CCC Data Summary

Asa Gold

4/30/2022

Basic Summary

The Crowd Counting Consortium (CCC), updated weekly, currently contains ~100k observations between 2017-01-20 and 2022-12-10. Take a quick look at the their GitHub page and their website.

Below is a glimpse at all 69 variables (plus several I have added: **year** and **state_code**).

```
## Rows: 100,836
## Columns: 71
## $ date                 <date> 2017-01-20, 2017-01-21, 2017-01-21, 2017-01-21, ~
## $ locality             <chr> "Seattle", "Adak", "Anchorage", "Bethel", "Cordov~
## $ state                <chr> "WA", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "~
## $ location_detail      <chr> "University of Washington campus", NA, NA, NA, NA~
## $ online               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ type                 <chr> "protest; demonstration", "march", "march", "marc~
## $ title                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ macroevent           <chr> NA, "20170121-womensmarch", "20170121-womensmarch~
## $ actors               <chr> "general protestors", NA, NA, NA, NA, NA, NA, NA,~
## $ organizations        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ participants         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ claims               <chr> "anti-Yiannopolous; anti-Breitbart; anti-Trump", ~
## $ valence              <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ issues               <chr> "executive", "women's rights", "women's rights", ~
## $ size_text            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ size_low             <int> 200, 10, 2000, 40, 108, 2000, 105, 150, 900, 1000~
## $ size_high            <int> 700, 10, 4000, 100, 114, 2000, 105, 170, 900, 100~
## $ size_mean            <dbl> 450, 10, 3000, 70, 111, 2000, 105, 160, 900, 1000~
## $ size_cat             <int> 2, 1, 3, 1, 2, 3, 2, 2, 2, 3, 2, 2, 1, 1, 2, 1, 1~
## $ arrests              <chr> "1", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ arrests_any          <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ injuries_crowd       <chr> "1", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ injuries_crowd_any   <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ injuries_police      <chr> "0", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ injuries_police_any  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ property_damage      <chr> "1", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ property_damage_any  <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ chemical_agents      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ participant_measures <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ police_measures      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ participant_deaths   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ police_deaths        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
## $ source_1             <chr> "http://infoweb.newsbank.com.du.idm.oclc.org/reso~
## $ source_2             <chr> "http://www.seattletimes.com/seattle-news/northwe~
## $ source_3             <chr> "http://www.seattletimes.com/seattle-news/educati~
## $ source_4             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_5             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_6             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_7             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_8             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_9             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_10            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_11            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_12            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_13            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_14            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_15            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_16            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_17            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_18            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_19            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_20            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_21            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_22            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_23            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_24            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_25            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_26            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_27            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_28            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_29            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ source_30            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ notes                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ lat                  <dbl> 47.60621, 51.88000, 61.21806, 60.79222, 60.54241,~
## $ lon                  <dbl> -122.3321, -176.6581, -149.9003, -161.7558, -145.~
## $ resolved_locality    <chr> "Seattle", "Adak", "Anchorage", "Bethel", "Cordov~
## $ resolved_county      <chr> "King County", "Aleutians West", "Anchorage", "Be~
## $ resolved_state       <chr> "WA", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "~
## $ fips_code            <chr> "53033", "02016", "02020", "02050", "02261", "020~
## $ year                 <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2~
## $ state_code           <chr> "53", "02", "02", "02", "02", "02", "02", "02", "~
```

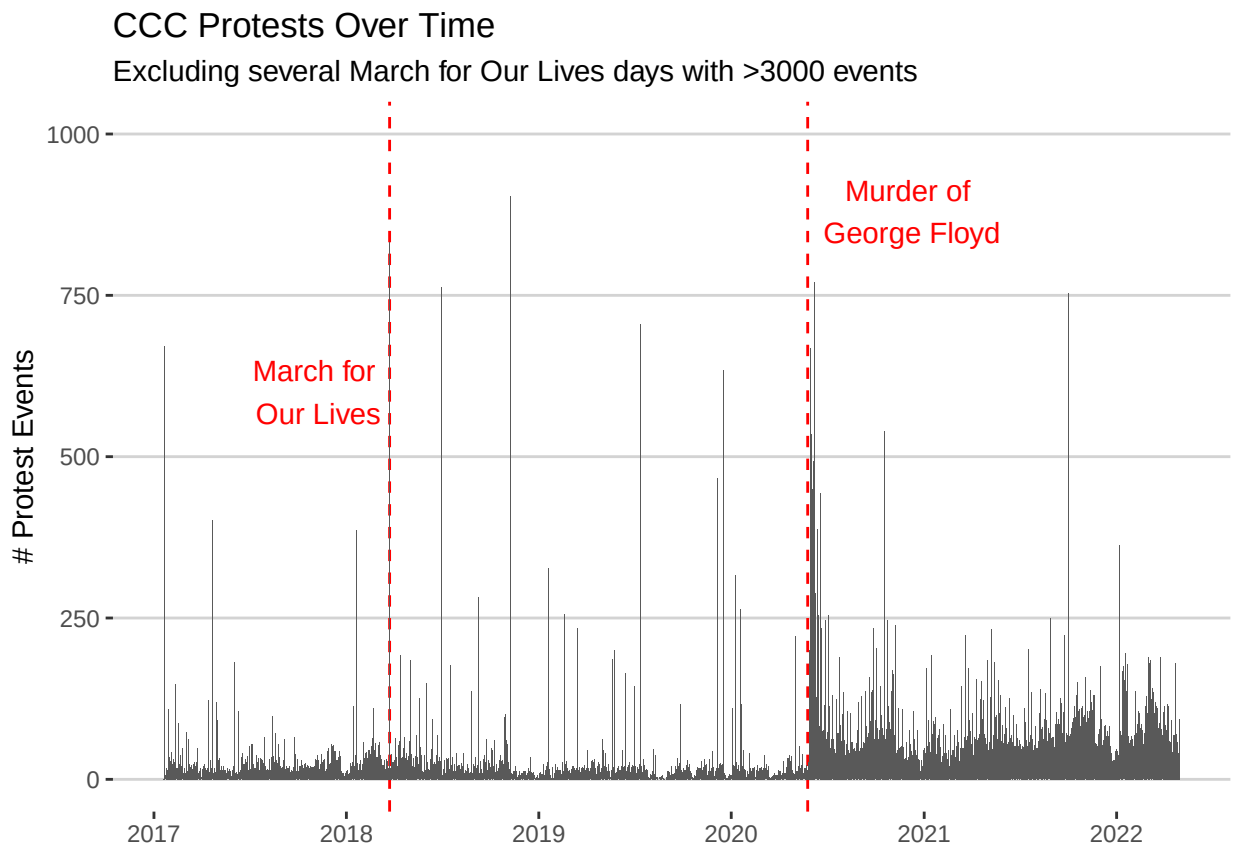
Several are particularly relevant.

- Descriptive Variables
  - **location_detail**: non-standardized descriptions of the place of the protest. Ex: Downtown Dayton, OSU Campus, Comcast Company Headquarters; City Hall
  - **actors**: Who engaged in the protest. Slightly more standardized than location_detail but still highly variable (25252 unique values). Ex: MAGA March, Counter-protesters; socialist Party USA, Democratic Club of the High Desert, Colorado MAGA March
  - **claims**: main grievances, goals, or rallying cries of the protestors. Ex: support LGBTQ, Pro Trump, oppose anti-Semitism
  - **valance**: political leaning of the protest; 0 = neutral; 1 = left; 2 = right.
  - I don't focus on **issues** and **macroevent** here, as neither will help us directly filter for college/university protests, but both may be helpful later on in figuring out larger goals or move-

ments associated with specific protests. Similarly, **type** may serve to further narrow our dataset if we want to exclude, say, prayer services.

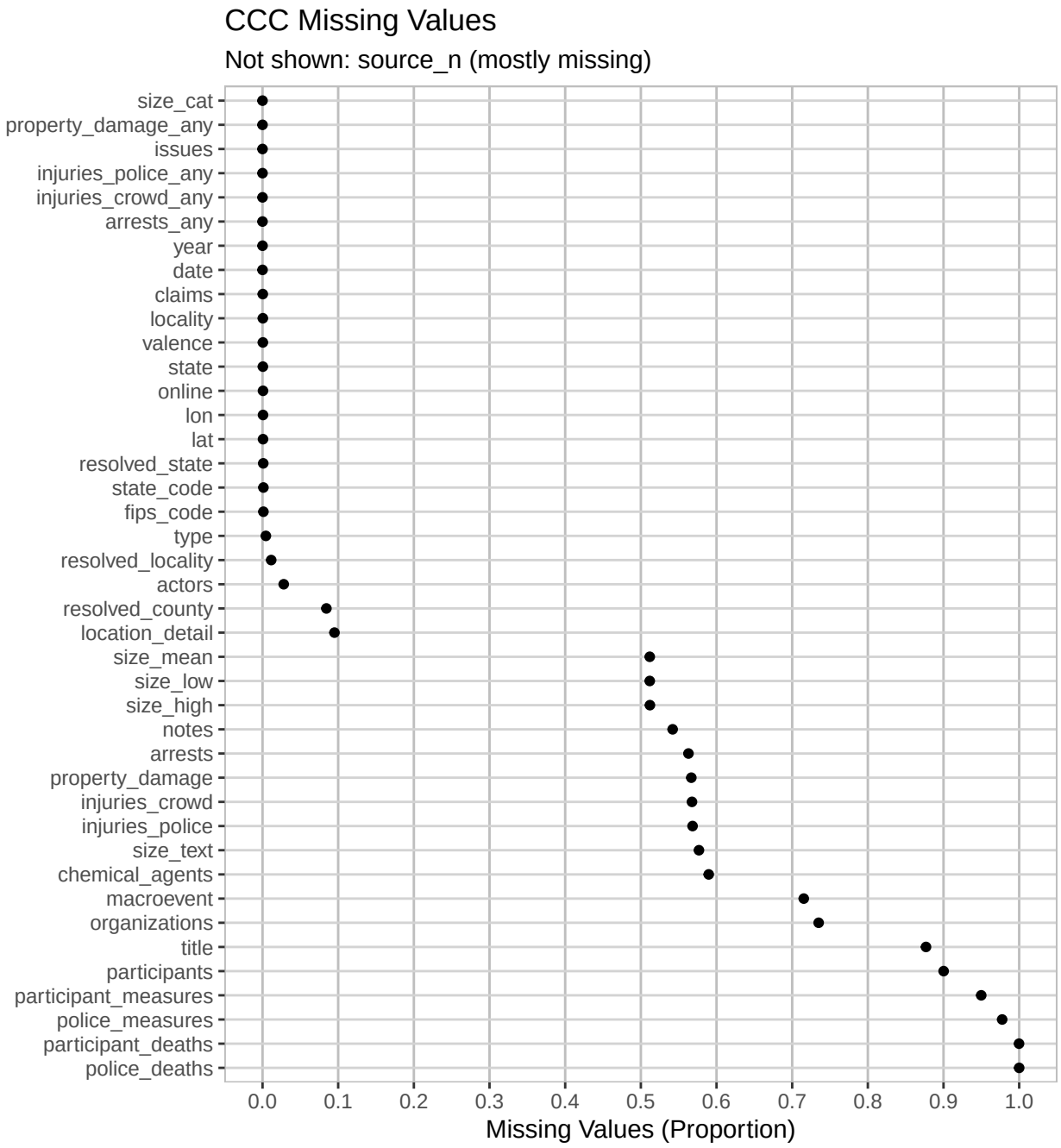
- Location Variables
  - **locality**: normally a city in or around which the protest event took place. Ex: Lancaster, Meadville, Philadelphia, Pittsburgh.
  - **fips_code**: county code associated with the locality

It is important to note that the relative frequency of protest events have ebbed and flowed in response to major nation-wide events. The plot below reports daily count of protest events between Jan 2017 and present. Note the spikes during and directly after several headline-making national events, like March for Our Lives (gun control) and George Floyd's murder (racism and police violence). This plot does not indicate the relative *size* of the protests; for more on size estimates, see **size_[low/mean/high]** variables (do note that size estimates have ~50% coverage; see Missing Values).



Missing Values

Below are proportions of missing values for all variables, except the *source_n* variables, which are largely missing.



Filtering

To restrict our dataset to only include protests occurring on college or university campuses or organized by college students, we use **location_detail** and **actors**. The code below filters the master dataset such that only those observations that reference university or college in *either* **location_detail** or **actors**. Further, we restrict the date range to before the present, remove non 50 state fips codes, and remove any online protests.

```
# NOTE: 'ccc' is the master dataset
college_filtered <- ccc |>

  filter(
    str_detect(actors,
               regex('(universit|college)', ignore_case = TRUE)
               ) |
      str_detect(location_detail,
                 regex('(universit|college)', ignore_case = TRUE))
  ) |>

  filter(date <= Sys.Date()) |>

  filter(!(state_code %in% c("66","72",NA))) |>
  filter(online != 1)
```

With this initial, naive filtering, we are left with 5028 observations. Below, inspect several **location_detail** and **actors** observations.

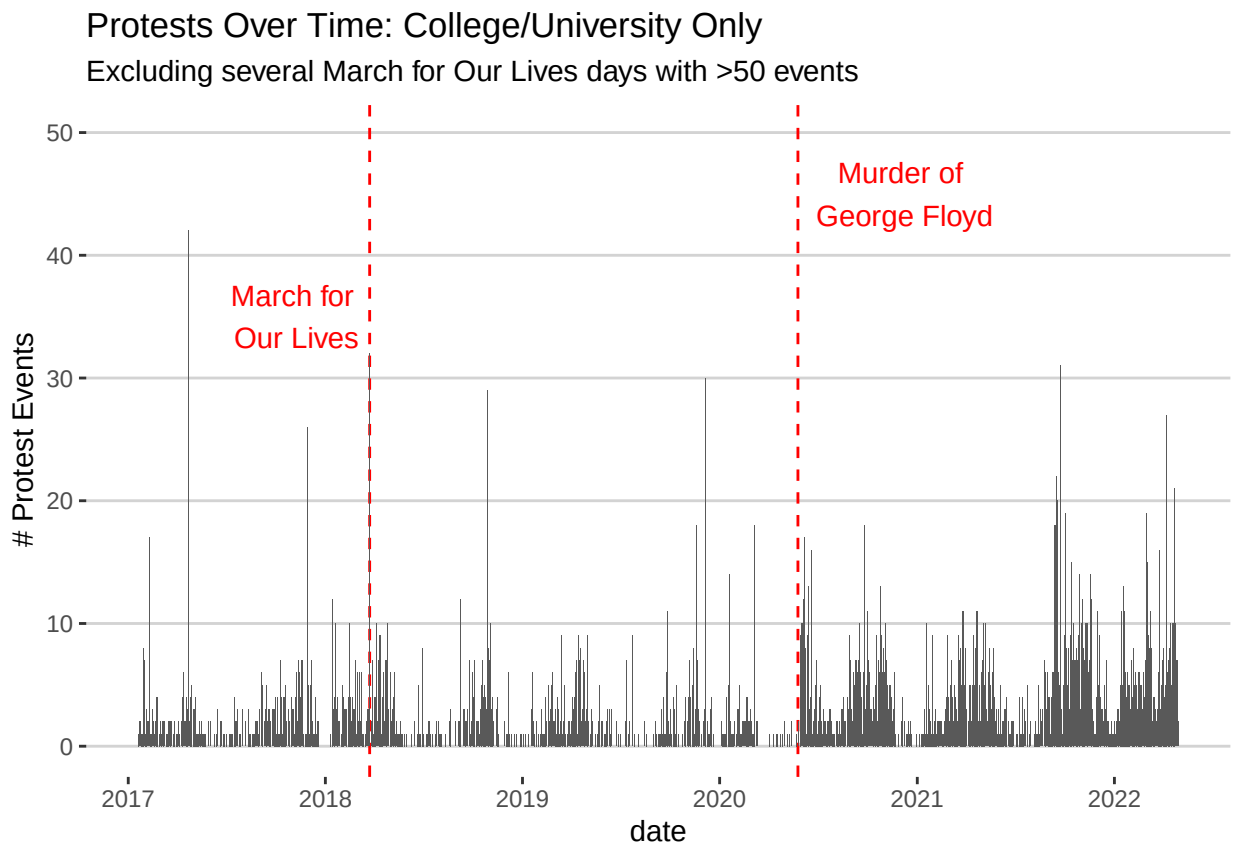
location_detail	actors
University of Washington campus	general protestors
Dartmouth College	students
Dartmouth College	mostly veterans; community members
University of Colorado; Watershed School	students
Oklahoma State University; outside of President Hargis' office	African American Student Alliance
University of Colorado	mostly students

While this initial filtering is highly effective, manual filtering may still be necessary. For instance, one March for Our Lives protest occurred near "College Lakes Recreation Center." Such observations cannot be excluded by naive filtering alone.

It is also appears that CCC considers multiple days of ongoing protests (e.g., encampment style) as distinct events. For instance, the Atlanta Student Movement staged an extended protest over the course of a week in October 2021. CCC has an entry for each day. These observations, however, are not mere duplicates; they differ slightly in participant makeup and sources. **We should discuss how best to handle such cases.**

Further, we should consider whether to include **claims** in our filtering effort. If **claims** references universities or colleges but neither **location_detail** nor **actors** does, should we include the observation in our dataset? There are ~330 such cases. For instance, a student researcher union at UC Berkeley stages protests against rollbacks in rights and benefits. Both **actor** and **location_detail** made no explicit reference to university or college.

The below plot indicates daily protest occurence in/around college or university campuses between Jan 2017 and present.



Geography

The college/university dataset contains observations from all 50 states and the District of Columbia. Below, observe the distribution of college protest events across states in each year since CCC began collecting data.

State	2017	2018	2019	2020	2021	2022	Total
CA	34	106	47	76	130	95	488
MI	22	52	28	52	123	115	392
NY	29	70	33	28	119	27	306
PA	31	57	42	72	52	22	276
IL	24	49	24	67	37	37	238
FL	31	42	20	59	54	22	228
TX	18	42	29	39	49	13	190
GA	16	19	19	21	106	8	189
OH	16	42	20	30	39	34	181
MA	17	42	19	27	46	29	180
IN	14	29	16	19	34	35	147
NC	12	38	17	40	26	9	142
WA	22	33	29	22	19	10	135
VA	21	24	14	24	29	14	126
CT	28	25	9	12	19	26	119
WI	14	17	10	24	27	18	110
CO	15	22	11	8	12	39	107

State	2017	2018	2019	2020	2021	2022	Total
DC	8	15	3	4	58	12	100
MN	14	15	14	17	31	7	98
NJ	8	17	21	13	25	6	90
OR	10	32	14	19	10	5	90
IA	7	27	13	11	25	6	89
MO	9	19	10	12	25	3	78
TN	9	22	8	12	19	4	74
MD	4	12	15	15	19	7	72
AZ	7	17	10	11	17	3	65
SC	8	15	6	12	13	5	59
UT	6	9	10	12	11	4	52
OK	10	5	6	15	10	3	49
KY	5	7	8	13	8	3	44
VT	4	12	11	9	6	1	43
NE	6	10	4	4	12	6	42
KS	10	9	5	6	10	0	40
AL	8	6	5	2	14	4	39
MS	7	3	12	9	7	1	39
NH	5	8	4	1	9	7	34
MT	6	9	5	2	9	2	33
WV	5	4	4	3	4	10	30
ME	4	9	5	4	1	4	27
LA	2	6	0	6	7	3	24
NM	6	2	0	6	7	3	24
AR	3	2	4	4	9	1	23
DE	6	4	3	1	3	2	19
NV	1	10	3	1	2	2	19
HI	3	7	4	1	2	0	17
RI	3	7	2	1	1	1	15
ID	2	0	3	3	3	0	11
AK	1	5	2	0	1	1	10
WY	3	1	2	2	1	1	10
ND	1	0	1	2	5	0	9
SD	1	0	2	0	2	1	6

Next, let's inspect county coverage. Taking as a case study California, which saw the most protests between 2017 and 2022, we can see which counties saw protests (left) and the raw count by county (right).