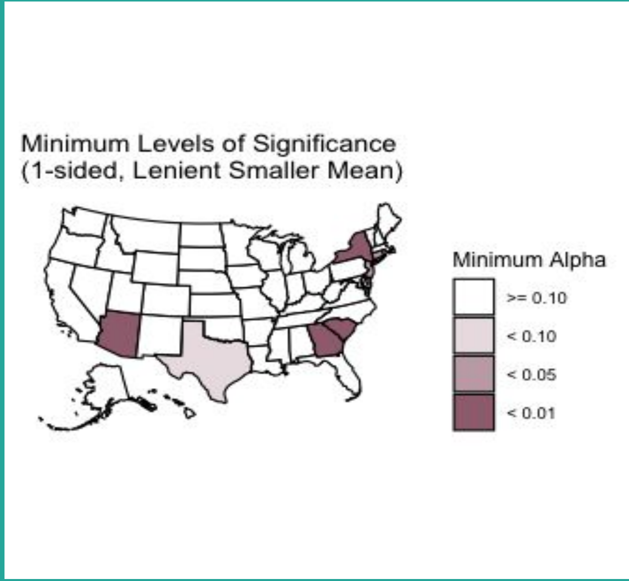


Average Daily Cases by State (2021)



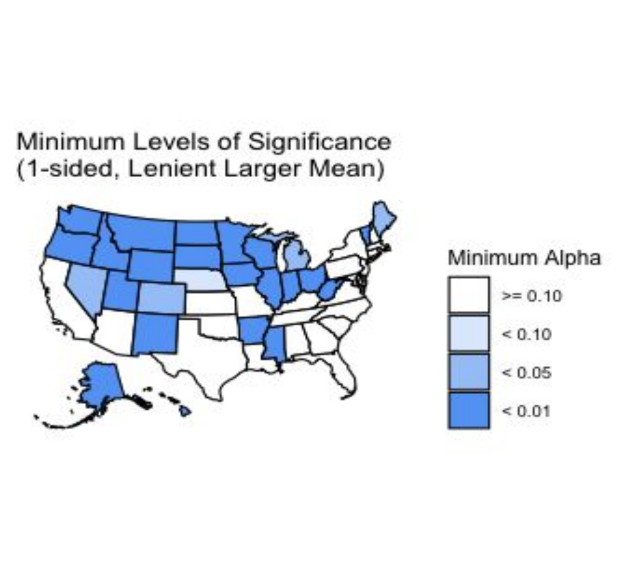
Average Daily Deaths by State (2021)

Statistical Significances of Lenient-Strict Mean Daily Cases Difference for each State (2021)



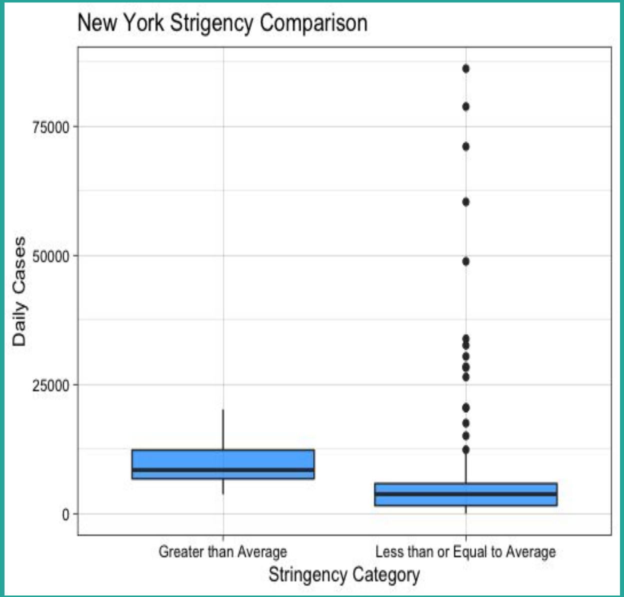
1-sided T-test Significance Levels (States with Significantly Smaller Mean Daily Cases During Periods of Lenient Regulations)

2-sided T-test Significance Levels (States with Significantly Different Mean Daily Cases Between Periods of Strict and Lenient Regulations)

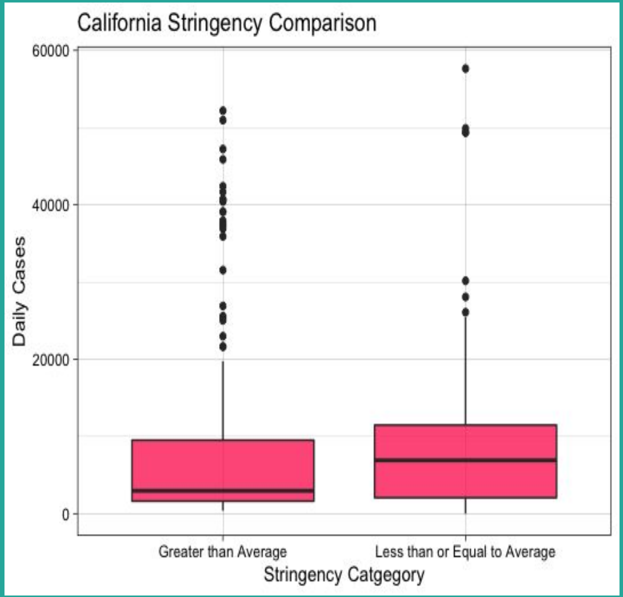


1-sided T-test Significance Levels (States with Significantly Larger Mean Daily Cases During Periods of Lenient Regulations)

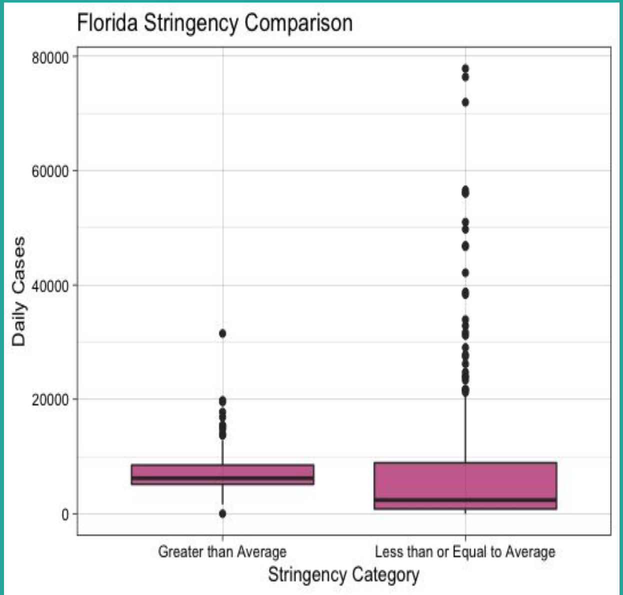
Box Plots of Daily Cases in Strict and Lenient Periods (2021)



New York State

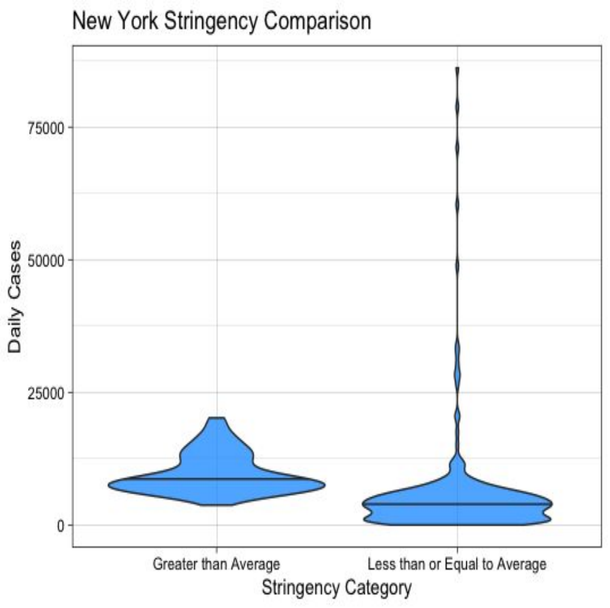


California

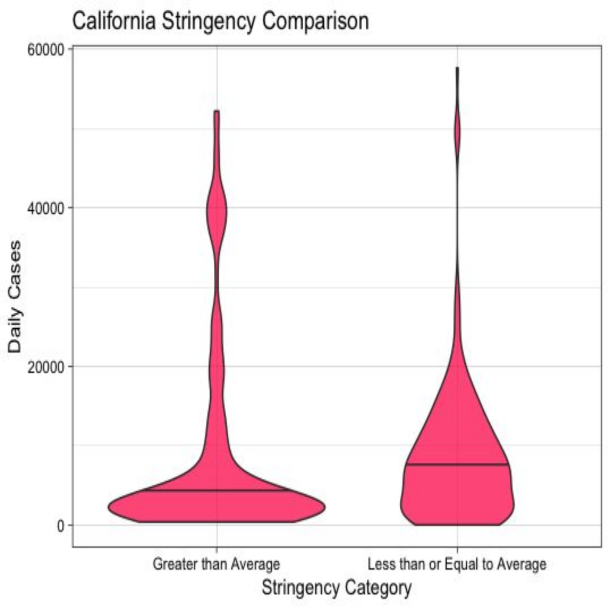


Florida

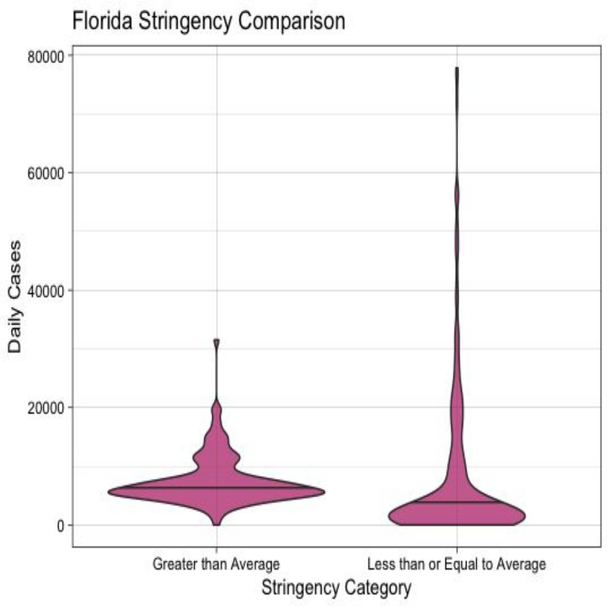
Violin Plots of Daily Cases in Strict and Lenient Periods (2021)



New York State

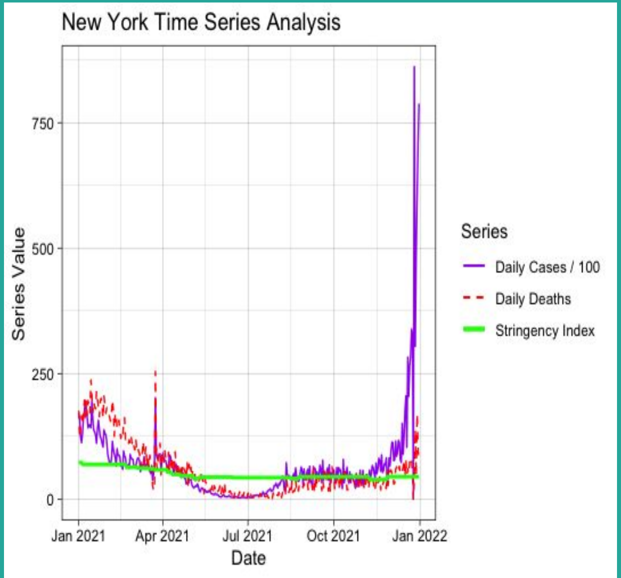


California

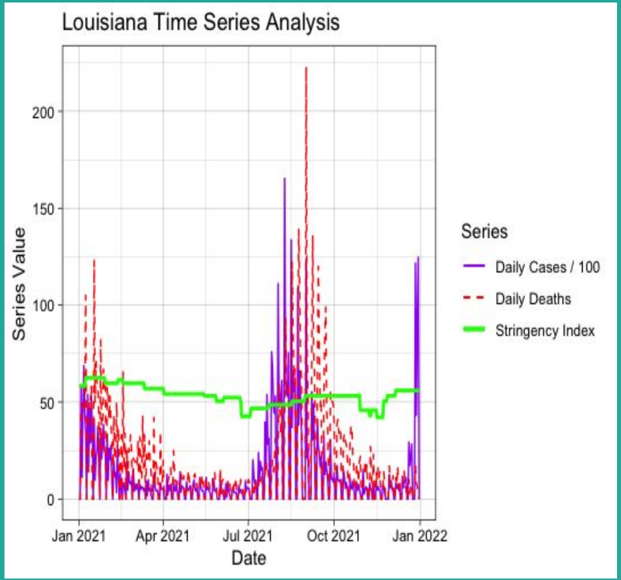


Florida

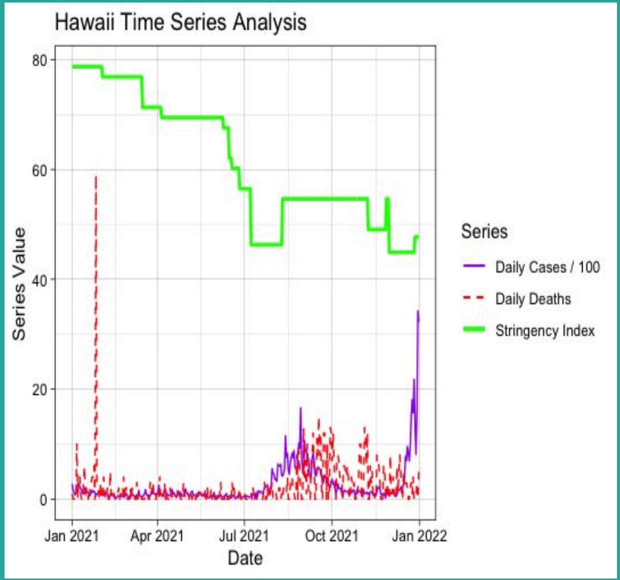
Time Series Plots of Daily Cases, Daily Deaths, and Stringency Index (2021)



New York State



Louisiana



Hawaii

Statistical Significance of Lenient-Strict Mean Daily Cases Difference: Summary and Conclusion

1-sided T-test (Lenient Smaller Mean) Summary:

Significance at...

- alpha = 0.01: 8% of states
- alpha = 0.05: 12% of states
- alpha = 0.10: 14% of states

2-sided T-test Summary:

Significance at...

- alpha = 0.01: 50% of states
- alpha = 0.05: 60% of states
- alpha = 0.10: 64% of states

1-sided T-test (Lenient Larger Mean) Summary:

Significance at...

- alpha = 0.01: 44% of states
- alpha = 0.05: 52% of states
- alpha = 0.10: 54% of states

Conclusion

Based on the 150 different T-tests conducted (3 for every state), it appears that at least half of the United States had a statistically significant difference between the mean number of cases during periods of lenient regulations and the mean number of cases during periods of strict regulations, specifically when considering the year 2021 (2021-01-01 to 2021-12-31). Furthermore, it seems that a larger proportion of the US states (at least 44% of them) had lenient-period mean numbers of daily cases that were larger than their strict-period mean numbers of daily cases, statistically speaking. Therefore, it appears that stricter regulations did have the desired effect of reducing COVID-19 transmission in a sizeable fraction of US states during 2021.