DATA605 Project

Predict National Policy Changes based on Case Data

Group Members:
Chao Cao
David Haft
Miao Zhang
Yibo Zhang

Introduction

COVID-19 is an ongoing, multi-year event. Public health policies change over time, as do infection rates.

This leads to the questions:

- Is there any way we can visualize the case data and policy data at the same time?
- Can we predict policy changes based on the case data?

Data Source (1)

Policy data

Oxford Covid-19 Government Response Tracker (OxCGRT)

The Oxford Covid-19 Government Response Tracker (OxCGRT) collects systematic information on which governments have taken which measures, and when. This can help decision-makers and citizens understand governmental responses in a consistent way, aiding efforts to fight the pandemic. The OxCGRT systematically collects information on several different common policy responses governments have taken, records these policies on a scale to reflect the extent of government action, and aggregates these scores into a suite of policy indices.

This is a project from the Blavatnik School of Government. More information on the OxCGRT is available on the school's website: https://www.bsg.ox.ac.uk/covidtracker. This README contains information about using the database.

https://github.com/OxCGRT/covid-policy-tracker

Policy data

```
Index(['CountryName', 'CountryCode', 'RegionName', 'RegionCode',
       'Jurisdiction', 'Date', 'C1_School closing', 'C1_Flag', 'C1_Notes',
       'C2_Workplace closing', 'C2_Flag', 'C2_Notes',
       'C3_Cancel public events', 'C3_Flag', 'C3_Notes',
       'C4_Restrictions on gatherings', 'C4_Flag', 'C4_Notes',
       'C5_Close public transport', 'C5_Flag', 'C5_Notes',
       'C6_Stay at home requirements', 'C6_Flag', 'C6_Notes',
       'C7_Restrictions on internal movement', 'C7_Flag', 'C7_Notes',
       'C8_International travel controls', 'C8_Notes', 'E1_Income support',
       'E1_Flag', 'E1_Notes', 'E2_Debt/contract relief', 'E2_Notes',
       'E3_Fiscal measures', 'E3_Notes', 'E4_International support',
       'E4_Notes', 'H1_Public information campaigns', 'H1_Flag', 'H1_Notes',
       'H2_Testing policy', 'H2_Notes', 'H3_Contact tracing', 'H3_Notes',
       'H4_Emergency investment in healthcare', 'H4_Notes',
       'H5_Investment in vaccines', 'H5_Notes', 'H6_Facial Coverings',
       'H6_Flag', 'H6_Notes', 'H7_Vaccination policy', 'H7_Flag', 'H7_Notes',
       'H8_Protection of elderly people', 'H8_Flag', 'H8_Notes', 'M1_Wildcard',
       'M1_Notes', 'V1_Vaccine Prioritisation (summary)', 'V1_Notes',
       'V2A_Vaccine Availability (summary)', 'V2_Notes',
       'V2B_Vaccine age eligibility/availability age floor (general population summary)',
       'V2C_Vaccine age eligibility/availability age floor (at risk summary)',
       'V2D_Medically/ clinically vulnerable (Non-elderly)', 'V2E_Education',
       'V2F_Frontline workers  (non healthcare)',
       'V2G_Frontline workers  (healthcare)',
       'V3_Vaccine Financial Support (summary)', 'V3_Notes',
       'V4_Mandatory Vaccination (summary)', 'V4_Notes', 'ConfirmedCases',
       'ConfirmedDeaths', 'StringencyIndex', 'StringencyIndexForDisplay',
       'StringencyLegacyIndex', 'StringencyLegacyIndexForDisplay',
       'GovernmentResponseIndex', 'GovernmentResponseIndexForDisplay',
       'ContainmentHealthIndex', 'ContainmentHealthIndexForDisplay',
       'EconomicSupportIndex', 'EconomicSupportIndexForDisplay'],
      dtype='object')
```

Policy data

```
dfP
```

	Unnamed: 0	CountryName	CountryCode	RegionName	RegionCode	Jurisdiction	Date	C1_School closing	C1_Flag	C1_Notes	...	StringencyIndex	StringencyI
0	30784	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-01	2.0	0.0	NaN	...	53.24	
1	30785	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-02	2.0	0.0	NaN	...	53.24	
2	30786	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-03	2.0	0.0	NaN	...	53.24	
3	30787	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-04	2.0	0.0	Several school districts across the country ar...	...	53.24	
4	30788	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-05	2.0	0.0	NaN	...	53.24	
...	...	...	...	...	...	...	...	...	...	...	...	...	
5403	36187	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-04-10	NaN	NaN	NaN	...	NaN	
5404	36188	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-04-11	NaN	NaN	NaN	...	NaN	
5405	36189	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-04-12	NaN	NaN	NaN	...	NaN	
5406	36190	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-04-13	NaN	NaN	NaN	...	NaN	
5407	36191	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-04-14	NaN	NaN	NaN	...	NaN	

5408 rows × 87 columns

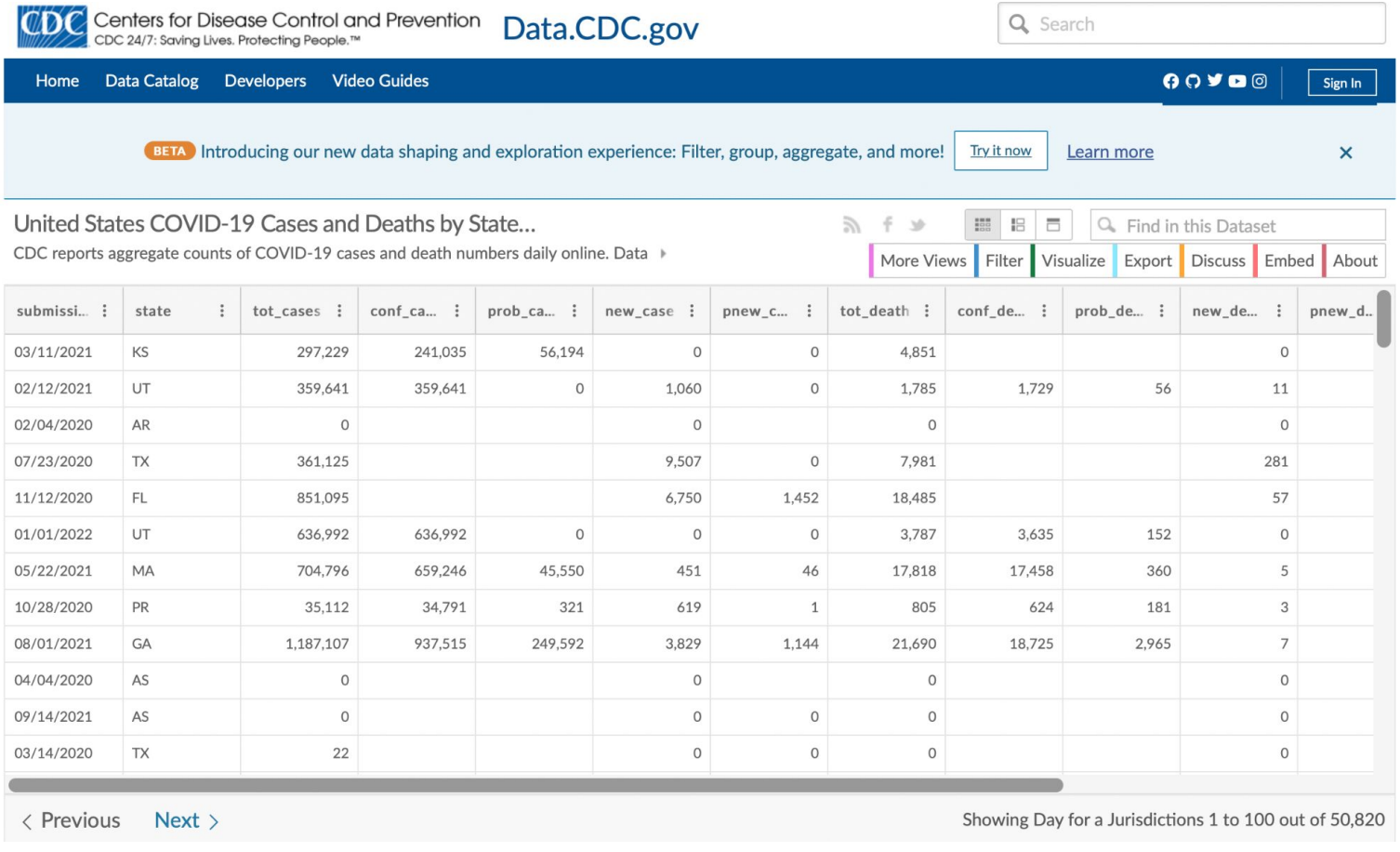
Data Source (2)

Case data

United States COVID-19 Cases and Deaths by State

- CDC reports aggregate counts of COVID-19 cases and death numbers daily online.
- However, because many municipalities do batch-reporting, in order to keep the data smooth, we use a 10-day window, not the daily numbers

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data


Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™
Data.CDC.gov
Search
Home
Data Catalog
Developers
Video Guides
Sign In
BETA
Introducing our new data shaping and exploration experience: Filter, group, aggregate, and more!
Try it now
Learn more
United States COVID-19 Cases and Deaths by State...
CDC reports aggregate counts of COVID-19 cases and death numbers daily online. Data
Find in this Dataset
More Views
Filter
Visualize
Export
Discuss
Embed
About
submissi...
state
tot_cases
conf_ca...
prob_ca...
new_case
pnew_c...
tot_death
conf_de...
prob_de...
new_de...
pnew_d..
03/11/2021 KS 297,229 241,035 56,194 0 0 4,851 0
02/12/2021 UT 359,641 359,641 0 1,060 0 1,785 1,729 56 11
02/04/2020 AR 0 0 0 0
07/23/2020 TX 361,125 9,507 0 7,981 281
11/12/2020 FL 851,095 6,750 1,452 18,485 57
01/01/2022 UT 636,992 636,992 0 0 0 3,787 3,635 152 0
05/22/2021 MA 704,796 659,246 45,550 451 46 17,818 17,458 360 5
10/28/2020 PR 35,112 34,791 321 619 1 805 624 181 3
08/01/2021 GA 1,187,107 937,515 249,592 3,829 1,144 21,690 18,725 2,965 7
04/04/2020 AS 0 0 0 0
09/14/2021 AS 0 0 0 0 0
03/14/2020 TX 22 0 0 0
Previous
Next
Showing Day for a Jurisdictions 1 to 100 out of 50,820

In [14]:
```
df = pd.read_csv('../data/db.csv')
df.head()
```

	submission_date	state	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_death	conf_death	prob_death	new_d
0	12/01/2021	ND	163565	135705.0	27860.0	589	220.0	1907	NaN	NaN	9
1	09/01/2021	ND	118491	107475.0	11016.0	536	66.0	1562	NaN	NaN	1
2	05/12/2022	CT	777064	696528.0	80536.0	1963	173.0	10883	8906.0	1977.0	0
3	10/04/2020	MD	127290	NaN	NaN	471	0.0	4092	3933.0	159.0	3
4	03/11/2021	MD	390490	NaN	NaN	924	0.0	8549	8345.0	204.0	19

In [15]:
```
df.columns
```

```
Index(['submission_date', 'state', 'tot_cases', 'conf_cases', 'prob_cases',
       'new_case', 'pnew_case', 'tot_death', 'conf_death', 'prob_death',
       'new_death', 'pnew_death', 'created_at', 'consent_cases',
       'consent_deaths'],
      dtype='object')
```

4Vs

Volume	Variety
The size of all datasets is ~500MB	pandas dataframe ⟶ spark dataframe
Velocity	**Veracity**
Real-time response (daily CDC data, results can be re-generated to use new information)	Intentionally sought out highly rated and independently verified sources

Step1: Exploring Policy Data

1. Separate national policy and state policy

In [3]:
```
# National Policies
dfP[dfP['Jurisdiction']=='NAT_TOTAL']
```

	Unnamed: 0	CountryName	CountryCode	RegionName	RegionCode	Jurisdiction	Date	C1_School closing	C1_Flag	C1_Notes	...	S
0	41144	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-01	2.0	0.0	NaN	...	5
1	41145	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-02	2.0	0.0	NaN	...	5
2	41146	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-03	2.0	0.0	NaN	...	5
3	41147	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-04	2.0	0.0	Several school districts across the country ar...	...	5
4	41148	United States	USA	NaN	NaN	NAT_TOTAL	2022-01-05	2.0	0.0	NaN	...	5
...	...	...	...	...	...	...	...	...	...	...	...	.
134	41278	United States	USA	NaN	NaN	NAT_TOTAL	2022-05-15	NaN	NaN	NaN	...	N
135	41279	United States	USA	NaN	NaN	NAT_TOTAL	2022-05-16	NaN	NaN	NaN	...	N
136	41280	United States	USA	NaN	NaN	NAT_TOTAL	2022-05-17	NaN	NaN	NaN	...	N
137	41281	United States	USA	NaN	NaN	NAT_TOTAL	2022-05-18	NaN	NaN	NaN	...	N
138	41282	United States	USA	NaN	NaN	NAT_TOTAL	2022-05-19	NaN	NaN	NaN	...	N

139 rows × 87 columns

In [4]:
```
# State Policies
dfP[dfP['Jurisdiction']=='STATE_TOTAL']
```

	Unnamed: 0	CountryName	CountryCode	RegionName	RegionCode	Jurisdiction	Date	C1_School closing	C1_Flag	C1_Notes	...
139	41283	United States	USA	Alaska	US_AK	STATE_TOTAL	2022-01-01	1.0	0.0	NaN	...
140	41284	United States	USA	Alaska	US_AK	STATE_TOTAL	2022-01-02	1.0	0.0	NaN	...
141	41285	United States	USA	Alaska	US_AK	STATE_TOTAL	2022-01-03	1.0	0.0	NaN	...
142	41286	United States	USA	Alaska	US_AK	STATE_TOTAL	2022-01-04	1.0	0.0	NaN	...
143	41287	United States	USA	Alaska	US_AK	STATE_TOTAL	2022-01-05	1.0	0.0	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
7223	48367	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-05-15	0.0	NaN	NaN	...
7224	48368	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-05-16	0.0	NaN	NaN	...
7225	48369	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-05-17	0.0	NaN	NaN	...
7226	48370	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-05-18	NaN	NaN	NaN	...
7227	48371	United States	USA	Wyoming	US_WY	STATE_TOTAL	2022-05-19	NaN	NaN	NaN	...

7089 rows × 87 columns

Step1: Exploring Policy Data

2. focus on instereted policy Sort by date

	Date	C1_School closing	C2_Workplace closing	C3_Cancel public events	C4_Restrictions on gatherings	C5_Close public transport	H2_Testing policy
0	2022-01-01	2.0	2.0	1.0	4.0	1.0	3.0
1	2022-01-02	2.0	2.0	1.0	4.0	1.0	3.0
2	2022-01-03	2.0	2.0	1.0	4.0	1.0	3.0
3	2022-01-04	2.0	2.0	1.0	4.0	1.0	3.0
4	2022-01-05	2.0	2.0	1.0	4.0	1.0	3.0
...	...	...	...	...	...	...	...
134	2022-05-15	NaN	NaN	NaN	NaN	NaN	NaN
135	2022-05-16	NaN	NaN	NaN	NaN	NaN	NaN
136	2022-05-17	NaN	NaN	NaN	NaN	NaN	NaN
137	2022-05-18	NaN	NaN	NaN	NaN	NaN	NaN
138	2022-05-19	NaN	NaN	NaN	NaN	NaN	NaN

139 rows × 7 columns

Step1: Exploring Policy Data

3. Calculate the difference to know when the policies change

```
In [9]: dfNP.loc[:,'C1_School closing'].rolling(window=2).apply(lambda x: x.iloc[1] - x.iloc[0]).unique()

array([nan,  0.])
```

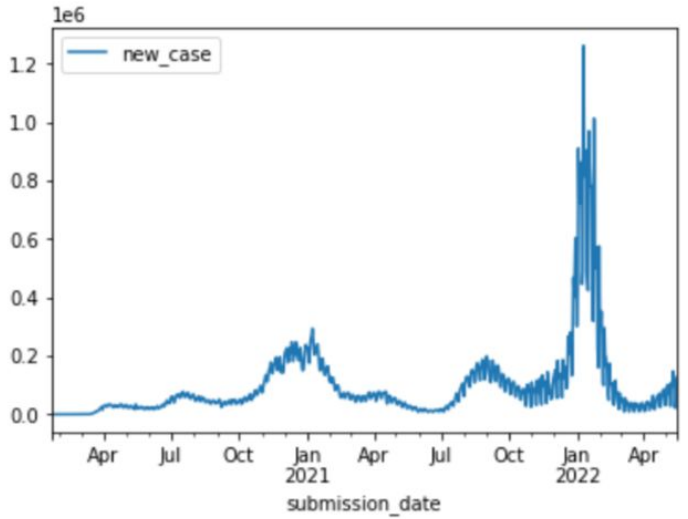
```
In [10]: dfNP.loc[:,'C4_Restrictions on gatherings'].rolling(window=2).apply(lambda x: x.iloc[1] - x.iloc[

array([nan,  0., -2.])
```

Step2: Exploring Case Data

1. Visualization of cases

```
dfC_proc.plot(y='new_case')
```

```
<AxesSubplot:xlabel='submission_date'>
```

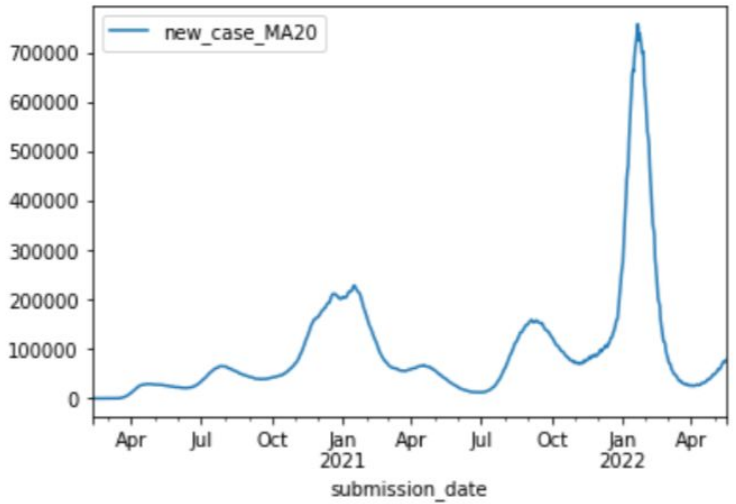


Step2: Exploring Case Data

2. Make it flatter and smoother by calculating the average number of each 20 days

submission_date	new_case	new_death	new_case_MA20	new_death_MA20
2020-02-10	0	0	0.60	0.00
2020-02-11	0	0	0.60	0.00
2020-02-12	1	0	0.60	0.00
2020-02-13	1	0	0.60	0.00
2020-02-14	0	0	0.60	0.00
...	...	...	...	...
2022-05-12	101200	253	70014.55	292.65
2022-05-13	116893	429	74490.90	311.00
2022-05-14	21733	46	74320.85	310.55
2022-05-15	45491	40	73985.40	302.90
2022-05-16	127814	231	77109.50	296.75

827 rows × 4 columns



Step3: Fusion

1. reset 0 in specfic columns

2. get the value and datelist

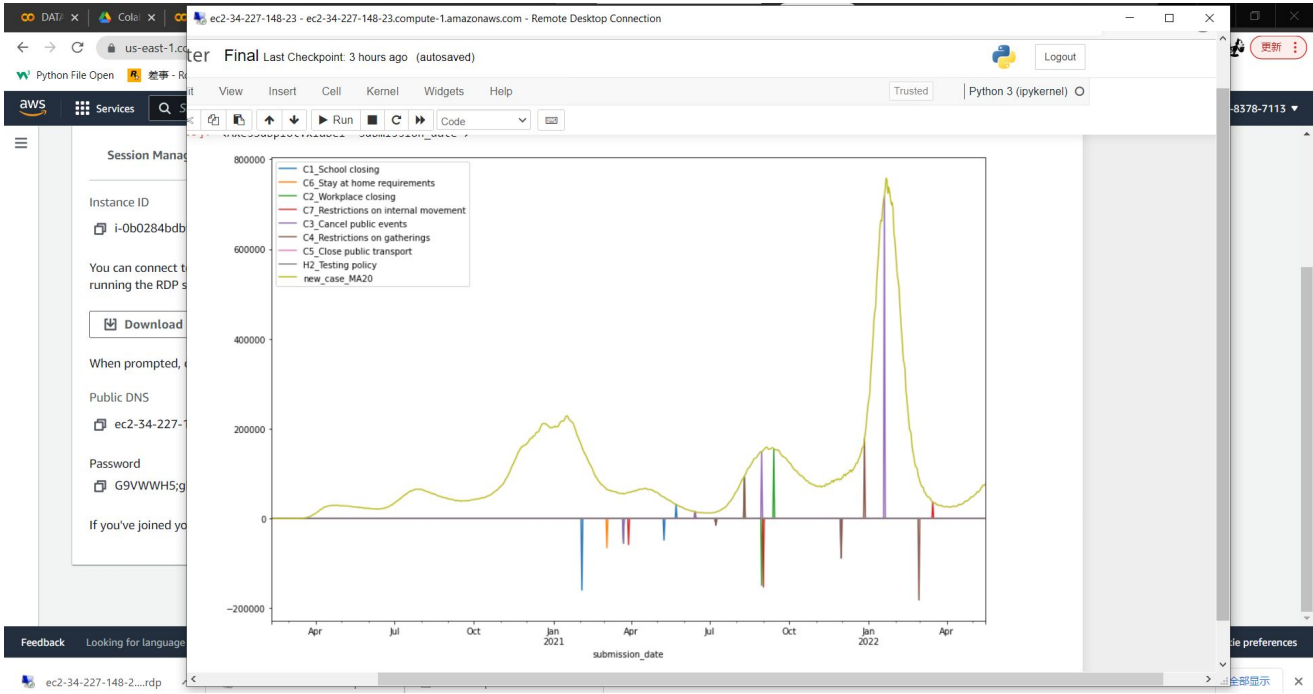
3. Plot

DATELIST

```
[{'policy_name': 'C1_School closing',
  'dates_of_change': [Timestamp('2020-03-05 00:00:00'),
   Timestamp('2021-02-03 00:00:00'),
   Timestamp('2021-05-09 00:00:00'),
   Timestamp('2021-05-23 00:00:00')],
  'value': [3.0, -1.0, -1.0, 1.0]},
 {'policy_name': 'C6_Stay at home requirements',
  'dates_of_change': [Timestamp('2020-03-15 00:00:00'),
   Timestamp('2020-07-20 00:00:00'),
   Timestamp('2020-10-13 00:00:00'),
   Timestamp('2020-10-26 00:00:00'),
   Timestamp('2020-11-16 00:00:00'),
   Timestamp('2021-03-04 00:00:00')],
  'value': [2.0, -1.0, 1.0, -1.0, 1.0, -1.0]},
```

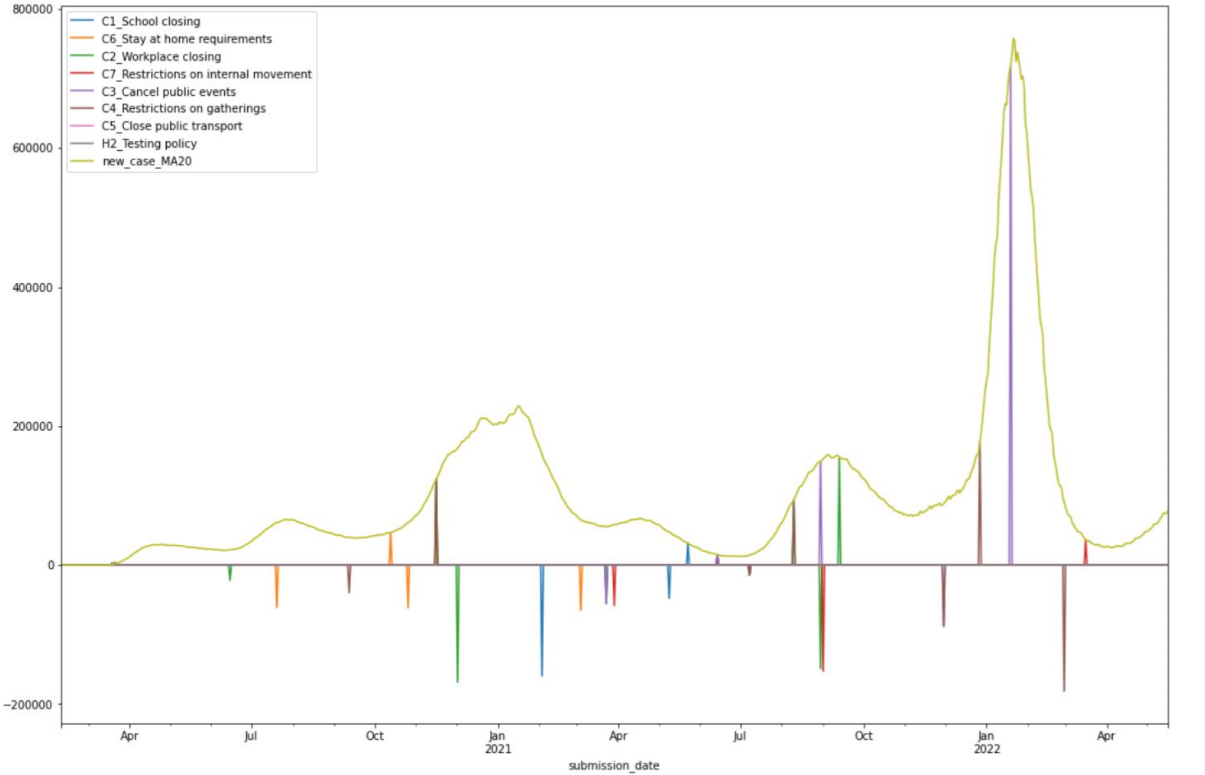
Techenology

1. AWS
2. Spark



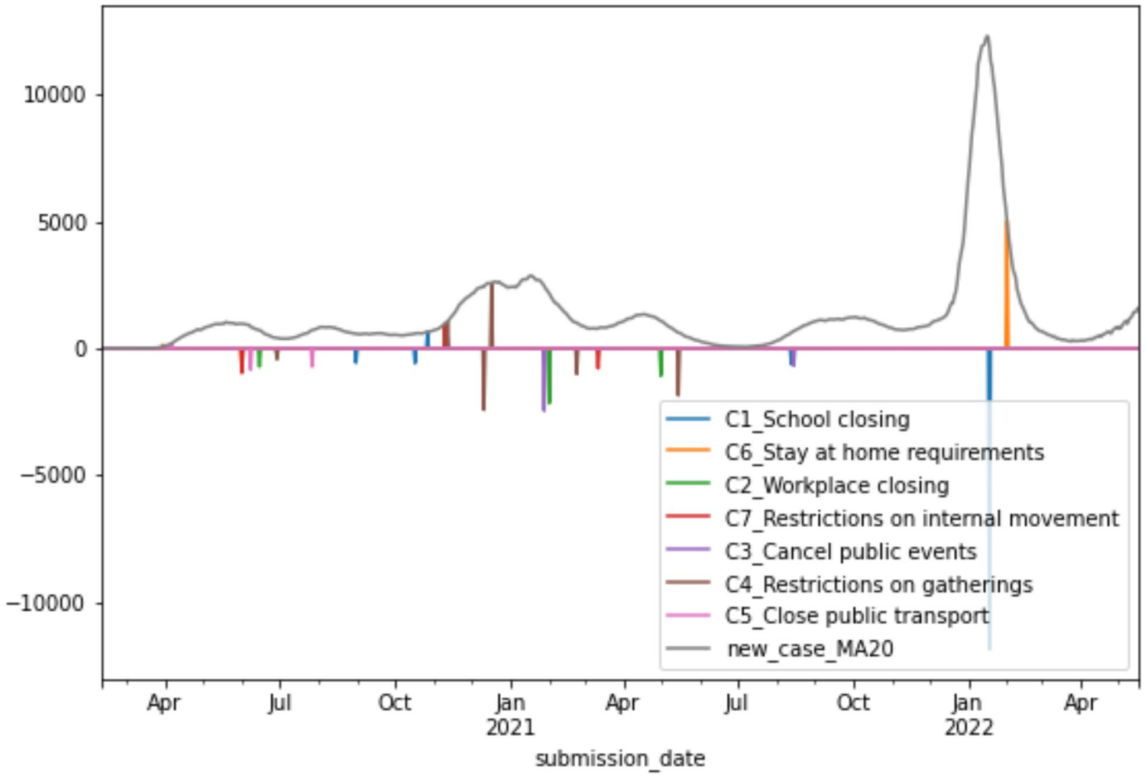
Results

COVID-19 Cases and Public Policy, United States



Results

COVID-19 Cases and Public Policy, Maryland



Conclusion

1. Responses tend to have a one-month lag time
2. Last peak was caused by the Omicron variant.
3. No policy changes made this year with regard to school closure.
4. Following the existing pattern, we expect to see additional policy changes in June following from recent COVID numbers.

Challenges

Compatibility

- Java 8/11 works for spark (version conflicts)
- Environment Variables / Paths
- Mac/ Windows/ Colab

Challenges

Data

- Evaluation (Sources (backup))
- Cleaning
- Visualization
- Missing data / facts?

Future work

1. Docker or other deployment technologies
2. Study relationships between differerent policies
3. Study specfic states (Using current cases)
4. Apply recommendation system algorithms

Github Repository

https://github.com/HenryVarro666/policy-changes-prediction

Thank you!