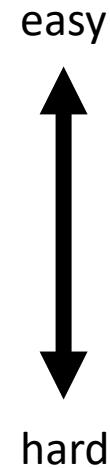


Online Data Sources

Lab 2

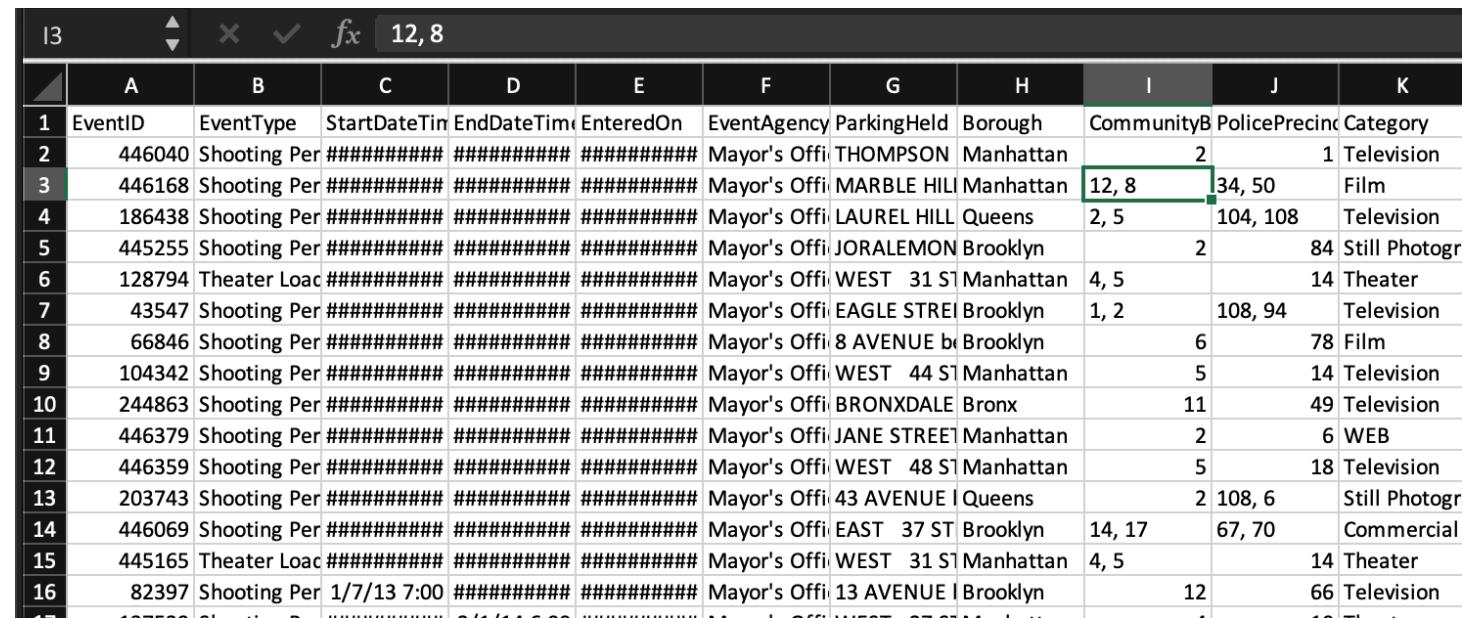
Agenda

- Publicly available tabular datasets
 - Data from research papers, repositories
 - Getting data with API's
 - Data within web pages
-
- R Markdown basics
 - Downloading and parsing web pages with R
 - Dealing with unstructured data: **regular expressions**



Publicly available tabular data

- [50 Amazing Free Data Sources You Should Know](#)
- [Data.gov](#)
- [NYC OpenData](#)
- [Google Dataset Search](#)
- [US Census Table viewer](#)



	A	B	C	D	E	F	G	H	I	J	K
1	EventID	EventType	StartDateTim	EndDateTime	EnteredOn	EventAgency	ParkingHeld	Borough	CommunityB	PolicePrecinct	Category
2	446040	Shooting Per	#####	#####	#####	Mayor's Offi	THOMPSON	Manhattan	2	1	Television
3	446168	Shooting Per	#####	#####	#####	Mayor's Offi	MARBLE HILL	Manhattan	12, 8	34, 50	Film
4	186438	Shooting Per	#####	#####	#####	Mayor's Offi	LAUREL HILL	Queens	2, 5	104, 108	Television
5	445255	Shooting Per	#####	#####	#####	Mayor's Offi	JORALEMON	Bronx	2	84	Still Photogr
6	128794	Theater Loac	#####	#####	#####	Mayor's Offi	WEST 31 ST	Manhattan	4, 5	14	Theater
7	43547	Shooting Per	#####	#####	#####	Mayor's Offi	EAGLE STREET	Brooklyn	1, 2	108, 94	Television
8	66846	Shooting Per	#####	#####	#####	Mayor's Offi	8 AVENUE	Brooklyn	6	78	Film
9	104342	Shooting Per	#####	#####	#####	Mayor's Offi	WEST 44 ST	Manhattan	5	14	Television
10	244863	Shooting Per	#####	#####	#####	Mayor's Offi	BRONXDALE	Bronx	11	49	Television
11	446379	Shooting Per	#####	#####	#####	Mayor's Offi	JANE STREET	Manhattan	2	6	WEB
12	446359	Shooting Per	#####	#####	#####	Mayor's Offi	WEST 48 ST	Manhattan	5	18	Television
13	203743	Shooting Per	#####	#####	#####	Mayor's Offi	43 AVENUE	Queens	2	108, 6	Still Photogr
14	446069	Shooting Per	#####	#####	#####	Mayor's Offi	EAST 37 ST	Brooklyn	14, 17	67, 70	Commercial
15	445165	Theater Loac	#####	#####	#####	Mayor's Offi	WEST 31 ST	Manhattan	4, 5	14	Theater
16	82397	Shooting Per	1/7/13 7:00	#####	#####	Mayor's Offi	13 AVENUE	Brooklyn	12	66	Television

Data from research papers

- Nature repository guidelines
- Example: Harvard Dataverse

Ethics. Permits were obtained from the Bavarian government and the Bavarian regional office for forestry (LWF).

Data accessibility. Data available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.rv15dv44s> [65].

Authors' contributions. All authors conceived the idea and designed the study; B.K. conducted the paternity analyses; K.B.B. and D.R.F. analysed the data with input from B.K.; K.B.B., D.R.F. and B.K. wrote the manuscript.

Getting data with API's

Application Programming Interface

The screenshot shows the Twitter Developer Platform documentation page. The left sidebar has a dark theme with white text. It includes sections like 'Documentation', 'Search the docs', 'Twitter API' (with a house icon), 'Getting started', 'Tutorials', 'Tools and libraries', 'What to build', 'Migrate', and 'API reference index'. Below this is the 'Twitter API v2' section with 'Fundamentals' expanded, showing 'Data dictionary'. The main content area has a light background. It starts with a title 'Retrieving a Tweet object' and a 'Sample Request' section containing a curl command. The command is:

```
curl --request GET 'https://api.twitter.com/2/tweets?ids=1212092628029698048&tweet.fields=attachments,author_id,context_annotations,created_at,&--header 'Authorization: Bearer $BEARER_TOKEN'
```

Below this is a 'Sample Response' section showing a JSON object:

```
1 {  
2   "data": [  
3     {  
4       "id": "1212092628029698048",  
5       "text": "We believe the best future version of our API will come from building it",  
6       "possibly_sensitive": false,  
7       "referenced_tweets": [  
8         {  
9           "type": "quoted_tweet"  
10          "id": "1212092628029698048"  
11        }  
12      ]  
13    }  
14  ]  
15}
```

Getting data with API's



1. Start your query with the host name:

<https://api.census.gov/data>

2. Add the data year to the URL:

<https://api.census.gov/data/2019>

This is the year that the data were estimated.

3. Add the dataset name acronym:

<https://api.census.gov/data/2019/pep/charagegroups>

This is the base URL for this dataset. You can find dataset names by browsing the discovery tool: <https://api.census.gov/data.html>

4. Add ?get= to the query

<https://api.census.gov/data/2019/pep/charagegroups?get=>

5. Add your variables:

<https://api.census.gov/data/2019/pep/charagegroups?get=NAME,POP>

Data within web pages

 www.eFloras.org

Flora of North America

FNA Vol. 7 Page 465, 483

[FNA | Family List](#) | [FNA Vol. 7](#) | [Brassicaceae](#) | [Cardamine](#)

37. *Cardamine rotundifolia* Michaux, Fl. Bor.-Amer. 2: 30. 1803.

Dentaria rotundifolia (Michaux) Greene

Perennials; glabrous throughout. Rhizomes slender, to 2 mm diam. Stems usually procumbent, sometimes (not flexuous), unbranched or branched, (1-) 1.5-3(-5) dm. Rhizomal leaves absent. Basal leaves (soon withered), not rosulate. Cauline leaves simple, petiolate, petiole (0.3-)0.5-2.5(-4) cm, base not auriculate; blade oblong, suborbicular, or cordate, (0.5-) 1-4.5(-5.5) cm × (5-)10-40(-54) mm, base cordate, rounded, or truncate, margin repand, or sinuate, (distally with shorter petiole, blade smaller). Racemes ebracteate. Fruiting pedicels divaricating, (6-)10-15(-20) mm. Flowers: sepals oblong, 2-3 × 1-1.7 mm, lateral pair not saccate basally; petals broadly oblanceolate, spreading, 5-7(-8) × 2-3 mm, (not clawed, apex rounded); filaments: median pairs 3.5-4.5 mm, lateral pair 2.7-3.5 mm; anthers oblong, 0.8-1.2 mm. Fruits linear, (torulose), 1-1.5(-2) cm × 0.8-1.1 mm; ovules per ovary; style 1.5-2.5 mm. Seeds brown, oblong to ovoid, 0.9-1.1 × 0.5-0.6 mm.

Flowering Apr-Jun. Stream banks, swamps, low woodland, wet rocky areas, seepage areas; 150-400 m; Del. Md., N.J., N.Y., N.C., Ohio, Pa., Tenn., Va., W.Va.

Related Objects

- [Distribution Map](#) Map

Related Links (opens in a new window)

Other Databases

- [W³TROPICOS](#)
- [IPNI](#)

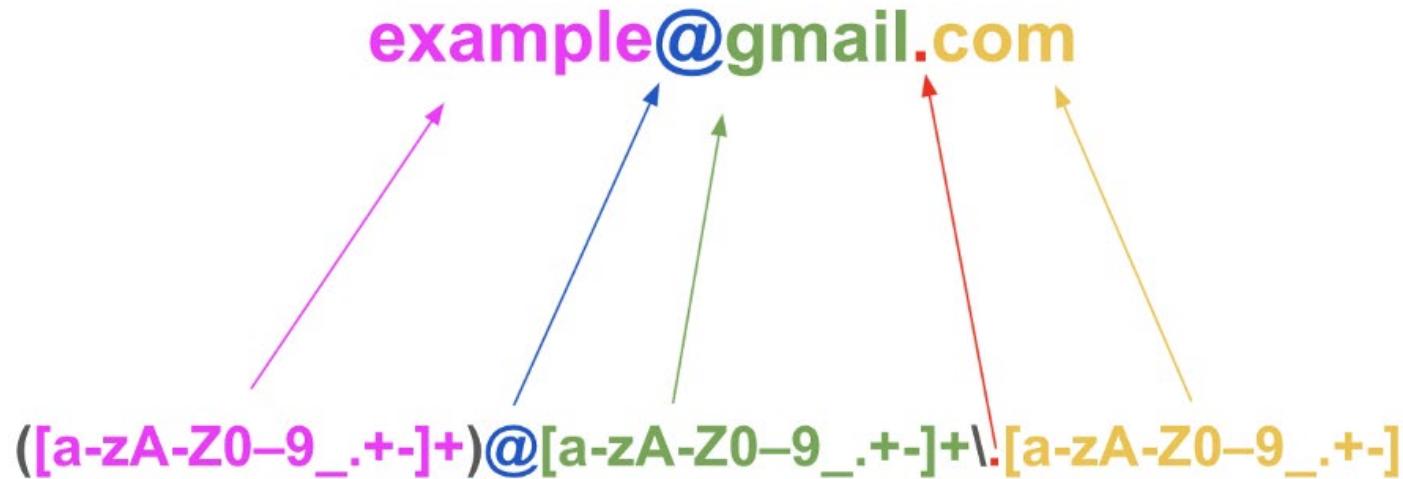
| [eFlora Home](#) | [People Search](#) | [Help](#) | [ActKey](#) | [Hu Cards](#) | [Glossary](#) |

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<!-- OutputCache Duration=60 VaryByParam="flora_id;taxon_id" %--&gt;
&lt;head&gt;...&lt;/head&gt;
&lt;!-- body onload="formMOMap.submit()" --&gt;
&lt;body&gt; == $0
  &lt;table id="TableMain" cellspacing="1" cellpadding="1" width="640" border="0"&gt;
    &lt;tbody&gt;
      &lt;tr&gt;...&lt;/tr&gt;
      &lt;tr&gt;...&lt;/tr&gt;
      &lt;tr&gt;
        &lt;td id="tdBody"&gt;
          &lt;table id="tableBody" cellspacing="0" cellpadding="0" width="100%" border="0"&gt;
            &lt;tbody&gt;
              &lt;tr id="tdLeftBar" valign="top" width="0"&gt;&lt;/tr&gt;
              &lt;tr id="tdContent" valign="top" width="1930"&gt;
                &lt;table width="100%" border="0"&gt;
                  &lt;tbody&gt;
                    &lt;tr&gt;
                      &lt;td valign="top"&gt;
                        &lt;table id="tableTaxonChain" cellspacing="0" cellpadding="1" width="100%" border="0"&gt;...&lt;/table&gt;
                        &lt;p&gt;&lt;/p&gt;
                      &lt;/td&gt;
                    &lt;/tr&gt;
                  &lt;/tbody&gt;
                &lt;/table&gt;
                &lt;div id="panelTaxonTreatment"&gt;
                  &lt;span id="lblTaxonDesc"&gt;
                    &lt;!-- FNA Template --&gt;
                    " 37. "
                    &lt;b&gt;Cardamine rotundifolia&lt;/b&gt;
                    " Michaux, Fl. Bor.-Amer. 2: 30. 1803. "
                    &lt;p&gt;&lt;/p&gt;
                    &lt;p&gt;&lt;/p&gt;
                    &lt;p&gt;&lt;/p&gt;
                    &lt;p&gt;&lt;/p&gt;
                    &lt;p&gt;&lt;/p&gt;
                    &lt;p&gt;
                      &lt;b&gt;Perennials;&lt;/b&gt;
                      " glabrous throughout. "
                      &lt;b&gt;Rhizomes&lt;/b&gt;
                      " slender, to 2 mm diam. "
                      &lt;b&gt;Stems&lt;/b&gt;
                      " usually procumbent, sometimes erect, (not flexuous), unbranched or branched, (1-) 1.5-3(-5) dm. "
                      &lt;b&gt;Rhizomal leaves&lt;/b&gt;
                      " absent. "
                      &lt;b&gt;Basal leaves&lt;/b&gt;
                      " (soon withered), not rosulate. "
                    &lt;/p&gt;
                  &lt;/span&gt;
                &lt;/div&gt;
              &lt;/tr&gt;
            &lt;/tbody&gt;
          &lt;/table&gt;
        &lt;/td&gt;
      &lt;/tr&gt;
    &lt;/tbody&gt;
  &lt;/table&gt;</pre>
```

Managing text with regular expressions

search_person	search_vehicle	
TRUE	TRUE	[39] "VIOLATION OF TRANSPORTATION/VEHICLE LAWS WATER SAFETY ACT"
FALSE	FALSE	[40] "PRE-EXISTING KNOWLEDGE VIOLATION OF CITY ORDINANCE"
FALSE	FALSE	[41] "CONSENSUAL CONTACT PRE-EXISTING KNOWLEDGE"
FALSE	FALSE	[42] "PRE-EXISTING KNOWLEDGE OTHER"
FALSE	FALSE	[43] "CALL FOR SERVICE PRE-EXISTING KNOWLEDGE"
FALSE	FALSE	[44] "VIOLATION OF PENAL CODE PRE-EXISTING KNOWLEDGE"
TRUE	TRUE	[45] "PRE-EXISTING KNOWLEDGE MOTOR VEHICLE DRIVER"
FALSE	FALSE	[46] "VIOLATION OF CITY ORDINANCE MOTOR VEHICLE DRIVER"
FALSE	FALSE	[47] "CALL FOR SERVICE WATER SAFETY ACT"
FALSE	FALSE	[48] "CONSENSUAL CONTACT SUSPICIOUS PERSON / VEHICLE CALL FOR SERVICE"
FALSE	FALSE	[49] "VIOLATION OF TRANSPORTATION/VEHICLE LAWS CALL FOR SERVICE PRE-EXISTING KNOWLEDGE"
TRUE	TRUE	[50] "CONSENSUAL CONTACT SUSPICIOUS PERSON / VEHICLE PRE-EXISTING KNOWLEDGE"
FALSE	FALSE	[51] "CALL FOR SERVICE MOTOR VEHICLE DRIVER"
FALSE	FALSE	[52] "SUSPICIOUS PERSON / VEHICLE MOTOR VEHICLE DRIVER"
FALSE	FALSE	[53] NA
FALSE	FALSE	[54] "CALL FOR SERVICE PRE-EXISTING KNOWLEDGE MOTOR VEHICLE DRIVER"
FALSE	FALSE	[55] "CONSENSUAL CONTACT MOTOR VEHICLE DRIVER"
FALSE	FALSE	[56] "NA VIOLATION OF TRANSPORTATION/VEHICLE LAWS NA CONSENSUAL CONTACT VIOLATION OF TR
FALSE	FALSE	[57] "VIOLATION OF TRANSPORTATION/VEHICLE LAWS SUSPICIOUS PERSON / VEHICLE PRE-EXISTING
TRUE	FALSE	[58] "OTHER WATER SAFETY ACT"
TRUE	FALSE	[59] "CONSENSUAL CONTACT CALL FOR SERVICE"
FALSE	FALSE	[60] "SUSPICIOUS PERSON / VEHICLE PRE-EXISTING KNOWLEDGE OTHER"
FALSE	FALSE	[61] "VIOLATION OF TRANSPORTATION/VEHICLE LAWS CONSENSUAL CONTACT CALL FOR SERVICE"
		[62] "NA VIOLATION OF PENAL CODE VIOLATION OF TRANSPORTATION/VEHICLE LAWS PRE-EXISTING

Managing text with regular expressions

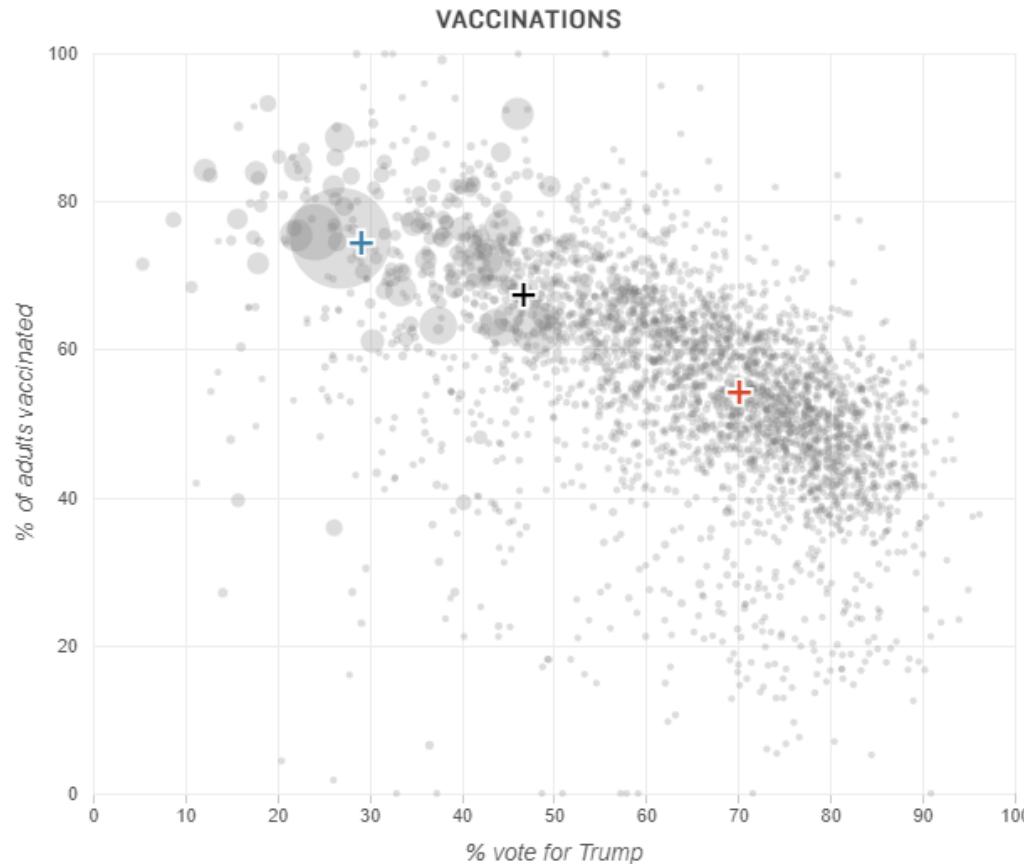


- More about learning regular expressions (regex)
- Regex cheat sheet

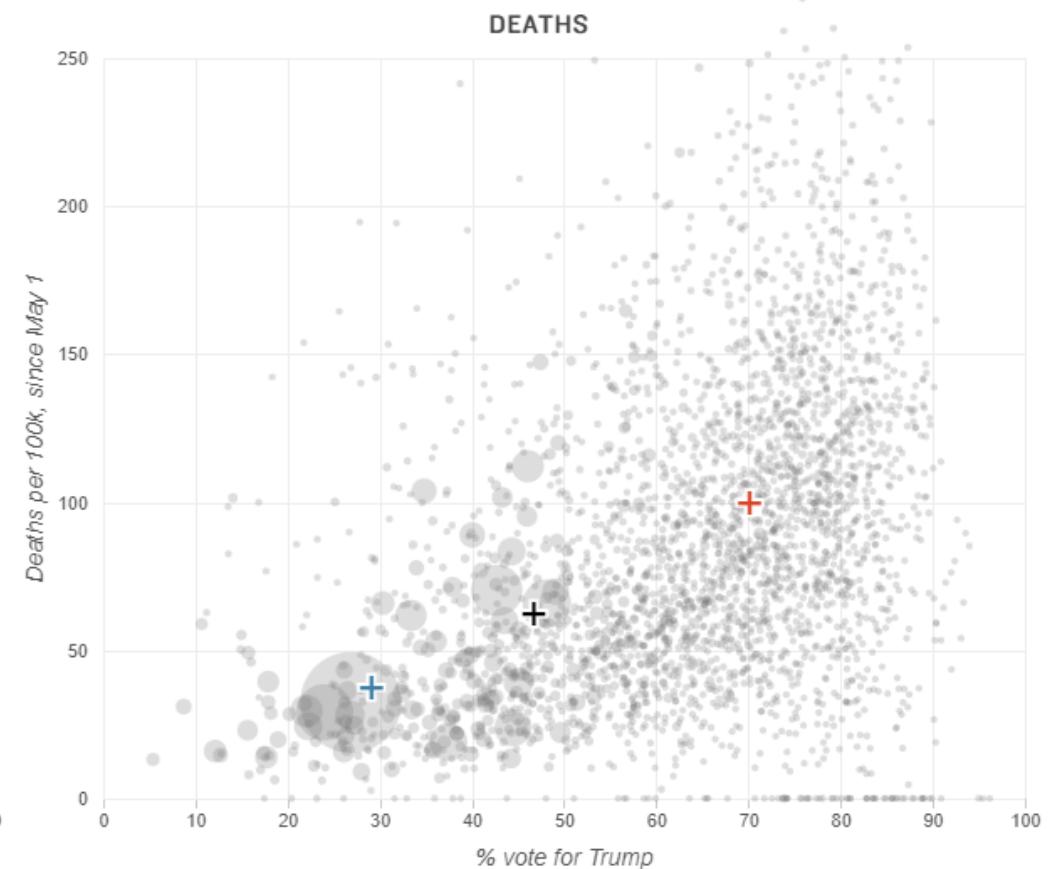
Pro-Trump higher Misinfo

Updated December 10
Heard on Morning Edition

DANIEL WOOD



Counties that went heavily for Donald Trump have seen much lower vaccination rates and much higher death rates from COVID



Integrating data from multiple sources

Methodology

Vaccination rate data are the rate of vaccination among all people 18 years of age or older, as of Nov. 30. They are from the [Centers for Disease Control and Prevention](#).

COVID-19 deaths per 100,000 residents are calculated by dividing the deaths from COVID-19 in a county since May 1 by the county's population. County population data come from the [U.S. Census Bureau's 2020 decennial census](#). May 1 was chosen as the start date of our analysis because that is roughly the time when vaccines became universally available to adults ages 18 and older. COVID-19 death data is collected by the [Center for Systems Science and Engineering \(CSSE\)](#) at Johns Hopkins University and is current as of Nov. 30. COVID-19 death data for Florida and Utah are from the [May 2 and December 1 editions of the COVID-19 Community Profile Report](#), produced by the White House COVID-19 Team.

2020 election result data are from [MIT Election Data and Science Lab](#).

Integrating data from multiple sources

Adding variables (columns)
needs **same # rows**, and a linking variable

	A	B	C
1			
2			
3			
4			
5			
6			

+

	C	D	E
1			
2			
3			
4			
5			
6			

Adding observations (rows)
needs **same # columns**

	A	B	C
1			
2			
3			

+

	A	B	C
4			
5			
6			

Integrating data from multiple sources

What do we do here?

	A	B	C
1			
2			
3			
4			
5			
6			



	C	D	E
1			
2			
3			



	A	B	C	D	E
1					
2					
3					
4				NA	NA
5				NA	NA
6				NA	NA