CCES Cumulative Common Content (2006 - 2018)

Shiro Kuriwaki*

Guide last updated: 2019-09-01

To cite the dataset:

> Kuriwaki, Shiro, 2018, "Cumulative CCES Common Content (2006-2017)", doi:10.7910/DVN/II2DB6, Harvard Dataverse

This dataset combines thirteen years (2006 - 2018) of the Cooperative Congressional Election Study (Principal Investigators: Stephen Ansolabehere, Sam Luks, Brian Schaffner).

The Cooperative Congressional Election Study (CCES) is an online survey conducted around November of each year, asking a range of questions on political behavior and public opinion. Questions can change from year to year; this cumulative file includes standard questions asked multiple years.

This dataset was constructed from CCES datasets from each year. The final product is a `tibble`-style data frame (built in R) that is also available as a Stata `dta` file. In addition, the same dataset is available on `Crunch`, an analytics interface optimized for survey datasets.

Please note that this cumulative dataset makes some modifications to the original CCES datasets for comparability. across years These modifications are only made when differences are deemed sufficiently minor, and are documented in source code (see below). However, for details on the survey methodology and a list of all questions, readers should consult the guides for each year.

- **To see the source code,** report a bug, or ask a question about the data, please feel free to file an issue from the source code page: https://github.com/kuriwaki/cces_cumulative. Alternatively, please contact me by email.

- **To obtain the individual year's CCES datasets,** search the CCES dataverse (https://dataverse.harvard.edu/dataverse/cces) or access the CCES homepage (https://cces.gov.harvard.edu/). Sign-up to the Crunch dataset from the homepage as well.

- **To examine the survey methodology,** consult the Frequentl y Asked Questions Page (https://cces.gov.harvard.edu/frequently-asked-questions) or the Methodology section of a recent Common Content's codebook (https://doi.org/10.7910/DVN/GDF6Z0).

*Department of Government, Harvard University. ORCID: https://orcid.org/0000-0002-5687-2647. Bug reports welcome. My thanks to Alexander Agadjanian, Steve Ansolabehere, Stephen DiMauro, Bernard Fraga, Nathan Kaplan, Mayya Komisarchik, and Stephen Pettigrew for their contributions. Thanks to Joe Williams at YouGov, and Jon Keane, Mike Malecki, and Gordon Shotwell atCrunch for their help.

Contents

Getting Started

The `.Rds` format can be read into R. This format preserves dataset properties such as the distinction between integers and doubles, and labelled variables. Unlike a `.Rdata` file, an `.Rds` file is assigned to an object.

```
df <- readRDS("cumulative_2006_2018.Rds")
```

The dataset in R is best viewed in `dplyr`, although it can be analyzed as a standard data frame.

```
library(tidyverse)
df
```

```
# A tibble: 452,755 x 80
    year case_id weight weight_cumulati~ state st    cd     dist dist_up
   <int>   <int>  <dbl>            <dbl> <chr> <chr> <chr> <int>   <int>
 1  2006  439219  1.85             1.67  Nort~ NC    NC-10    10      10
 2  2006  439224  0.968            0.872 Ohio  OH    OH-3      3       3
 3  2006  439228  1.59             1.44  New ~ NJ    NJ-1      1       1
 4  2006  439237  1.40             1.26  Illi~ IL    IL-9      9       9
 5  2006  439238  0.903            0.813 New ~ NY    NY-22    22      22
 6  2006  439242  0.839            0.756 Texas TX    TX-11    11      11
 7  2006  439251  0.777            0.700 Minn~ MN    MN-3      3       3
 8  2006  439254  0.839            0.756 Neva~ NV    NV-2      2       2
 9  2006  439255  0.331            0.299 Texas TX    TX-24    24      24
10  2006  439263  1.10             0.993 Mary~ MD    MD-2      2       2
# ... with 452,745 more rows, and 71 more variables: cong <int>,
#   cong_up <int>, zipcode <chr>, county_fips <chr>, tookpost <int+lbl>,
#   weight_post <dbl>, rvweight <dbl>, rvweight_post <dbl>,
#   starttime <dttm>, pid3 <int+lbl>, pid3_leaner <int+lbl>,
#   pid7 <int+lbl>, ideo5 <fct>, gender <int+lbl>, birthyr <int>,
#   age <int>, race <int+lbl>, hispanic <int+lbl>, educ <int+lbl>,
#   faminc <fct>, marstat <int+lbl>, economy_retro <int+lbl>,
#   newsint <int+lbl>, approval_pres <int+lbl>, approval_rep <fct>,
#   approval_sen1 <fct>, approval_sen2 <fct>, approval_gov <int+lbl>,
#   intent_pres_08 <fct>, intent_pres_12 <fct>, intent_pres_16 <fct>,
#   voted_pres_08 <fct>, voted_pres_12 <fct>, voted_pres_16 <fct>,
#   vv_regstatus <fct>, vv_party_gen <fct>, vv_party_prm <fct>,
#   vv_turnout_gvm <fct>, vv_turnout_pvm <fct>, intent_rep <fct>,
#   intent_rep_party <fct>, voted_rep <fct>, voted_rep_party <fct>,
#   intent_gov <fct>, intent_gov_party <fct>, voted_gov <fct>,
#   voted_gov_party <fct>, intent_sen <fct>, intent_sen_party <fct>,
#   voted_sen <fct>, voted_sen_party <fct>, intent_rep_chosen <chr>,
#   intent_rep_fec <chr>, intent_sen_chosen <chr>, intent_sen_fec <chr>,
#   intent_gov_chosen <chr>, intent_gov_fec <chr>, voted_rep_chosen <chr>,
#   voted_rep_fec <chr>, voted_sen_chosen <chr>, voted_sen_fec <chr>,
#   voted_gov_chosen <chr>, voted_gov_fec <chr>, rep_current <chr>,
#   rep_icpsr <dbl>, sen1_current <chr>, sen1_icpsr <dbl>,
#   sen2_current <chr>, sen2_icpsr <dbl>, gov_current <chr>, gov_fec <chr>
```

A Stata dta file is provided as well. `cumulative_2006_2018.dta` can be read by Stata, or in R by the haven package

```
library(haven)
df <- read_dta("cumulative_2006_2018.dta")
```

Labelled variables (for analysis in R)

A note on variable types. The R dataset stores variables in `numeric`, `character`, `factor`, or `labelled` class.[1] The first three classes are commonly used, but the `lablelled` format is more novel. `labelled` classes are numeric integers where each integer is associated with a label (See vignette here). This makes it the same as `factor` but alwasys ordered and referenceable by its numeric value. It is essentially the same idea as labels in Stata and SPSS. It is built around R's `haven` package, which includes more documentation.

A labelled variable's labels are not usually shown. Recent versions of the `haven` package (version 2.1.0 or above) will display the associated labels in the Console if selected within a tibble (a dataframe in tidyverse). This makes it immediately obvious which value is associated with which label.

```
select(df, year, case_id, pid3)
```

```
# A tibble: 452,755 x 3
    year case_id            pid3
   <int>   <int>       <int+lbl>
 1  2006  439219 1 [Democrat]
 2  2006  439224 4 [Other]
 3  2006  439228 1 [Democrat]
 4  2006  439237 1 [Democrat]
 5  2006  439238 1 [Democrat]
 6  2006  439242 3 [Independent]
 7  2006  439251 2 [Republican]
 8  2006  439254 1 [Democrat]
 9  2006  439255 1 [Democrat]
10  2006  439263 1 [Democrat]
# ... with 452,745 more rows
```

But labels can be made explicit transforming the labelled vector into a factor. On the other hand, removes the numerical ordering and value codes in the labelled class.

```
library(haven)
select(df, year, case_id, pid3) %>%
  mutate(pid3_fct = as_factor(pid3))
```

```
# A tibble: 452,755 x 4
    year case_id            pid3 pid3_fct
   <int>   <int>       <int+lbl> <fct>
 1  2006  439219 1 [Democrat]    Democrat
 2  2006  439224 4 [Other]       Other
 3  2006  439228 1 [Democrat]    Democrat
 4  2006  439237 1 [Democrat]    Democrat
 5  2006  439238 1 [Democrat]    Democrat
```

[1] Technically, this is now called a `labelled_haven` class, to disambiguate from an unrelated but older use of `labelled` in the Hmisc package.

```
 6  2006  439242 3 [Independent] Independent
 7  2006  439251 2 [Republican]  Republican
 8  2006  439254 1 [Democrat]    Democrat
 9  2006  439255 1 [Democrat]    Democrat
10  2006  439263 1 [Democrat]    Democrat
# ... with 452,745 more rows
```

and unlike factors, they can be referenced by their underlying numeric value. In short, it is sometimes useful to treat survey values as numbers rather than raw text, and the labelled calss allows you to do that.

```
select(df, year, case_id, pid3) %>%
  filter(pid3 == 1)
```

```
# A tibble: 160,637 x 3
    year case_id         pid3
   <int>   <int>    <int+lbl>
 1  2006  439219 1 [Democrat]
 2  2006  439228 1 [Democrat]
 3  2006  439237 1 [Democrat]
 4  2006  439238 1 [Democrat]
 5  2006  439254 1 [Democrat]
 6  2006  439255 1 [Democrat]
 7  2006  439263 1 [Democrat]
 8  2006  439304 1 [Democrat]
 9  2006  439338 1 [Democrat]
10  2006  439390 1 [Democrat]
# ... with 160,627 more rows
```

Some variables are stored as labelled values, and some are not and stored as factors. Most of this is because the latter variables were different enough in their value codings across years that summarizing them into a single numeric value was difficult.

Adding more variables

As noted, the cumulative dataset only uses key variables from each year's common content. However,common content variables can be merged in.

In R, one could use the `merge` or `left_join` functions. Here is sample code for merging with the latter. In Stata, users can use `merge 1:1`. In all cases, `year` and `case_id` uniquely identify every row in the common content, so merges should merge on year and the case identifier.

Features of the 2006 - 2018 Cumulative Dataset

Unified Variable Names

Most variables in this dataset come straight from each year's CCES. However, it renames and standardizes variable names, making them accessible in one place. Please see the rest of this guide or the Crunch dataset for a full list and description of variables.

Chosen Candidate Names and Identifiers

One addition to this cumulative dataset are variables of candidate names and identifiers that a respondent chose. In the individual year's CCES datasets, typically the response values for a vote choice question is a generic label, e.g. `Candidate1` and `Candidate2`. Then, separate variables of names and parties correspond to each `Candidate1` and `Candidate2`.

Instead, the cumulative dataset shows both the generic label *and* the chosen candidate's name, party, and identifier, which will vary across individuals.

```
select(df, year, case_id, st, matches("voted_sen"))
```

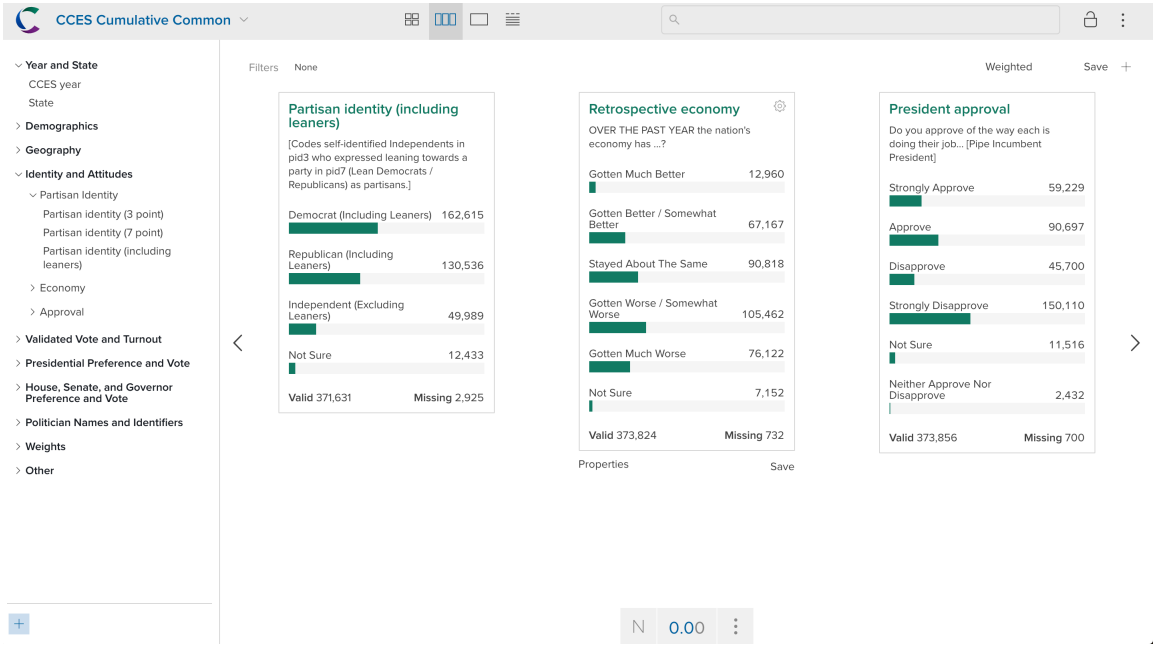
```
# A tibble: 452,755 x 7
    year case_id st    voted_sen voted_sen_party voted_sen_chosen
   <int>   <int> <chr> <fct>     <fct>           <chr>
 1  2006  439219 NC    <NA>      <NA>            <NA>
 2  2006  439224 OH    [Democra~ Democratic      Sherrod C. Brow~
 3  2006  439228 NJ    [Democra~ Democratic      Robert Menendez~
 4  2006  439237 IL    <NA>      <NA>            <NA>
 5  2006  439238 NY    [Democra~ Democratic      Hillary Rodham ~
 6  2006  439242 TX    I Did No~ <NA>            <NA>
 7  2006  439251 MN    [Republi~ Republican      Mark Kennedy (R)
 8  2006  439254 NV    [Democra~ Democratic      Jack Carter (D)
 9  2006  439255 TX    [Democra~ Democratic      Barbara Ann Rad~
10  2006  439263 MD    I Did No~ <NA>            <NA>
# ... with 452,745 more rows, and 1 more variable: voted_sen_fec <chr>
```

Crunch

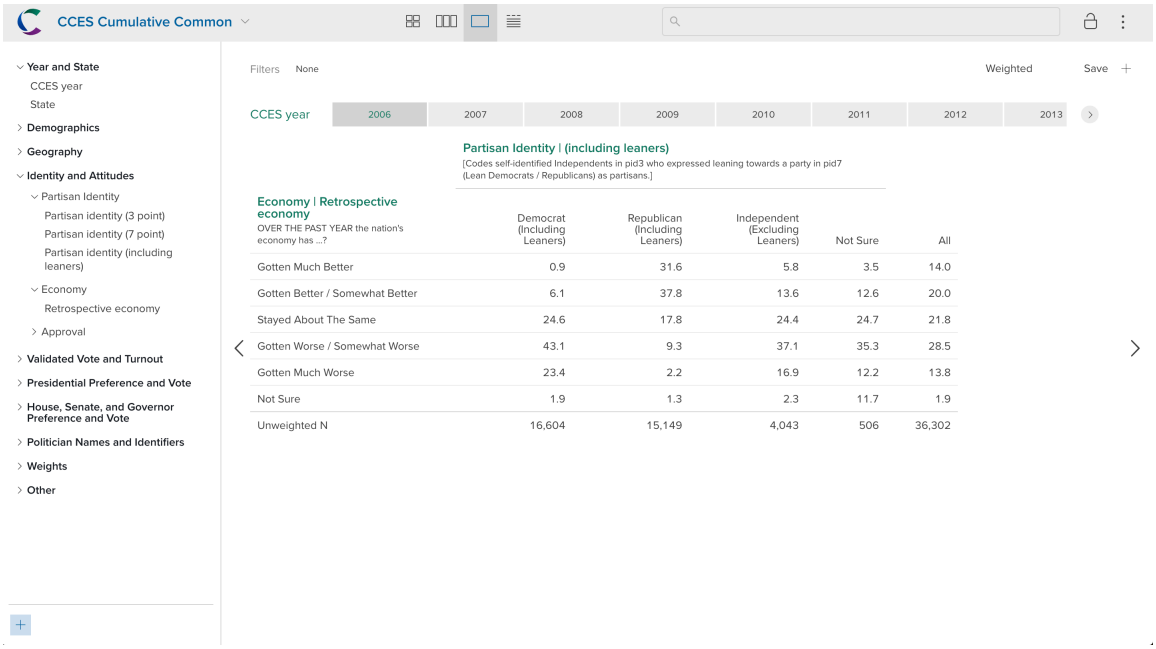
A version of the dataset is also included in Crunch, a database platform that makes it easy to view and analyze survey data either with our without any programming experience. Crunch is in beta at the time of writing.

1. Obtain Access: For View access to the dataset (free), please sign up here: https://harvard.az1.qualtrics.com/jfe/form/SV_066hQi4Eeco3Kap. For questions and more access, please contact the CCES Team.

2. Browse: Crunch offers a web GUI for quickly browsing variables:



3. Analyze: The crunch interface allows Viewers to make cross-tabs and bar graphs quickly.



Crunch datasets can also be manipulated from a R package, `crunch` https://github.com/Crunch-io/rcrunch.

Variables

The sections below provide summary more information on each variable.

- The title shows the name as used in the dataset, suitable for coding ("alias" in Crunch terminology). followed by a more descriptive name suitable for presentation ("name" in Crunch terminology).
- Question wordings, where applicable, immediately follow. Otherwise a description is provided in square brackets (`[ ]`). All square or brackets, both in the description and the response options, indicate descriptions that are summaries of what the respondent saw rather than the question verbatim.
- A tabulation of response options (or summary statistics for numeric variables) follow. Numbers are unweighted counts.
- The "Years" bullet lists the years of the CCES in which data on the variable is available at all. If a year is not listed, either the question was not asked in the year or was not incorporated in the creation of this dataset.
- Finally, the "Limitations" bullet notes some of the caveats required when interpreting this variable. As this dataset is combinations of different surveys, some year-specific details on implementation are inevitably lost. For example, for all 2016 responses "Not Asked" and "Skipped" are both coded as a `NA` (missing) to stay consistent with past years that did not make that finer distinction.

Administration

`year`: CCES year

[Year of CCES Common Content]

2,006	36,421
2,007	9,999
2,008	32,800
2,009	13,800
2,010	55,400
2,011	20,150
2,012	54,535
2,013	16,400
2,014	56,200
2,015	14,250
2,016	64,600
2,017	18,200
2,018	60,000

`starttime`: Start time

[Pre-election wave start time (up to second)]

```
             Min.                1st Qu.                 Median
```

"2006-10-07 00:02:34" "2012-10-12 14:48:08" "2014-10-20 15:49:08" Mean 3rd Qu. Max. "2014-08-30 23:43:39" "2016-10-21 08:26:27" "2018-11-05 23:27:49" NA's "118349" * Years: 2006, 2009, 2012, 2013, 2014, 2015, 2016, 2017, 2018

`tookpost`: Took post-election wave

[Whether or not the respondent took the post-election wave of the survey (in even years)]

Did Not Take Post-Election Survey	59,064
Took Post-Election Survey	300,892
(Missing)	92,799

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018 (Post-election wave only exists for even years)

Weights

`weight`: Survey weight (Year-Specific)

[weights for pre-election survey of each year]
Min. 1st Qu. Median Mean 3rd Qu. Max. 0.0000 0.4259 0.7281 1.0000 1.1741 15.0006

- Years: All of 2006-2018
- In even years, they are re-computed after vote validation has been computed and those re-computed weights are taken here when available. The weights applied to the sample (which is originally drawn from a matched sample) are constructed to make each year's respondents' pool representative of the national adult population. See the methodology section of the 2016 Guide for details.
- Limitations: Only specific to each year. Built off of the entire pre-election wave sample, but not necessarily to adjust post-election wave respondents. See `weight_post`

`weight_cumulative`: Survey weight (Cumulative)

[weight variable with simple adjustment: multiplied a constant within year to make years comparable]
Min. 1st Qu. Median Mean 3rd Qu. Max. 0.0000 0.2989 0.5562 0.9418 1.0754 24.0297

- Years: All of 2006-2018
- Limitations: Only a simple transformation of `weight`. Specifically, `weight_cumulative` is `weight` divided by the year-specific factors shown in the following table. For example, all weights in the 2016 common content are divided by about 1.97, because it has about twice as many observations as the other datasets.

Year	Observations	Factor
2006	36,421	1.11
2007	9,999	0.30
2008	32,800	1.00
2009	13,800	0.42
2010	55,400	1.69
2011	20,150	0.61
2012	54,535	1.66
2013	16,400	0.50
2014	56,200	1.71
2015	14,250	0.43
2016	64,600	1.97
2017	18,200	0.55
2018	60,000	1.83

`rvweight`: Survey weights to validated registered voters

[weights to validated registered voter population]

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     0.6     0.8     1.0     1.2    15.0  412738
```

- Years: 2018
- In 2018, YouGov computed weights after vote validation to weight to the target population of registered voters. See the methodology section of the 2018 Guide for details.
- Limitations: Only specific to each year. Built off of the entire pre-election wave sample, but not necessarily to adjust post-election wave respondents. See `rvweight_post`

`rvweight_post`: Survey weights to validated registered voters, post-election wave

[weights to validated registered voter population, post-election wave]

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     0.5     0.8     1.0     1.2    15.0  415806
```

- Years: 2018
- Limitations: Only available for some even years.

`weight_post`: Survey weight for post-election wave

[weight for post-election wave respondents. Only available for some of the even years.]

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    0.45    0.70    1.00    1.10   15.00  303170
```

- Years: 2012, 2016, 2018
- Limitations: Only available for some even years.

Geography

A series of variables for the respondent's location

- `state`: State (FIPS): [State (Imputed from input zipcode)]
- `st`: State abbreviation (FIPS): [State (Imputed from input zipcode)]
- `dist`: Congressional district number in current Congress: [Current Congressional District Number (Imputed from input zipcode)]
- `dist_up`: Congressional district number for upcoming Congress: [Upcoming Congressional District Number (Imputed from input zipcode)]
- `cd`: Congressional district in current Congress: [Current Congressional District (Imputed from input zipcode)]
- `zipcode`: Zipcode of residence: "So that we can ask you about the news and events in your area, in what zip code do you currently reside?"
- `county_fips`: County of residence: [County (Imputed from input zipcode)]

```
Observations: 452,755
Variables: 7
$ state       <chr> "California", "Pennsylvania", "Texas", "Texas", "T...
$ st          <chr> "CA", "PA", "TX", "TX", "TX", "NY", "NC", "NC", "M...
$ cd          <chr> "CA-2", "PA-5", "TX-16", "TX-19", "TX-6", "NY-28",...
```

```
$ dist        <int> 2, 5, 16, 19, 6, 28, 11, 7, 1, 17, 15, 1, 2, 6, 1,...
$ dist_up     <int> 1, 3, 16, 19, 6, 27, 11, 7, 2, 20, 12, 1, 2, 8, 1,...
$ zipcode     <chr> "95969", "16255", "79924", "79423", "76123", "1413...
$ county_fips <chr> "06007", "42031", "48141", "48303", "48439", "3606...
```

- Years: All of 2006-2018
- Limitations: Some years do not provide the variable relevant to `dist_up`, in which case the current district (`dist`) is assigned automatically. Thus, `dist_up` may not reflect district changes in off-cycle redistricting. Only residence (not registration) geographies included here; see individual years' for registration geographies.

Demographics

`gender`: Gender

"Are you male or female?"

Male	210,102
Female	242,653

- Years: All of 2006-2018

`birthyr`: Year of birth

"In what year were you born?"

Min. 1st Qu. Median Mean 3rd Qu. Max. 1900 1950 1961 1963 1977 2000

- Years: All of 2006-2018

`age`: Age

[Approximate age computed from the year of survey minus Year of Birth]

Min. 1st Qu. Median Mean 3rd Qu. Max. 18.00 36.00 51.00 49.61 62.00 109.00

- Years: All of 2006-2018

`educ`: Education

"What is the highest level of education you have completed?"

No HS	13,939
High School Graduate	125,323
Some College	112,290
2-Year	43,606
4-Year	103,011
Post-Grad	54,519
(Missing)	67

- Years: All of 2006-2018

`race`: Race

"What racial or ethnic group best describes you?"

White	337,793
Black	49,162
Hispanic	35,924
Asian	9,134
Native American	3,543
Mixed	8,919
Other	7,554
Middle Eastern	726

- Years: All of 2006-2018
- Limitations: The "Hispanic" value may undercount self-identified Hispanics. See `hispanic`

`hispanic`: Hispanic

"Are you of Spanish, Latino, or Hispanic origin or descent? [Asked if response to race is not Hispanic]"

Yes	11,113
No	321,994
(Missing)	119,648

- Years: 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018
- In years in which this question was fielded, this question supplements the `race` variable by asking those who did *not* respond "Hipsanic" in the `race` question.

`faminc`: Family Income

"Thinking back over the last year, what was your family's annual income? [Brackets coarsened]"

Less than 10k	18,780
10k - 20k	32,995
20k - 30k	46,019
30k - 40k	46,728
40k - 50k	41,768
50k - 60k	40,863
60k - 70k	30,063
70k - 80k	32,459
80k - 100k	37,809
100k - 120k	27,566
120k - 150k	22,108
150k+	25,075
Prefer not to say	48,955
Skipped	12
(Missing)	1,555

- Years: All of 2006-2018

- Limitations: The income brackets provided changed slightly over time. The brackets in this cumulative dataset coarsens certain brackets, losing some granularity. In particular, from 2011-2016, respondents answering "over 150k" were asked a follow-up question to select one of several brackets above 150k. Here, these are top-coded and only labelled as "over 150k."
- The 2009 CCES did not have an option for 60-70k.

`marstat`: Marital Status

"What is your marital status?"

Married	250,629
Separated	7,606
Divorced	49,651
Widowed	21,287
Single / Never Married	101,416
Domestic Partnership	20,611
(Missing)	1,555

- Years: All of 2006-2018
- The option "Single" was used till 2016, which was then replaced by "Never Married" in 2017 and 2018.
- The option "Domestic Partnership" was used till 2016, which was then replaced by "Domestic / Civl Partnership" in 2017 and 2018.

Validations

Observations in even years include (or will include) indicators for validated voting, which means that YouGov has matched survey respondents' personal identifiable information to public voter files, which in turn officially record whether a person has voted or not. Validation is often completed in the summer following the election; 2018 validation data is not available as of March. For more information, see Ansolabehere and Hersh (2012).

`vv_regstatus`: Validated registration status

[Validation results. Missing if validation was not conducted in the year. Categories are aggregated. Both Matched-not registered and unmatched are labeled as a no record.]

Active	218,373
No Record Of Registration	61,861
Unregistered	15,869
Dropped	6,607
Inactive	3,565
Multiple Appearances	1,600
(Missing)	144,880

- Years: 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: Collapses some response options

`vv_party_gen`: Validated registered party

[Validation results. Only available for some staets and years]

Unknown	68,895
No Record Of Party Registration	60,890
Democratic Party	37,600
Republican Party	29,494
No Party Affiliation	13,874
Declined To State	2,376
Other	1,635
Independent Party	1,511
Liberatarian Party	537
Green Party	265
Cns	44
Constitution Party	38
Reform Party	11
Wor	9
Socialist Party	5
(Missing)	235,571

- Years: 2012, 2014, 2016, 2018
- Limitations: Not available for some even years

`vv_party_prm`: Validated registered Primary party

[Validation results. Only available for some staets and years]

No Record Of Party Registration	157,120
Republican Party	14,486
Democratic Party	12,783
No Party Affiliation	16
Liberatarian Party	11
Other	8
Green Party	4
(Missing)	268,327

- Years: 2012, 2014, 2016, 2018
- Limitations: Not available for some even years

Turnout

`vv_turnout_gvm`: Validated turnout General Election

[Validation results. All vote methods (polling, mail, early, unknown, etc..) are aggregated as a vote.]

Voted	202,966
No Record Of Voting	129,019
No Voter File	1,733
(Missing)	119,037

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: Collapses most response options. For example, the particular voting method is collapsed into one category, even though `gvm` stands for General Election voting *method*. Also, the result of not matching to a voter file is collapsed with the result of matching to a voter file and having no indication of turning out to vote. The distinction is unclear in earlier years, and is thus collapsed for all years here. For finer distinctions, see the individual year's CCES.

`vv_turnout_pvm`: Validated turnout Primary Election (Congressional)

[Validation results]

No Record Of Voting	185,927
Voted	96,435
No Voter File	1,363
(Missing)	169,030

- Years: 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: See `vv_turnout_gvm`

Identity and Attitudes

Partisan Identity

`pid3`: Partisan identity (3 point)

"Generally speaking, do you think of yourself as a . . . ?"

Democrat	160,637
Republican	118,907
Independent	126,270
Other	17,975
Not Sure	20,012
(Missing)	8,954

- Years: All of 2006-2018
- Limitations: Response options offer slightly by year. For example, the `Not Sure` option is not a response option in years 2006 and 2010. Open-text responses not included. 2010 values are from the post-election wave.

`pid7`: Partisan identity (7 point)

[Based on branching from Partisan Identity question]

Strong Democrat	107,733
Not Very Strong Democrat	54,838
Lean Democrat	45,411
Independent	60,946
Lean Republican	47,739
Not Very Strong Republican	43,810
Strong Republican	75,782
Not Sure	13,480
(Missing)	3,016

- Years: All of 2006-2018
- Limitations: See `pid3`

`pid3_leaner`: Partisan identity (including leaners)

[Codes self-identified Independents in pid3 who expressed leaning towards a party in pid7 (Lean Democrats / Republicans) as partisans.]

Democrat (Including Leaners)	207,982
Republican (Including Leaners)	167,331
Independent (Excluding Leaners)	60,946
Not Sure	13,480
(Missing)	3,016

- Years: All of 2006-2018
- Limitations: See `pid3`

`ideo5`: Ideology (5 point)

"In general, how would you describe your own political viewpoint?"

Very Liberal	40,142
Liberal	79,591
Moderate	141,633
Conservative	105,268
Very Conservative	52,243
Not Sure	32,104
(Missing)	1,774

- Years: All of 2006-2018

Economy

`economy_retro`: Retrospective economy

"OVER THE PAST YEAR the nation's economy has . . . ?"

Gotten Much Better	29,702
Gotten Better / Somewhat Better	100,766
Stayed About The Same	118,229
Gotten Worse / Somewhat Worse	114,965
Gotten Much Worse	77,949
Not Sure	10,294
(Missing)	850

- Years: All of 2006-2018
- Limitations: Response options varies by year. Some are collapsed into one category (e.g. `Gotten Better`, presented in some years, and `Gotten Somewhat Better`, presented in other years, are collapsed into `Gotten Better / Somewhat Better`). Some are left as is. For example, `Not Sure` was not an option in 2009.

News Interest

`newsint`: News Interest

"Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs .."

Most Of The Time	225,200
Some Of The Time	105,337
Only Now And Then	50,141
Hardly At All	24,621
Don't Know	10,449
(Missing)	37,007

- Years: 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018
- Limitations: Not asked in 2006. Similar questions about watching TV news was asked in 2006, but not included in this cumulative file.

Approval

`approval_pres`: President approval

"Do you approve of the way each is doing their job... [Pipe Incumbent President]"

Strongly Approve	92,872
Approve / Somewhat Approve	104,677
Disapprove / Somewhat Disapprove	47,005
Strongly Disapprove	194,516
Never Heard / Not Sure	12,507
Neither Approve Nor Disapprove	443
(Missing)	735

- Years: All of 2006-2018
- Limitations: `Neither approve nor disapprove` only included in 2007.
- This question is asked in a grid format, along with Governors, Congress, and Courts.

`approval_rep`: House Representative approval

"Do you approve of the way each is doing their job... [Pipe Incumbent Representative's Name]"

Strongly Approve	65,102
Approve / Somewhat Approve	142,503
Disapprove / Somewhat Disapprove	79,595
Strongly Disapprove	70,663
Never Heard / Not Sure	85,772
Neither Approve Nor Disapprove	1,798
(Missing)	7,322

- Years: All of 2006-2018
- Limitations: `Neither approve nor disapprove` only included in 2007.
- This question is asked in a grid format, along with Senators (`approval_sen1`, `approval_sen2`).
- To see who [Representative] refers to for a particular respondent, see `rep_inc` (incumbent identifier in `rep_icpsr`)

`approval_sen1`: Senator 1 approval

"Do you approve of the way each is doing their job... [Pipe Incumbent Senator 1's Name]"

Strongly Approve	58,761
Approve / Somewhat Approve	144,415
Disapprove / Somewhat Disapprove	90,846
Strongly Disapprove	89,017
Never Heard / Not Sure	63,783
Neither Approve Nor Disapprove	1,413
(Missing)	4,520

- Years: All of 2006-2018
- Limitations: : Response options varies by year. Some are collapsed into one category (e.g. `Approve`, presented in some years, and `Somewhat Approve`, presented in other years, are

collapsed into `Approve / Somewhat Approve`). `Neither approve nor disapprove` only included in 2007.
- To see who [Senator 1] refers to for a particular respondent, see `sen1_inc` (incumbent identifier in `sen1_icpsr`)

`approval_sen2`: Senator 2 approval

"Do you approve of the way each is doing their job... [Pipe Incumbent Senator 2's Name]"

Strongly Approve	63,135
Approve / Somewhat Approve	139,337
Disapprove / Somewhat Disapprove	88,539
Strongly Disapprove	89,232
Never Heard / Not Sure	66,076
Neither Approve Nor Disapprove	1,158
(Missing)	5,278

- See `approval_sen2`

`approval_gov`: Governor approval

"Do you approve of the way each is doing their job... Governor of [Pipe State]"

Strongly Approve	67,404
Approve / Somewhat Approve	139,940
Disapprove / Somewhat Disapprove	84,271
Strongly Disapprove	117,282
Never Heard / Not Sure	40,318
Neither Approve Nor Disapprove	1,414
(Missing)	2,126

- Years: All of 2006-2018
- Limitations: See `approval_pres`
- To see who the Governor refers to for a particular respondent, see `gov_inc` (incumbent identifier in `gov_fec`, if applicable).

Presidential Vote

A note on `intent` and `voted` In this dataset we make the distinction between "intent" / "preference" vs. "voted" / "vote choice". "Intent" (or "preference") refers to the response to the prospective question of the sort "who would you vote for?" in the *pre-election* wave. "Vote choice" refers to the response to the retrospective question of the sort "in the election this November, who did you vote for?" Response to the vote choice questions coalesces both *post-election* wave responses (the bulk of the responses) and pre-election respondents who reported having already voted early.

`intent_pres_08`: 2008 President preference (before voting)

"For which candidate for President of the United States would you vote?"

John McCain	13,322
Barack Obama	12,897
Ron Paul	535
Ralph Nader	209
Bob Barr	258
Cynthia McKinney	74
Other	352
I Won't Vote In This Election	851
I'm Not Sure	1,697
(Missing)	422,560

- Years: 2008

`intent_pres_12`: 2012 President preference (before voting)

"In the race for President of the United States, who do you prefer?"

Mitt Romney (Republican)	20,738
Barack Obama (Democratic)	24,401
Other	1,781
I Will Not Vote In This Election	1,467
I'm Not Sure	3,856
(Missing)	400,512

- Years: 2012

`intent_pres_16`: 2016 President preference (before voting)

"Which candidate did you prefer for President of the United States?"

Donald Trump (Republican)	19,227
Hillary Clinton (Democrat)	27,502
Gary Johnson (Libertarian)	3,145
Jill Stein (Green)	1,400
Other	1,880
I Won't Vote In This Election	3,312
I'm Not Sure	6,536
(Missing)	389,753

- Years: 2016

`voted_pres_08`: 2008 President vote choice (after voting)

"2008: For which candidate for President of the United States did you vote? [see guide for wording in all years]"

Barack Obama (Democratic)	73,986
John McCain (Republican)	68,398
Someone Else	4,204
Did Not Vote	18,227
Don't Recall	1,787
(Missing)	286,153

- Years: 2008, 2009, 2010, 2011, 2012
- Limitations: Response options offer slightly by year; some are collapsed into one.

`voted_pres_12`: 2012 President vote choice (after voting)

"2012: For whom did you vote for President of the United States? 2016: In 2012, who did you vote for in the election for President? [see guide for wording in all years]"

Barack Obama	82,681
Mitt Romney	64,956
Other / Someone Else	5,890
Did Not Vote	2,758
Not Sure / Don't Recall	1,990
I Did Not Vote In This Race	81
(Missing)	294,399

- Years: 2012, 2013, 2014, 2015, 2016
- Limitations: Response options offer slightly by year; some are collapsed into one.
- This variable coalesces two variables: Either the response to the early vote question in the pre-election wave if the respondent indicates they have already voted, or if not, the response in the post-election wave.

`voted_pres_16`: 2016 President vote choice (after voting)

"2017: In the election for U.S. President, who did you vote for? [If reported voting] 2016: For whom did you vote for President of the United States? [Post-election]"

Donald Trump	43,891
Hilary Clinton	51,342
Other / Someone Else	10,091
Did Not Vote	627
Not Sure / Don't Recall	527
(Missing)	346,277

- Years: 2016, 2017, 2018

- This variable coalesces two variables in the CCES: Either the response to the early vote question in the pre-election wave if the respondent indicates they have already voted, or if not, the response in the post-election wave.

House, Senate and Governor Voting

Preference

`intent_rep`: House preference (before voting)

"In the general election for U.S. House of Representatives in your area, who do you prefer?"

[Democrat / Candidate 1]	128,231
[Republican / Candidate 2]	115,292
[Other / Candidate 3]	4,401
$HouseCand4Name ($HouseCand4Party)	37
Other	2,259
I'm Not Sure	70,460
No One	19,235
$HouseCand5Name ($HouseCand5Party)	23
I Won't Vote In This Election	2,269
$HouseCand6Name ($HouseCand6Party)	41
$HouseCand7Name ($HouseCand7Party)	20
$HouseCand8Name ($HouseCand8Party)	14
$HouseCand9Name ($HouseCand9Party)	1
$HouseCand10Name ($HouseCand10Party)	1
$HouseCand11Name ($HouseCand11Party)	3
(Missing)	110,468

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: Only available for even years. The third party candidate is not specified for early years. The fourth candidate and below are not shown for most years. Response options differ by year.
- Note that it is not always the case that 1 is a Democrat and 2 is a Republican. When two Democrats are on the general ballot (e.g. in top-two primary states like California), both candidates are Democrats. Use `intent_rep_party` to see the party affiliation of the chosen candidate.
- Note that for each respondent, a name (and party affiliation) is shown in place of the square bracket values. To see the name of the candidate chosen, see `intent_rep_chosen`.
- `[Other / Candidate 3]` refers to the third option presented, whereas `Other` refers to the unnamed choice after all numbered candidates.

`intent_sen`: Senate preference (before voting)

"In the race for U.S. Senator in your state, who do you prefer?"

[Democrat / Candidate 1]	97,220
[Republican / Candidate 2]	82,433
[Other / Candidate 3]	4,477
$SenCand4Name ($SenCand4Party)	19
Other	1,713
I'm Not Sure	38,112
No One	12,419
I Won't Vote In This Election	1,145
(Missing)	215,217

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: See `intente_rep`. When both senate seats are up for re-election in the same year, only responses to the first senate seat is incorporated. For the second senate seat, see individual year's CCES.
- See `intent_sen_party` for the party affiliation of the chosen candidate.

`intent_gov`: Governor preference (before voting)

"In the race for Governor in your state, who do you prefer?"

[Democrat / Candidate 1]	74,561
[Republican / Candidate 2]	66,292
[Other / Candidate 3]	4,055
Other	1,390
I'm Not Sure	24,296
No One	7,991
I Won't Vote In This Election	466
(Missing)	273,704

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- Limitations: See `intente_rep`. For governor elections in odd years, see individual year's CCES.
- See `intent_gov_party` for the party affiliation of the chosen candidate.

Vote Choice

`voted_rep`: House vote choice (after voting)

"For whom did you vote for U.S. House?"

[Democrat / Candidate 1]	117,581
[Republican / Candidate 2]	111,255
[Other / Candidate 3]	2,786
$HouseCand4Name ($HouseCand4Party)	27
Other	3,120
I Did Not Vote In This Race	12,535
$HouseCand5Name ($HouseCand5Party)	24
Not Sure	4,493
$HouseCand6Name ($HouseCand6Party)	39
$HouseCand7Name ($HouseCand7Party)	15
$HouseCand8Name ($HouseCand8Party)	16
$HouseCand9Name ($HouseCand9Party)	2
$HouseCand10Name ($HouseCand10Party)	2
$HouseCand11Name ($HouseCand11Party)	3
(Missing)	200,857

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- This variable coalesces two variables in the CCES for years 2012 and onward: Either the response to the early vote question in the pre-election wave if the respondent indicates they have already voted, or if not, the response in the post-election wave.
- Note that it is not always the case that 1 is a Democrat and 2 is a Republican. When two Democrats are on the general ballot (e.g. in top-two primary states like California), both candidates are Democrats. Use `voted_rep_party` for party affiliation
- See `voted_rep_party` for party affiliation.

`voted_sen`: Senate vote choice (after voting)

"For whom did you vote for U.S. Senator?"

[Democrat / Candidate 1]	86,668
[Republican / Candidate 2]	77,569
[Other / Candidate 3]	2,974
Other	1,967
Not Sure	2,094
$SenCand4Name ($SenCand4Party)	11
I Did Not Vote In This Race	4,789
(Missing)	276,683

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- This variable coalesces two variables in the CCES for years 2012 and onward: Either the response to the early vote question in the pre-election wave if the respondent indicates they have already voted, or if not, the response in the post-election wave.
- See `voted_sen_party` for party affiliation.
- Senate Special elections where both senate seats are up for election is often recorded as different columns in the year-specific CCES, but these are not collected in the cumulative.

`voted_gov`: Governor vote choice (after voting)

"For whom did you vote for Governor?"

[Democrat / Candidate 1]	68,445
[Republican / Candidate 2]	63,434
[Other / Candidate 3]	2,800
Other	1,817
I Did Not Vote In This Race	10,116
Not Sure	1,091
(Missing)	305,052

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- This variable coalesces two variables in the CCES for years 2012 and onward: Either the response to the early vote question in the pre-election wave if the respondent indicates they have already voted, or if not, the response in the post-election wave.
- See `voted_gov_party` for party affiliation.

Metadata and Identifiers

Identifiers

The case identifier `case_id` is unique within the year and is identical to the case identifiers in the individual year's CCES. It should be used in conjunction with `year` for a unique identifier for the whole dataset. Some individuals across years may be the same YouGov panel respondent with different identifiers; for example the 2007 CCES draws from the 2006 CCES respondents.

```
Observations: 452,755
Variables: 2
$ year    <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, ...
$ case_id <int> 439219, 439224, 439228, 439237, 439238, 439242, 439251...
```

Current Representatives

Name and Party

The four names in the three offices are representatives of the respondent *at the time of the survey*. Names are printed as shown, and similarly parties are shown if the particular year's CCES did not show party. For example, Senator Shelby is presented as `Richard Craig Shelby`, `Richard C. Shelby (R)`, `Richard Shelby (R)`, `Richard C. Shelby (R)`, depending on the year. Party names are abbreviated down to initials (`D` for Democrat, `R` for Republican, `I` for Independent) in this dataset.

```
Observations: 452,755
Variables: 4
$ rep_current  <chr> "Patrick T. McHenry (R)", "Michael R. Turner (R)"...
$ sen1_current <chr> "Elizabeth Dole (R)", "Mike DeWine (R)", "Robert ...
$ sen2_current <chr> "Richard Burr (R)", "George V. Voinovich (R)", "F...
$ gov_current  <chr> "Michael Easley (D)", "Bob Taft (R)", "Jon Corzin...
```

Incumbent Identifiers

Unique identifiers (ICPSR / Nominate for Congress, FEC for Governor) for the current representatives. Identifiers are not part of the individual year's CCES. Instead, I attempt to merge in these identifiers through a series of name and district merges.

The matching of identifiers to respondent occurs through matching by district, by district and last name, or both:

- For House representatives, we join on `cong`, `st`, and `dist` to a NOMINATE database that only consists of unique observations according to the key. For duplicates with regards to these three variables (e.g. in the rare case where a new representative comes into office mid-session), we match on `cong`, `st`, `dist` and last name.
- For Senators, we join entirely on `cong`, `st`, and last name
- For Governors, we join only on `st` and last name. In this period, there are no two governors in the same state that share the same last name.

```
Observations: 452,755
Variables: 4
$ rep_icpsr  <dbl> 20522, 20342, 29132, 29911, 29380, 20531, 29126, 29...
$ sen1_icpsr <dbl> 40303, 15020, 29373, 15021, 14858, 49306, 40101, 15...
$ sen2_icpsr <dbl> 29548, 49903, 14914, 40502, 40105, 40305, 40302, 29...
$ gov_fec    <chr> "NC5998", NA, "NJ6395", "IL7", NA, "TX3156", "MN472...
```

- Years: All of 2006-2018
- Limitations: Please note there may be some incorrect merges, especially for nontraditional names and representatives who were elected in special elections and may not be in some datasets.

The unique identifiers can be used to join with other databases to append additional information such as committee membership and ideology scores, such as

> Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2017). Voteview: Congressional Roll-Call Votes Database. https://voteview.com/

Candidates

The text responses that the respondent chose in each of the `intent_` / `voted_` questions, if the respondent was a candidate. For example, respondent with `case_id = 163051575` in the 2012 CCES chose the first option in the House representative preference question. `intent_rep_chosen` shows that for this particular respondent, the first option was `Maxine Waters (Democrat)` who has a FEC Identifier of `H4CA23011`.

```
df %>%
  filter(year == 2012, st == "CA", dist_up == 43) %>%
  select(matches("intent_rep"))
```

```
# A tibble: 91 x 4
   intent_rep              intent_rep_party intent_rep_chos~ intent_rep_fec
   <fct>                   <fct>            <chr>            <chr>
 1 [Democrat / Candidate ~ Democratic       Maxine Waters (~ H4CA23011
 2 I'm Not Sure            <NA>             <NA>             <NA>
 3 No One                  <NA>             <NA>             <NA>
 4 [Democrat / Candidate ~ Democratic       Maxine Waters (~ H4CA23011
 5 [Republican / Candidat~ Democratic       Bob Flores (D)   H2CA43385
 6 I'm Not Sure            <NA>             <NA>             <NA>
 7 Other                   <NA>             <NA>             <NA>
 8 [Republican / Candidat~ Democratic       Bob Flores (D)   H2CA43385
 9 [Republican / Candidat~ Democratic       Bob Flores (D)   H2CA43385
10 [Democrat / Candidate ~ Democratic       Maxine Waters (~ H4CA23011
# ... with 81 more rows
```

The name and party are those as provided in the CCES datasets (e.g. in the form `HouseCand1Name`). The FEC ID is not part of the CCES but joined in this dataset.

For all three offices, the matching generally occurs by `year`, `st`, `dist_up` (not `dist`, because `dist_up`, refers to the district of the upcoming session) and `party`. `party` is the party affiliation as indicated in the CCES. For years 2008 and 2010, the first option is automatically labelled as a Democrat and the second option as a Republican, although these may be inaccurate at times.

The FEC database runs up until 2014 (thus more recent candidates do *not* get a FEC ID) originates from

> Bonica, Adam , 2015, "Database on Ideology, Money in Politics, and Elections (DIME)", doi:10.7910/DVN/O5PX0B, Harvard Dataverse, V2

which helpfully includes candidates office sought, district (for House members), party affiliation, and cycle in which the candidate filed. The variable `cycle` in Bonica's data is used to join on the CCES dataset's `year` variable.

Only candidates who are unique within the district and party are considered for the first join. However, many candidates are not unique within the district-party, as many co-partisans may file in the same district. The second matching process thus considers the full name of the candidate listed in the CCES and the candidates in the FEC database. *Within* the subset of year, district, and party, a Jaro-Winker string distance (that ranges from 0 to 1) is computed for both last name and the first name - middle name. If the sum of the two string distances are more than 0.2 for all possible combinations, no match is returned. If there is a unique combination that achieves a unique minimum that is below 0.2, that combination is declared a match. If there are multiple matches with the same minimum string distance, one is randomly chosen.

Chosen

```
Observations: 452,755
Variables: 6
$ intent_rep_chosen <chr> "Richard C. Carsner (D)", "Stephanie Studeba...
$ intent_sen_chosen <chr> NA, "Sherrod C. Brown (D)", "Robert Menendez...
$ intent_gov_chosen <chr> NA, "Ted Strickland (D)", NA, "Rod Blagojevi...
$ voted_rep_chosen  <chr> "Richard C. Carsner (D)", "Stephanie Studeba...
$ voted_sen_chosen  <chr> NA, "Sherrod C. Brown (D)", "Robert Menendez...
$ voted_gov_chosen  <chr> NA, "Ted Strickland (D)", NA, "Rod Blagojevi...
```

- Years: 2006, 2008, 2010, 2012, 2014, 2016, 2018
- Early years may mislabel the candidate's party, especially when the two candidates are of the same party (as in top-two primary states)

Candidate Identifiers

```
Observations: 452,755
Variables: 6
$ intent_rep_fec <chr> "H6NC10141", "H6OH03142", "H0NJ01066", "H8IL090...
$ intent_sen_fec <chr> NA, "S6OH00163", "S6NJ00289", NA, NA, NA, "S6MN...
$ intent_gov_fec <chr> NA, "OH19691", NA, "IL7", "NY19490", NA, "MN472...
$ voted_rep_fec  <chr> "H6NC10141", "H6OH03142", "H0NJ01066", "H8IL090...
$ voted_sen_fec  <chr> NA, "S6OH00163", "S6NJ00289", NA, "S0NY00188", ...
$ voted_gov_fec  <chr> NA, "OH19691", NA, "IL7", "NY19490", NA, "MN472...
```

- Years: 2006, 2008, 2010, 2012, 2014, 2016
- Limitations: Matching may be inaccurate (see previous section on matching methodology). In particular, a lack of a FEC ID may either indicate a failure of the matching procedure, or that the candidate in question did not file under the FEC. The match rate in the current procedure is upwards of 80 percent in the current procedure.