Big Data Final Project Report

Group Name: Immortals

Project Participants

- Hongyu Zhai (`hz2162`)
- Yuhan Chen (`yc4184`)
- Sifan Chen (`sc7782`)

Project Description

Analysis of the relationship between medical resource distribution, policy stringency index and confirmed case(by country, by state).

In this project, we are trying the answer the following research questions:

- Which countries/regions are slow/fast to take actions?
- Are there any countries/regions that are ignoring the rising numbers?
- Are there any countries/regions being extra cautious?
- What patterns can we find when examining the data?
- Do countries/regions with low medical resources tend to take more stringent actions?
- Do countries/regions with high population density tend to take more stringent actions?
- Do countries/regions with higher percentages of older people tend to take more stringent actions?
- What about states? Can we find similar patterns in the state level?

Approach Overview

The project is roughly divided into three parts:

- Collecting data
    - policy information (government responses to COVID-19, stay at home order, mandatory quarantine for travelers, non-essential business closures, large gatherings ban, school closures, bar/restaurant limits and Primary Election Postponement) from various regions
    - medical resources (test number, hospital capacity, number of beds and ICU beds, population, population density and percentage of older people) of various regions
    - local virus outbreak curves, mortality rate, cure rate, bed utilization rate, etc.
- Cleaning and wrangling the data.
    - Remove the header/footnotes from the table.
    - Calculating policy stringency index for each state.
    - Every steps detailed in a Jupyter Notebook https://github.com/iamzhaihy/BD2020-Final-Project/blob/master/Data%20Processing.ipynb
- Using processed data to generate visualizations.Try answering the research questions list above after gaining some insights from the visualization/analysis.

List of datasets

Here we provide a simplified list of datasets we used. A more detailed list can be found in `/data/datasets_used.csv` in our github repo. We also recorded the specific version we used in our Jupyter Notebook.

- Country level
    - World Bank - Hospital Beds (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.BEDS.ZS
    - World Bank - Physicians (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.PHYS.ZS

    - World Bank - Nurses (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.NUMW.P3
    - World Bank - Percentage of Ages 65+: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS
    - JHU - Global Confirmed Cases: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
    - JHU - Global Recovered Cases: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv
    - JHU - Global Death Tolls: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv
    - Oxford - Government Responses to COVID-19: https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
- State level
    - The Covid Tracking Project - State Totals https://covidtracking.com/data
    - KFF - State Actions to Stop COVID-19 Spread: https://www.kff.org/health-costs/issue-brief/state-data-and-policy-actions-to-address-coronavirus/#stateleveldata
    - Wikipedia - State Regulations: https://en.wikipedia.org/wiki/U.S._state_and_local_government_response_to_the_2020_coronavirus_pandemic
- County level
    - KHN - Hospital by County: https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/

- KHN - ICU Beds by County: https://khn.org/wp-content/uploads/sites/2/2020/03/KHN-ICU-bed-county-analysis_2.zip

Data Cleaning and Integration

Processing Country Level Data

At the country level, we are creating two tables `country_indicators.csv` and `country_responses.csv`. For the first table, we will collect the following information for each country:

name	meaning	source
`country_name`	the name of the country/region	`world-bank-hospital-beds`
`country_code`	the ISO 3 country code	`world-bank-hospital-beds`
`hospital_beds_per_1000`	number of hospital beds per 1000 people	`world-bank-hospital-beds`
`physicians_per_1000`	number of physicians per 1000 people	`world-bank-physicians`
`nurses_per_1000`	number of nurses and midwives per 1000 people	`world-bank-nurses`
`percentage_65up`	population ages 65 and above (% of total population)	`world-bank-elderly-population`

For the second table `country_responses.csv`, we will collect the time series data, each containing the following attributes:

name	meaning	source
date	date of the data collected	oxford-government-responses
country_name	name of the country/region	oxford-government-responses
country_code	ISO 3 country code	oxford-government-responses
c1	closing of schools/universities	oxford-government-responses
c1_flag	whether c1 is general or targeted	oxford-government-responses
c2	closing of workplaces	oxford-government-responses
c2_flag	whether c2 is general or targeted	oxford-government-responses
c3	cancelling public events	oxford-government-responses
c3_flag	whether c3 is general or targeted	oxford-government-responses
c4	cut-off size for bans on private gathering	oxford-government-responses
c4_flag	whether c4 is general or targeted	oxford-government-responses
c5	closing public transport	oxford-government-responses
c5_flag	whether c5 is general or targeted	oxford-government-responses
c6	stay at home requirements	oxford-government-responses
c6_flag	whether c6 is general or targeted	oxford-government-responses
c7	restricting internal travel	oxford-government-responses
c7_flag	whether c7 is general or targeted	oxford-government-responses
c8	restricting international travel	oxford-government-responses
h1	public_info_campaigns	oxford-government-responses
h1_flag	whether h1 is general or targeted	oxford-government-responses
stringency_index	the sum of policy scores, measuring the strictness of the government policies	oxford-government-responses
confirmed	the number of confirmed cases	oxford-government-responses
recovered	the number of recovered cases	jhu-global-recovered
deaths	death toll	oxford-government-responses

This table uses data from Oxford COVID-19 Government Responses Tracker. The authors use a novel index to measure the stringency of government responses. A total of 18 indicators are used. Nine of them (the ones we chose) are used to compute the stringency index (a value to measure the strictness of government policies). A detailed explanation can be found here.

Some cleaning operation (details can be seen in `Data Processing.ipynb`):

- The JHU datasets does not have ISO 3 country codes, which means we need to use country/region name to do the join. It is possible that different datasets use different names for the same country/region. After inspecting the data, we performed some translations:
    - `"Slovakia"` to `"Slovak Republic"`
    - `"Korea, South"` to `"South Korea"`
    - `"Kyrgyzstan"` to `"Kyrgyz Republic"`
    - `"Congo (Kinshasa)"` to `"Democratic Republic of Congo"`

    - "US" to "United States"
    - "Czechia" to "Czech Republic"
- Standardized date format.
    - M/DD/YY to YYYYMMDD

Processing State Level Data

At the state level, we are creating a total of four tables: state_indicators.csv, state_responses_04-24.csv, state_reponses_05-04.csv and state_cases.csv.

For the first table, we will collect the following information for each state:

name	meaning	source
state_name	the name of the state	khn-icu-beds-by-county
state_code	the code of the state	khn-hospital-by-county
hospitals_per_1000	the number of hospitals per 1000 people	khn-hospital-by-county
icu_beds_per_1000	number of icu beds per 1000 people	khn-icu-beds-by-county
percentage_60up	population ages 60 and above (% of total population)	khn-icu-beds-by-county
population_density_km2	population density of state	population-density-state
state_population	population of state	khn-hospital-by-county

For the second table state_responses_04-24.csv and the third table state_reponses_05-04.csv, we will collect the following information for each state:

name	meaning	source
state_name	the name of the state	kff-state-actions
state_is_easing_social_distancing_measures	state is easing social distancing measures	kff-state-actions
stay_at_home_order	order scope	kff-state-actions
date_when_stay_at_home_ordered	date when stay at home ordered	wiki-state-regulations
mandatory_quarantine_for_travelers	mandatory quarantine for travelers	kff-state-actions
non-essential_business_closures	business closures	kff-state-actions
large_gatherings_ban	gatherings scope ban	kff-state-actions
school_closures	details about school closures	kff-state-actions
bar/restaurant_limits	detail about bar/resurant limits	kff-state-actions
primary_election_postponement	whether primary election is postponed	kff-state-actions
emergency_declaration	emergency declaration	kff-state-actions
date_of_state_emergency_declared	date of state emergency declared	wiki-state-regulations
waive_cost_sharing_for_COVID-19_treatment	waive cost sharing for COVID-19 treatment	kff-state-health-policy-actions
free_cost_vaccine_when_available	free cost vaccine when available	kff-state-health-policy-actions
state_requires_waiver_of_prior_authorization_requirements	state requires waiver of prior authorization requirements	kff-state-health-policy-actions
early_Prescription_Refills	early Prescription Refills	kff-state-health-policy-actions
premium_payment_grace period	premium payment grace period	kff-state-health-policy-actions
marketplace_special_enrollment_period(SEP)	marketplace special enrollment period (SEP)	kff-state-health-policy-actions
section_1135_waiver	if waiver is approved	kff-state-health-policy-actions
paid_sick_leave	paid sick leave	kff-state-health-policy-actions
daycares	daycares	wiki-state-regulations

Some cleaning operation (details can be seen in `Data Processing.ipynb`):

- There are similar columns between `wiki-state-regulations` and `kff-state-actions`:
  - `"Gatherings banned"` and `"Large Gatherings Ban"`
  - `"Out-of-state travel restrictions"` and `"Mandatory Quarantine for Travelers"`
  - `"Schools"` and `"School Closures"`
  - `"Bars & sit-down restaurants"` and `"Bar/Restaurant Limits"`
  - `"Non-essential retail"` and `"Non-Essential Business Closures"`

  So for similar columns we only choose columns from kff-state-actions.

- We need to use state names to do the join. It is possible that different datasets use different names for the same location. After inspecting the output, we need to perform some translations.
    - `"Washington, D.C."` to `"District of Columbia"`
    - `"Washington (state)"` to `"Washington"`
    - `"New York (state)"` to `"New York"`
    - `"Georgia (U.S. state)"` to `"Georgia"`

    The KFF dataset does not have the territory record, so the following territory records are removed.
    - Puerto Rico
    - Northern Mariana Islands
    - Guam
    - United States Virgin Islands
    - American Samoa

    We then rename the column names for joining:
    - change Wiki dataset column `State/territory` to `state_name`
    - change KFF datasets column `Location` to `state_name`

For the fourth table `state_cases.csv`, we will collect the time series data, each containing the following attributes:

name	meaning	source
`date`	date	`covid-tracking-states-daily`
`state_name`	state name	`covid-tracking-states-daily`
`totaltestresults`	the number of test results	`covid-tracking-states-daily`
`confirmed`	the number of confirmed cases	`covid-tracking-states-daily`
`recovered`	the number of recovered cases	`covid-tracking-states-daily`
`deaths`	death toll	`covid-tracking-states-daily`

Cleaning operation (details can be seen in `Data Processing.ipynb`):

After inspecting the output, we see that the KFF dataset does not have the territory record. Therefore, we remove the records for territories.

Data Analysis and Visualizations

Calculating State Level Policy Stringency index

Based on Oxford Variation in government responses to COVID-19, we came up with a similar method to calculate state government response stringency index.

First, classify the columns in files in the policy-data folder (stored `kff-state-actions` from 03-25 to 05-07), then set encoding instructions.

Indicator	Column name	Description	Coding Instruction
I1	i1_stay_at_home_requirements	Stay at Home Order	0 - No Action or Expired or Lifted 1 - Other 2 - High-risk Groups or Rolled Back to High Risk Groups 3 - Affected Counties 4 - Statewide
I2	i2_travel_controls	Mandatory Quarantine for Travelers	0 - No Action or Lifted 1 - Other 2 - From Certain States or Rolled Back to Certain States 3 - All Air Travelers 4 - All Travelers
I3	i3_non_essential_business_closures	Non-Essential Business Closures	0 - No Actions 1 - Open with Reduced Capacity or All Non-Essential Businesses Permitted to Reopen with Reduced Capacity 2 - Some Non-Essential Businesses Permitted to Reopen 3 - Some Non-Essential Businesses Permitted to Reopen with Reduced Capacity 4 - Some Non-Essential Businesses Closed or Certain Non-Essential Businesses or Certain Non-Essential Businesses or All Non-Essential Retail Businesses 5 - All Non-Essential Businesses or All Non-Essential Businesses Closed
I4	i4_large_gathering_bans	Large Gatherings Ban	0 - No Actions or Lifted or Expired 1 - 50+ People Prohibited 2 - Other 3 - Expanded to 25+ People Prohibited or 25+ People Prohibited or Expanded to >25 People Prohibited 4 - 20+ People Prohibited or Expanded to 20+ People Prohibited 5 - 10 People Prohibited or Expanded to >10 People Prohibited or >10 People Prohibited 6 - All Gatherings Prohibited
I5	i5_school_closures	School Closures	0 - No action or Rescinded 1 - Other or Recommended Closure for School Year or Recommended Closure 2 -Yes or Effectively Closed or Closed or Closed for School Year
I6	i6_bar_restaurant_limits	Bar/Restaurant Limits	0 - No Action 1 - Limited on-site service or Restaurants Reopened to Dine-in Service or Other 2 - Closed Except for Takeout/Delivery or Original Restaurant Closures Still in Place
I7	i7_primary_elaction_postponement	Primary Election Postponement	0 - No Actions 1 - Yes or Postponed 2 - Canceled

All indicators `I1 - I7` will be used to calculate state Stringency Index, which will be a value between 0 and 100. Assuming all these indicators are state wide effective.

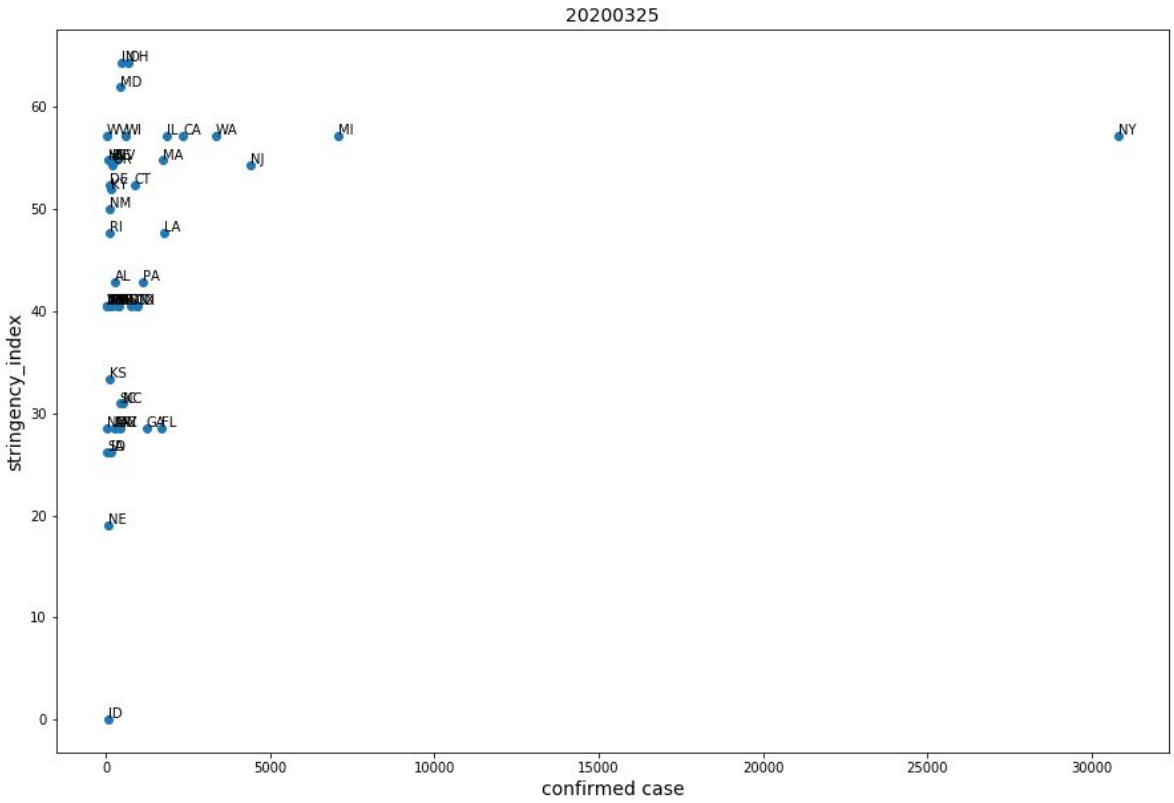
$$I = \frac{1}{7}\sum_{j=1}^{7} I_j$$

There are some indicators that have fewer ordinal points while other indicators have more. To avoid "over-contributing" and "under-contributing", we used a similar method as the researchers : calculating the weight of each indicator. N_j means the max ordinal points of jth indicators. C_j means the ordinal points of jth indicators.

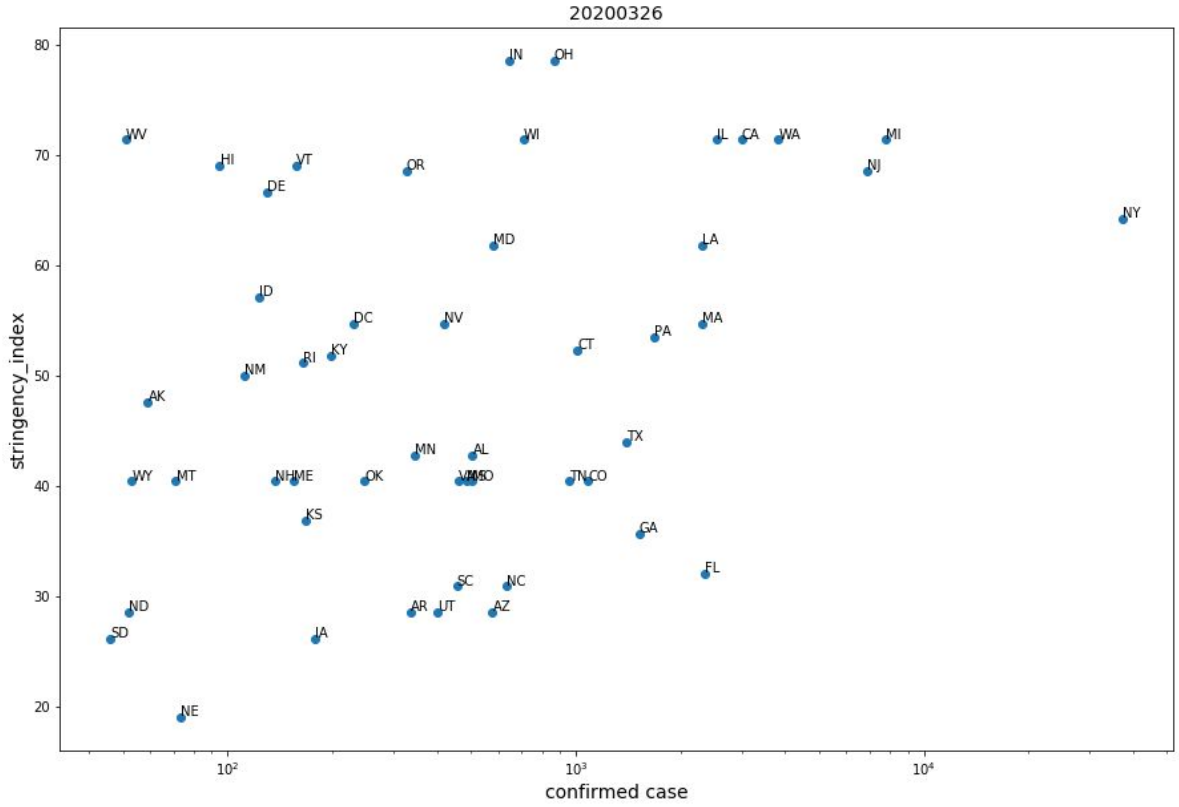
$$I_j = 100(\frac{C_j}{N_j})$$

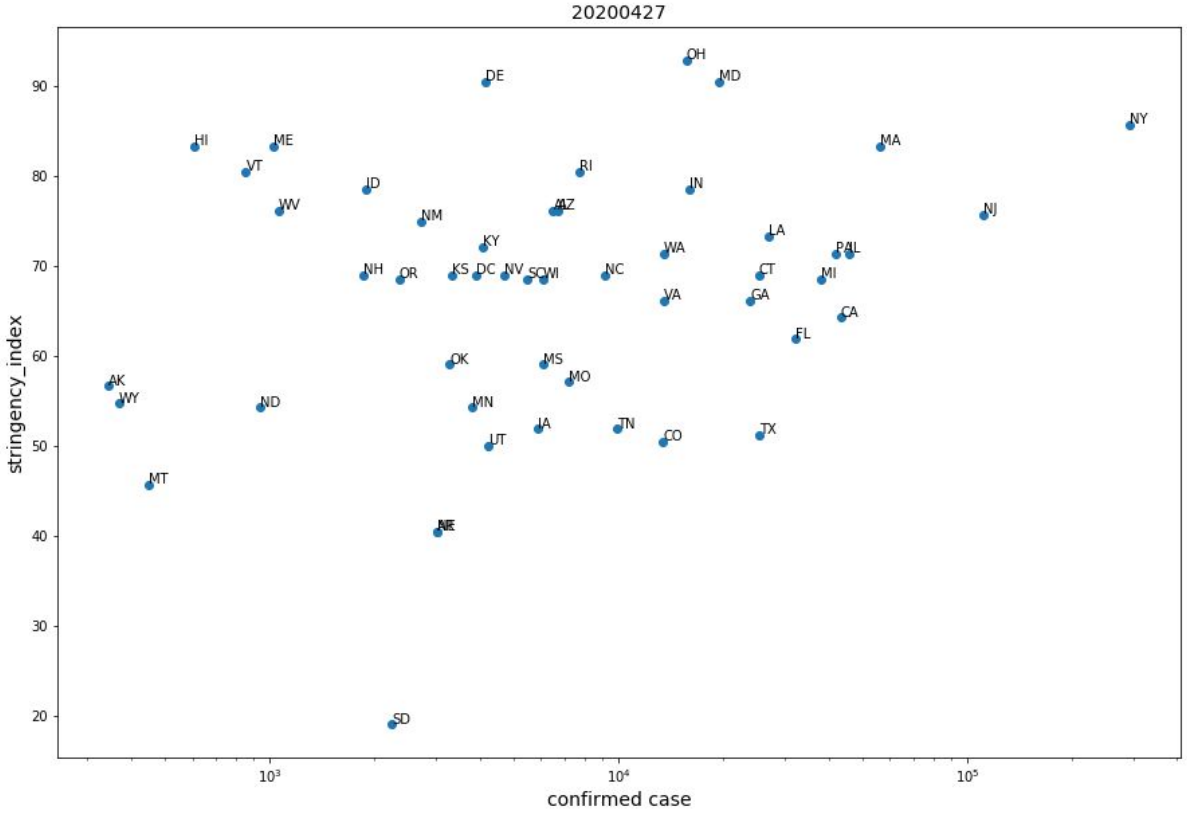
Visualization

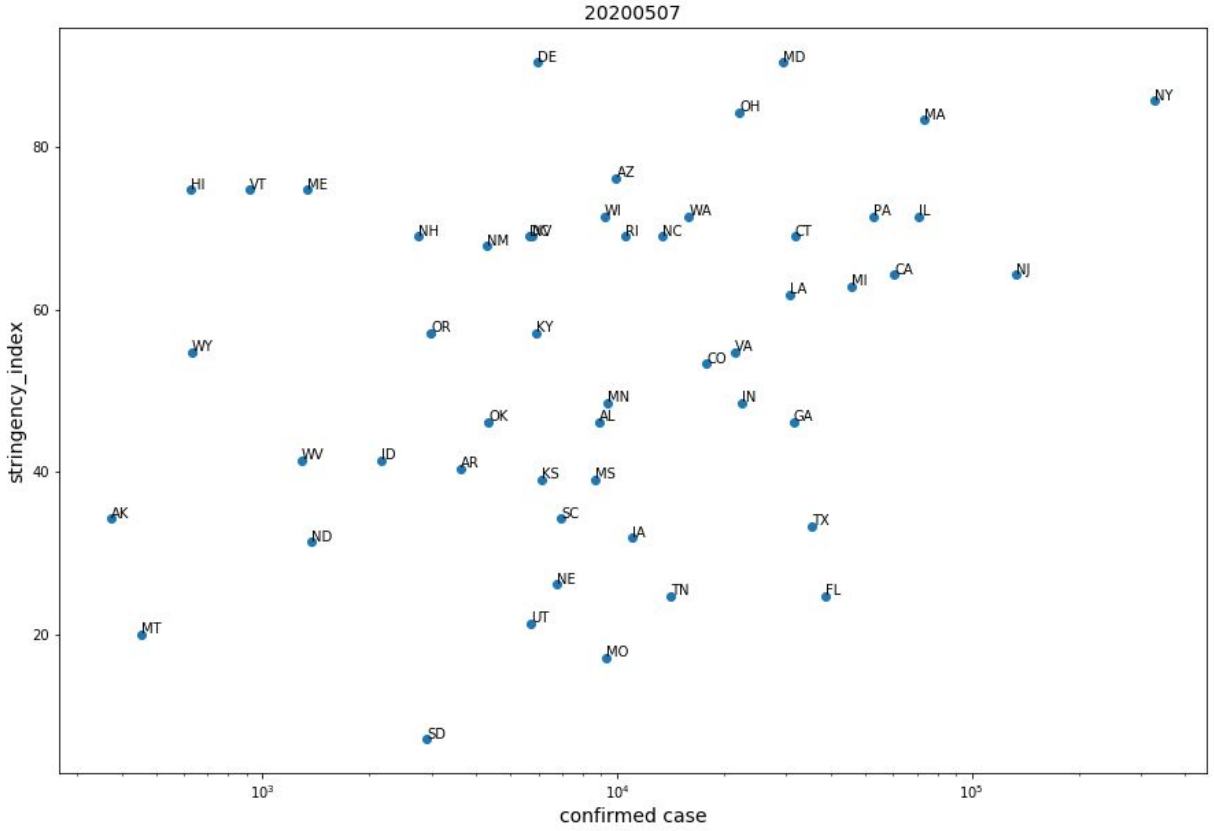
After calculating state level policy stringency index, we use it to merge with the confirmed case data in the `state_cases.csv` and generate confirmed_case- stringency index plot in time series to find the correlation. We save the visualization results in the plots folder.



As we can see in this plot, NY state has far more confirmed cases than other states. There are many dots clustered at the left side of the plot, so we cannot distinguish them well. Thus, we decided to use log scale on the x axis.


20200326
stringency_index
confirmed case
80
70
60
50
40
30
20
10^2
10^3
10^4
IN
OH
WV
WI
IL
CA
WA
MI
HI
VT
OR
NJ
DE
NY
MD
LA
ID
DC
NV
MA
PA
CT
RI
KY
NM
AK
TX
MN
AL
WY
MT
NH
ME
OK
TN
CO
KS
GA
FL
SC
NC
ND
AR
UT
AZ
SD
IA
NE

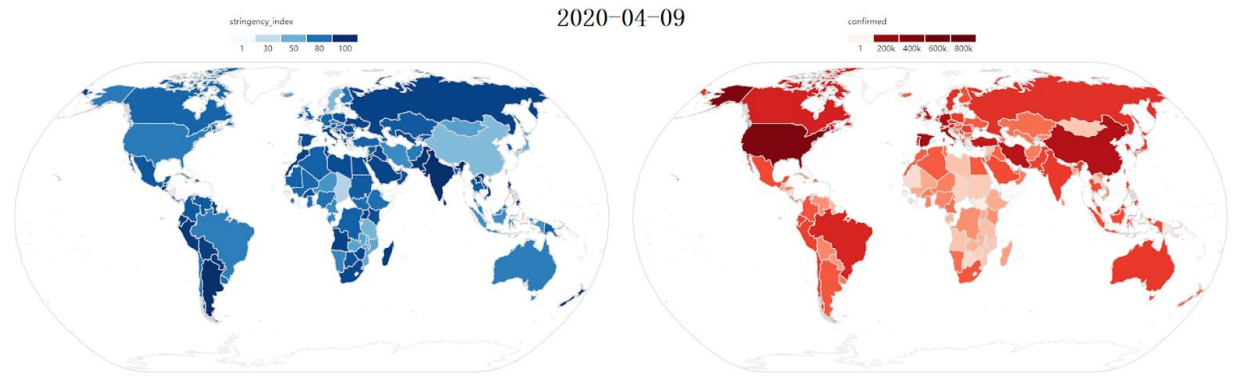

20200427
stringency_index
confirmed case
90
80
70
60
50
40
30
20
10^3
10^4
10^5
OH
DE
MD
NY
HI
ME
MA
VT
RI
ID
IN
WV
NM
NJ
LA
KY
WA
NH
OR
KS
DC
NV
NC
CT
MI
VA
GA
CA
FL
OK
MS
AK
MO
WY
ND
MN
IA
TN
TX
UT
CO
MT
SD



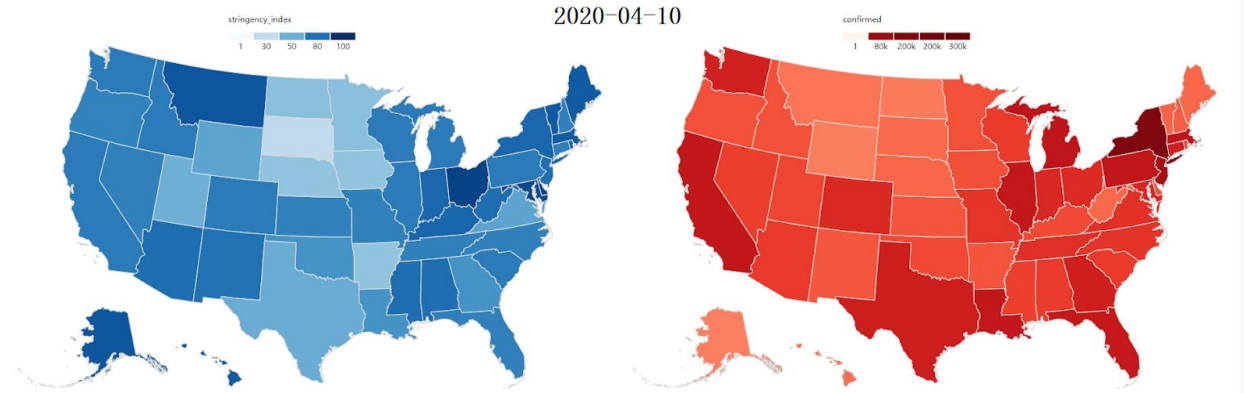
According to the plots above (there are twenty plots in all, we only showed three of them here) we can see how the number of confirmed cases and state stringency_index change over time.

To explore the data interactively, we used D3.js to generate interactive maps of stringency index and confirmed cases on both country level and state level. Such an interactive visualization enabled us to better explore the data. The user can change the date by dragging the range slider on top. It will show extra information if the user hovers the cursor on the country/state.

Visualization 1



Visualization 2



Challenges and Limitations

1. It is challenging to find data sources related to medical resources and government policy at the country level and state level. It is especially hard to collect state level policy data, since tracing daily policy changes is a nontrivial task. The sources for state actions are diverse and chaotic, and we also need to compute stringency_index for each state by ourselves.

Solution:

We looked for data from many sources, such as Google dataset, Github, and Kaggle. To increase data credibility, we also traced the origin of the datasets that we found. For state level policy data, we utilize the snapshots on `archive.org`, manually adjusted using information on Wikipedia, and came up with our own encoding rules to compute stringency_index for states.

2.The project spans several weeks, so the datasets we use might also change. For example, datasets of state actions responses to COVID-19 are constantly changing and the statistical methods of data sets are also changing.

Solution:
We focused on the latest and most accurate data source, updated our dataset regularly, and made necessary changes to our processed data.

3.When doing join operation, there exist columns that have the same meaning, but with totally different names.

Solution:
We examined the datasets carefully and took extra care when we decided to modify. We removed columns that are not related to the problem, compared column names between different datasets, and standardized column names before doing join operation.

4.When it comes to government policy, it is usually very complex and subtle. Government responses may vary from different regions. Notably, some strategies could only adapt to particular countries or states. In this case, focusing on stringency index analysis may not reflect the whole picture of responses and lose the detailed information when making decisions. Besides, policy effectiveness and adherence of the public is also hard to be measured and taken into consideration.

Solution:

We combined the state stringency index with the number of cases to generate visualization results, analyzed the correlation between them, and tried to explain the potential reasons. We made educated guesses based on the best data we can find.

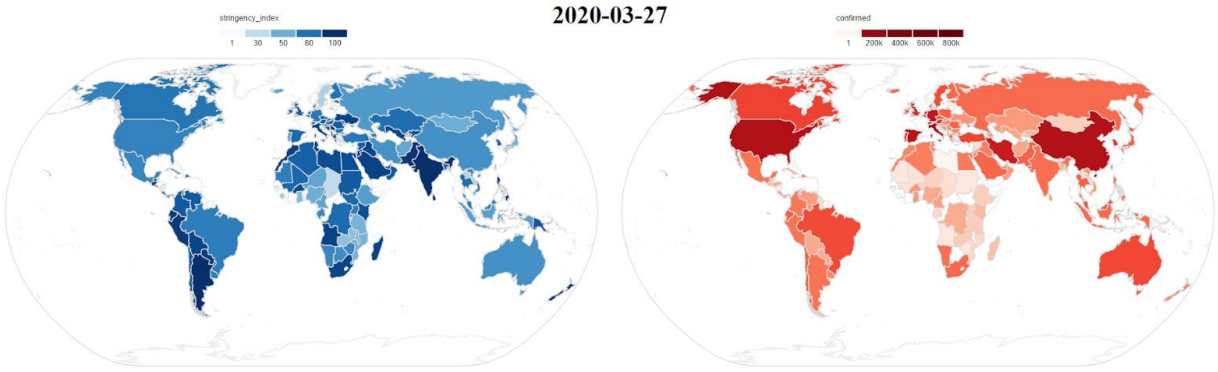
5. As for the maps, the shape and size of a region may influence the perception of color.

Solution:
Each map projection has pros and cons. We are aware of the potential problem and chose a projection that does not distort the map too much.

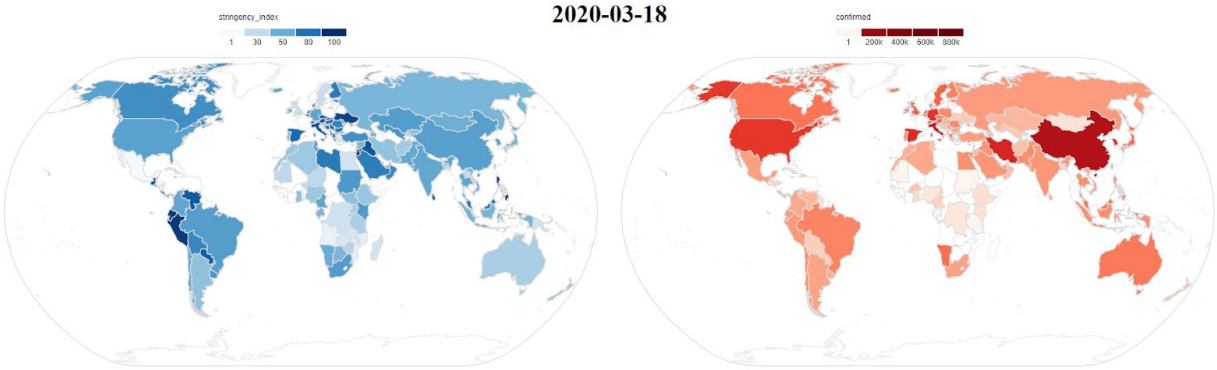
Findings and Answers to Research Questions

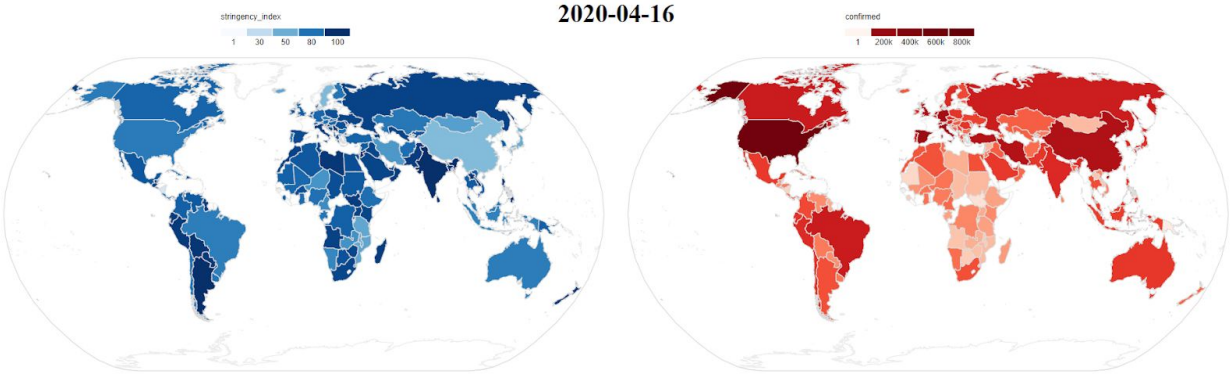
Country level analysis

Figure1: Choropleth maps of confirmed cases and stringency index in country level

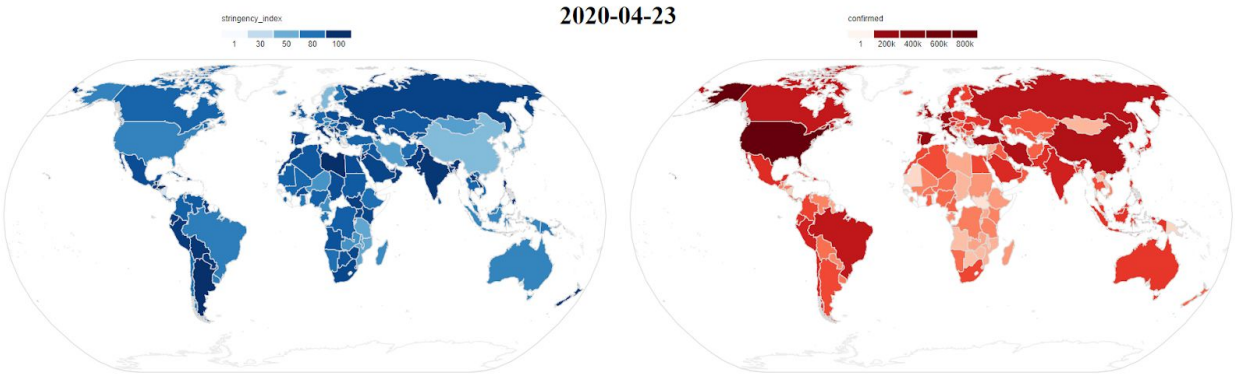


(2020-03-27) Most African countries adopted strict policies at the beginning of the pandemic.





Iran (2020-03-18) and Brazil (2020-04-16) responded relatively slower than most countries.

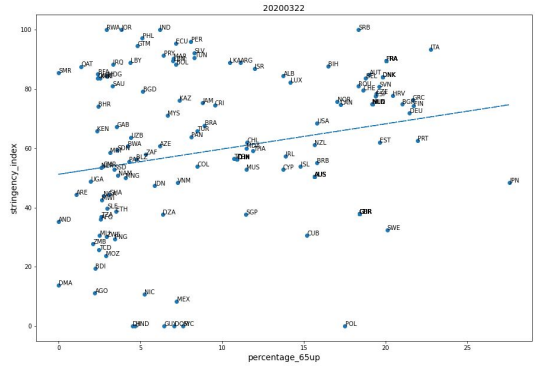


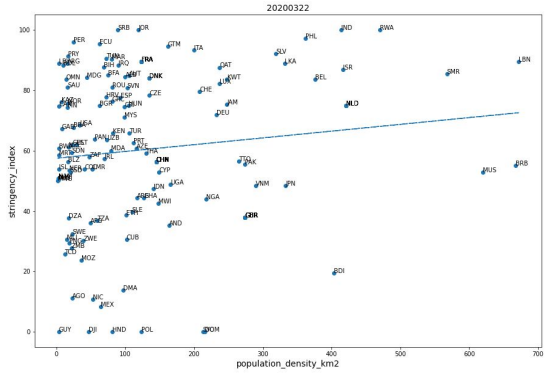
(2020-04-23) Compared Japan to Russia, the number of confirmed cases is close but Japan is taking measures much more lighter than Russia. Probably because of the 2020 Tokyo Olympic Games.

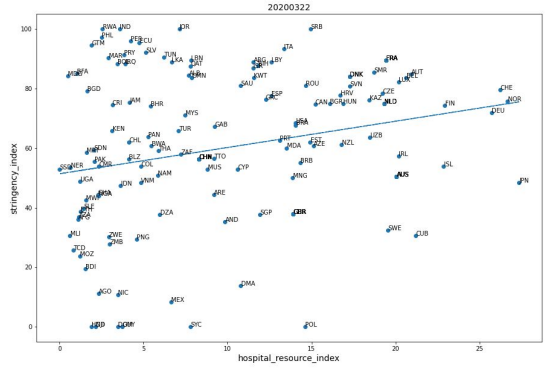
(2020-04-23) Compared the USA to Canada and Mexico, the situation in the USA is much more serious. However, the USA is taking less stringent measures than Canada and Mexico.

Based on the visualization, it seems that the earlier the measures taken, the slower the confirmed cases increase. We can see that the number of confirmed cases in China and Mongolia are now relatively small compared to many other countries. It is therefore reasonable to infer that fast stringent actions are helpful to control the spread of COVID-19.

Figure2: Relations between hospital resources, population density, percentage of the elder and stringency index in country level:





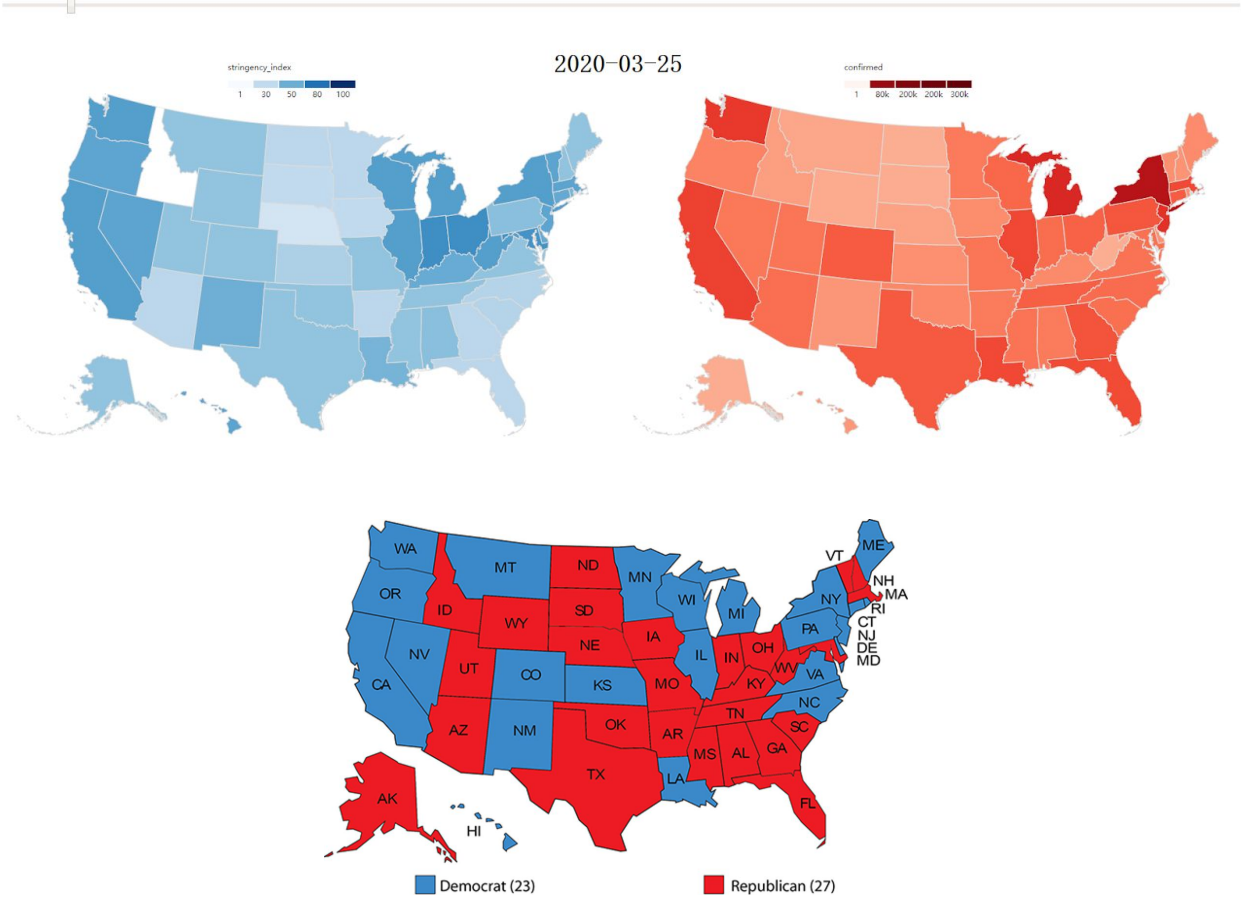


To explore how different factors (population density, hospital resources, percentage of 65+ years old people, etc.) influence the policy stringency, we draw three charts of the relationship between the policy stringency index and these features respectively. As we can see, the positive slopes of the regression lines for three figures indicate that the strength of these features is positively

related to the policy stringency. It suggests that when making policy, the governments did consider their own status rather than blindly follow the early successful template. It is not wise and unfair for some countries to blame others for not taking the same policy as they did.

State level analysis

Figure1: Choropleth maps of confirmed cases and stringency index in state level:

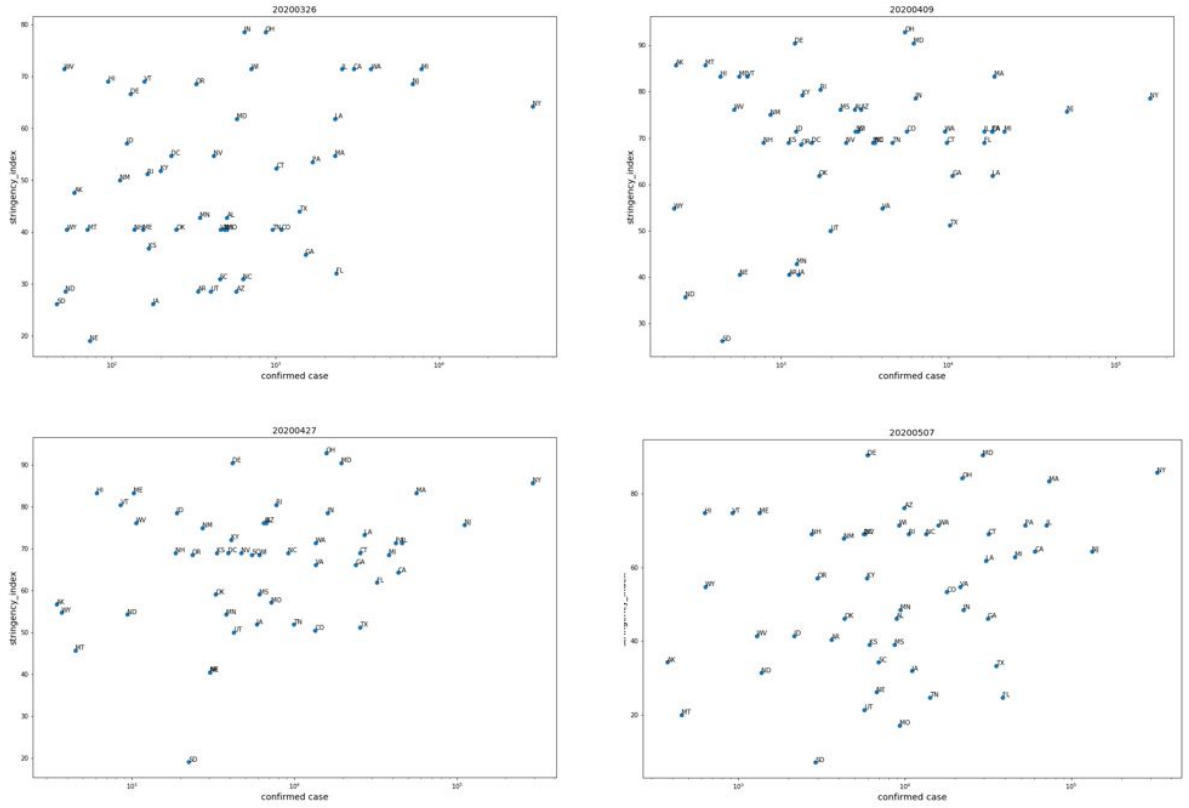


(source:https://en.wikipedia.org/wiki/State_attorney_general)

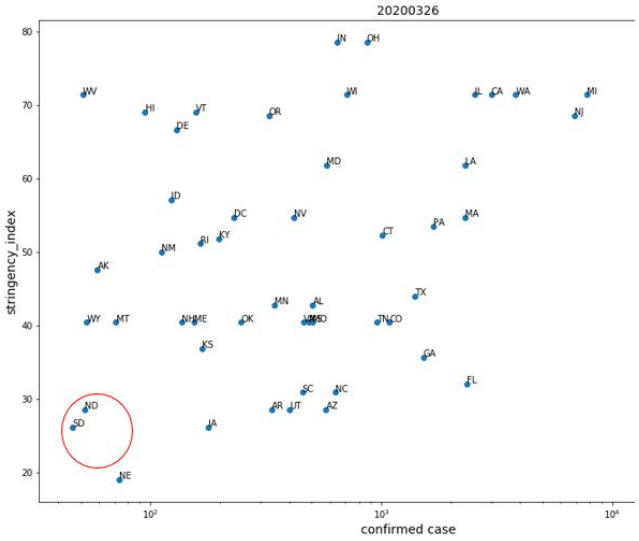
- The stringency map looks similar to the one below: blue states are generally taking more stringent actions than red states. Since the number of confirmed cases is related to the number of tests, it might be because the Republican states have fewer tests and are not actively testing. One interesting exception is Ohio, which is taking fast and strict actions.

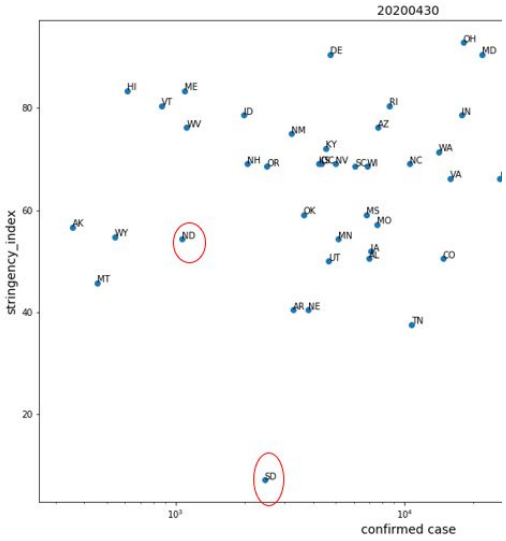
- Since Apr 30th, we see something similar. The country is reopening and red states are generally acting faster to loosen their policies. For example: (2020-05-07) Texas and California : the number of comfirmed cases are almost the same. However, their stringency index is very different. It might be, again, due to the difference of two political parties.

Figure2 Relations between confirm cases and stringency index in state level:

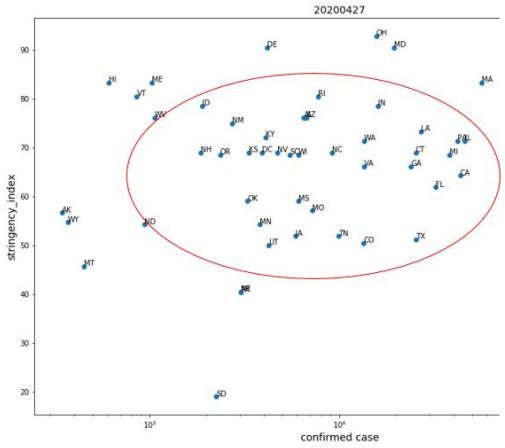


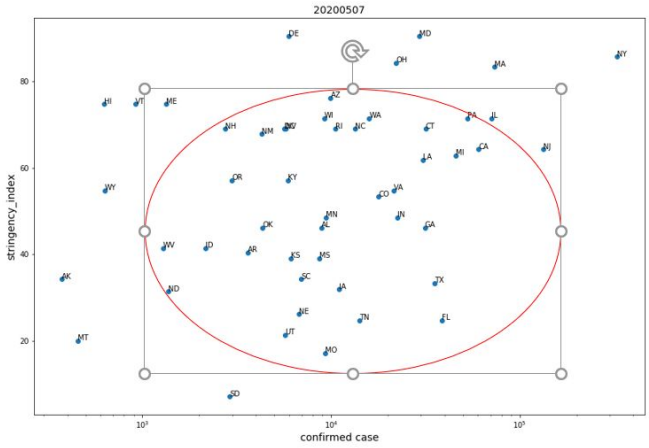
- States with a high number of confirmed cases (such as NY, NJ, and MA) remain at a relatively high level of stringency index.
- For states with a moderate number of confirmed cases, the stringency indices are more unstable.



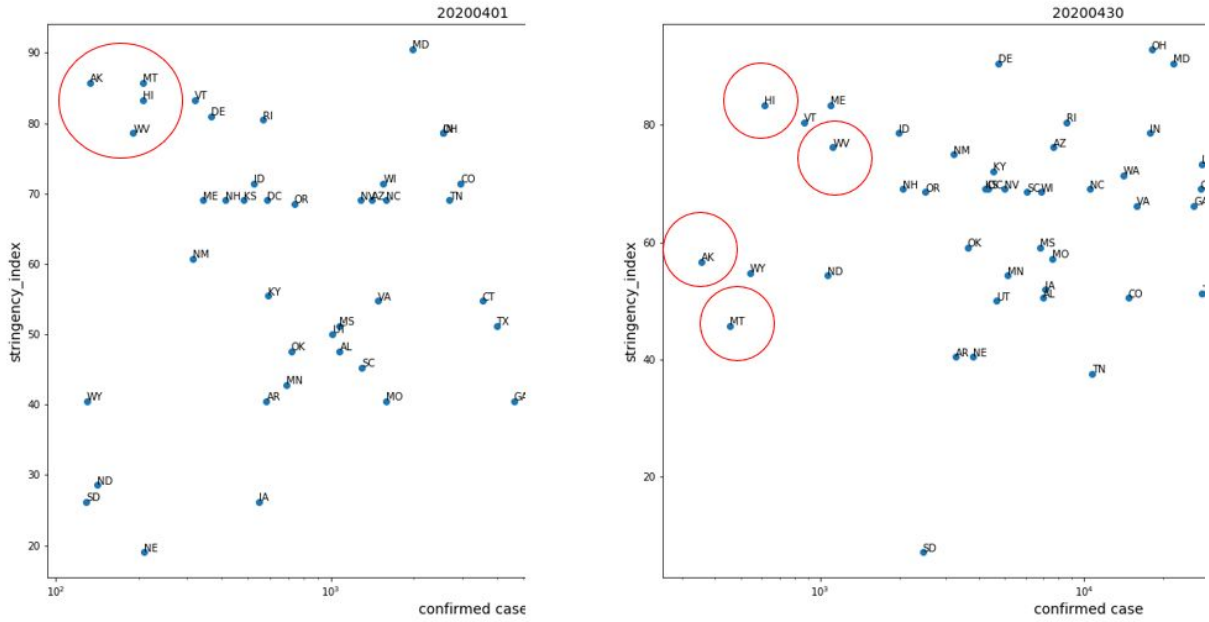


- Compared SD to ND, both of them have a relatively low number of confirmed cases in the United States on Mar 25th. As confirmed cases increase, ND takes more strict actions and stringency index increases accordingly, while the low stringency index of SD remains low. The number of confirmed cases in this SD is increasing faster than ND during April.



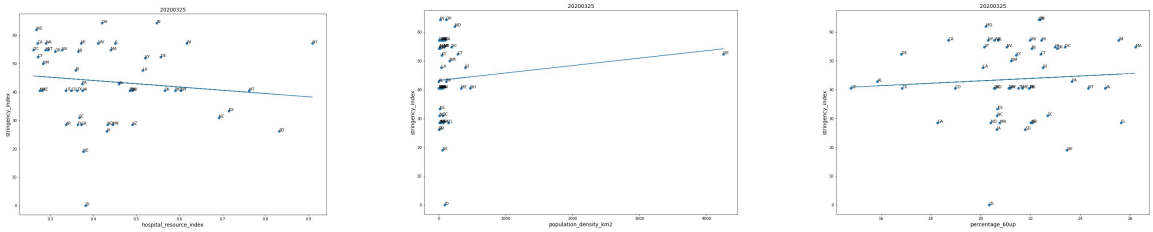


- In April, most states adopted more strict actions as the number of confirmed cases increased. The stringency indices increased accordingly.
- Many states loosen the policies in May, and stringency indices drop.



- AK, WV, HI and MT are extra cautious. These states take strict actions at the beginning of April when they still have relatively low confirmed cases. And they make great achievements in slowing down the spread speed of COVID-19.

Figure3: Relations between hospital resources, population density, percentage of the elder and stringency index in state level:



We take the same analysis on the state's data. However, there is no obvious relationship between these features. The result seems to suggest that the difference between two political parties is the main reason that states are acting differently.