

# Package ‘fivethirtyeight’

May 8, 2026

**Title** Data and Code Behind the Stories and Interactives at  
‘FiveThirtyEight’

**Description** Datasets and code published by the data journalism website  
‘FiveThirtyEight’ available at <https://github.com/fivethirtyeight/data>.  
Note that while we received guidance from editors at ‘FiveThirtyEight’, this  
package is not officially published by ‘FiveThirtyEight’.

**Version** 0.6.2

**Maintainer** Albert Y. Kim <[albert.y.s.kim@gmail.com](mailto:albert.y.s.kim@gmail.com)>

**Depends** R (>= 3.2.4)

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**URL** <https://github.com/rudeboybert/fivethirtyeight>

**BugReports** <https://github.com/rudeboybert/fivethirtyeight/issues>

**RoxygenNote** 7.1.1

**Suggests** ggplot2, dplyr, tidyr (>= 1.0.0), readr, tibble, stringr,  
broom, knitr, rmarkdown, patchwork, fivethirtyeightdata

**Additional\_repositories** <https://fivethirtyeightdata.github.io/drat>

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Albert Y. Kim [aut, cre] (ORCID:  
<<https://orcid.org/0000-0001-7824-306X>>),  
Chester Ismay [aut] (ORCID: <<https://orcid.org/0000-0003-2820-2547>>),  
Jennifer Chunn [aut],  
Meredith Manley [ctb] (ORCID: <<https://orcid.org/0000-0001-7707-0654>>),  
Maggie Shea [ctb],  
Starry Yujia Zhou [ctb],  
Andrew Flowers [ctb],  
Jonathan Bouchet [ctb],  
G. Elliott Morris [ctb],

Adam Spannbauer [ctb],  
 Pradeep Adhokshaja [ctb],  
 Olivia Barrows [ctb],  
 Jojo Miller [ctb],  
 Jayla Nakayama [ctb],  
 Ben Baumer [ctb] (ORCID: <<https://orcid.org/0000-0002-3279-0516>>),  
 Rana Gahwagy [ctb] (ORCID: <<https://orcid.org/0000-0002-6331-2840>>),  
 Natalia Iannucci [ctb] (ORCID: <<https://orcid.org/0000-0001-5663-1731>>),  
 Marium Tapal [ctb] (ORCID: <<https://orcid.org/0000-0001-5093-6462>>),  
 Irene Ryan [ctb],  
 Alina Barylsky [ctb],  
 Danica Miguel [ctb],  
 Sunni Raleigh [ctb],  
 Anna Ballou [ctb],  
 Jane Bang [ctb],  
 Jordan Moody [ctb],  
 Kara Van Allen [ctb],  
 Jessica Keast [ctb],  
 Lizette Carpenter [ctb],  
 Fatima Keita [ctb]

### Repository CRAN

**Date/Publication** 2021-10-07 13:40:02 UTC

## Contents

ahca_polls . . . . .	5
airline_safety . . . . .	6
antiquities_act . . . . .	7
august_senate_polls . . . . .	8
avengers . . . . .	9
bachelorette . . . . .	10
bad_drivers . . . . .	11
bechdel . . . . .	12
biopics . . . . .	13
bob_ross . . . . .	14
cabinet_turnover . . . . .	17
candy_rankings . . . . .	17
cand_events_20150114 . . . . .	18
cand_events_20150130 . . . . .	19
cand_state_20150114 . . . . .	20
cand_state_20150130 . . . . .	21
chess_transfers . . . . .	21
classic_rock_raw_data . . . . .	22
classic_rock_song_list . . . . .	23
college_all_ages . . . . .	23
college_grad_students . . . . .	24
college_recent_grads . . . . .	25

comma_survey . . . . .	27
congress_age . . . . .	28
cousin_marriage . . . . .	29
daily_show_guests . . . . .	29
datasets_master . . . . .	30
democratic_bench . . . . .	30
dem_candidates . . . . .	31
drinks . . . . .	34
drug_use . . . . .	35
elasticity_by_district . . . . .	37
elasticity_by_state . . . . .	38
elo_blatter . . . . .	38
endorsements . . . . .	39
endorsements_2020 . . . . .	40
fandango . . . . .	41
fifa_audience . . . . .	42
fight_songs . . . . .	43
fivethirtyeight . . . . .	44
flying . . . . .	44
food_world_cup . . . . .	46
forecast_results_2018 . . . . .	48
foul_balls . . . . .	49
generic_polllist . . . . .	50
generic_topline . . . . .	51
google_trends . . . . .	52
governor_national_forecast . . . . .	53
governor_state_forecast . . . . .	54
hate_crimes . . . . .	55
hiphop_cand_lyrics . . . . .	56
hist_ncaa_bball_casts . . . . .	56
hist_senate_preds . . . . .	57
house_national_forecast . . . . .	58
impeachment_polls . . . . .	59
librarians . . . . .	60
love_actually_adj . . . . .	61
love_actually_appearance . . . . .	62
mad_men . . . . .	63
male_flight_attend . . . . .	64
masculinity_survey . . . . .	64
mediacloud_hurricanes . . . . .	66
mediacloud_online_news . . . . .	67
mediacloud_states . . . . .	67
mediacloud_trump . . . . .	68
media_mentions_2020 . . . . .	69
mlb_as_play_talent . . . . .	70
mlb_as_team_talent . . . . .	71
mueller_approval_polls . . . . .	72
murder_2015_final . . . . .	73

murder_2016_prelim . . . . .	73
nba_draft_2015 . . . . .	74
nba_draymond . . . . .	75
nba_elo . . . . .	75
nba_tattoos . . . . .	76
ncaa_w_bball_tourney . . . . .	77
nftix_div_avgprice . . . . .	78
nftix_usa_avg . . . . .	78
nflwr_aging_curve . . . . .	79
nflwr_hist . . . . .	79
nfl_fandom_google . . . . .	80
nfl_fandom_surveymonkey . . . . .	81
nfl_fav_team . . . . .	83
nfl_suspensions . . . . .	84
nutrition_pvalues . . . . .	84
partisan_lean_district . . . . .	85
partisan_lean_state . . . . .	86
police_deaths . . . . .	87
police_killings . . . . .	87
police_locals . . . . .	89
pres_2016_trail . . . . .	90
pres_commencement . . . . .	91
pulitzer . . . . .	91
riddler_castles . . . . .	92
riddler_castles2 . . . . .	93
riddler_pick_lowest . . . . .	94
russia_investigation . . . . .	95
sandy_311 . . . . .	96
san_andreas . . . . .	98
senate_national_forecast . . . . .	99
senate_polls . . . . .	100
senate_seat_forecast . . . . .	100
spi_global_rankings . . . . .	101
state_info . . . . .	102
state_of_the_state . . . . .	103
steak_survey . . . . .	104
tarantino . . . . .	105
tennis_events_time . . . . .	106
tennis_players_time . . . . .	106
tennis_serve_time . . . . .	107
tenth_circuit . . . . .	108
trumpworld_issues . . . . .	109
trumpworld_polls . . . . .	110
trump_approval_poll . . . . .	112
trump_approval_trend . . . . .	113
trump_lawsuits . . . . .	114
trump_news . . . . .	115
trump_twitter . . . . .	116

tv_hurricanes . . . . .	116
tv_hurricanes_by_network . . . . .	117
tv_states . . . . .	118
undefeated . . . . .	119
unisex_names . . . . .	119
US_births_1994_2003 . . . . .	120
US_births_2000_2014 . . . . .	121
weather_check . . . . .	121
wwc_2019 . . . . .	122

<b>Index</b>	<b>125</b>
--------------	------------

---

ahca_polls	<i>American Health Care Act Polls</i>
------------	---------------------------------------

---

## Description

The raw data behind the story "Why The GOP Is So Hell-Bent On Passing An Unpopular Health Care Bill" <https://fivethirtyeight.com/features/why-the-gop-is-so-hell-bent-on-passing-an-unpopular-h>

## Usage

ahca\_polls

## Format

A data frame with 15 rows representing polls and 7 variables:

**start** Start date of the poll.

**end** End date of the poll.

**pollster** The entity that conducts and collects information from the poll.

**favor** The number of affirmative responses to the question at the pollster.

**oppose** The number of negative responses to the question at the pollster.

**url** The website associated with the polling question.

**text** The polling question asked at the pollster.

## Source

See <https://github.com/fivethirtyeight/data/blob/master/ahca-polls/README.md>

## Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
ahca_polls_tidy <- ahca_polls %>%
  pivot_longer(-c(start, end, pollster, text, url), names_to = "opinion", values_to = "count")
```

---

airline_safety	<i>Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?</i>
----------------	--

---

### Description

The raw data behind the story "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?" <https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had->

### Usage

```
airline_safety
```

### Format

A data frame with 56 rows representing airlines and 9 variables:

**airline** airline  
**incl\_reg\_subsidiaries** indicates that regional subsidiaries are included  
**avail\_seat\_km\_per\_week** available seat kilometers flown every week  
**incidents\_85\_99** Total number of incidents, 1985-1999  
**fatal\_accidents\_85\_99** Total number of fatal accidents, 1985-1999  
**fatalities\_85\_99** Total number of fatalities, 1985-1999  
**incidents\_00\_14** Total number of incidents, 2000-2014  
**fatal\_accidents\_00\_14** Total number of fatal accidents, 2000-2014  
**fatalities\_00\_14** Total number of fatalities, 2000-2014

### Source

Aviation Safety Network <https://aviation-safety.net>.

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
airline_safety_tidy <- airline_safety %>%
  pivot_longer(-c(airline, incl_reg_subsidiaries, avail_seat_km_per_week),
    names_to = "type", values_to = "count") %>%
  mutate(
    period = str_sub(type, start=-5),
    period = str_replace_all(period, "_", "-"),
    type = str_sub(type, end=-7)
  )
```

---

antiquities_act	<i>Trump Might Be The First President To Scrap A National Monument</i>
-----------------	--

---

## Description

The raw data behind the story "Trump Might Be The First President To Scrap A National Monument" <https://fivethirtyeight.com/features/trump-might-be-the-first-president-to-scrap-a-national-m>

## Usage

```
antiquities_act
```

## Format

A data frame with 344 rows representing acts and 9 variables (Note that 7 of the original rows failed to parse and are omitted here):

**current\_name** Current name of piece of land designated under the Antiquities Act

**states** State(s) or territory where land is located

**original\_name** If included, original name of piece of land designated under the Antiquities Act

**current\_agency** Current land management agency. NPS = National Parks Service, BLM = Bureau of Land Management, USFS = US Forest Service, FWS = US Fish and Wildlife Service, NOAA = National Oceanic and National Oceanic and Atmospheric Administration

**action** Type of action taken on land

**date** Date of action

**year** Year of action

**pres\_or\_congress** President or congress that issued action

**acres\_affected** Acres affected by action. Note that total current acreage is not included. National monuments that cover ocean are listed in square miles.

## Source

National Parks Conservation Association <https://www.npca.org/> and National Parks Service Archeology Program <https://www.nps.gov/history/archeology/sites/antiquities/MonumentsList.htm>

---

august\_senate\_polls    *How Much Trouble Is Ted Cruz Really In?*

---

### Description

The raw data behind the story "How Much Trouble Is Ted Cruz Really In?" <https://fivethirtyeight.com/features/how-much-trouble-is-ted-cruz-really-in/>.

### Usage

august\_senate\_polls

### Format

A data frame with 594 rows representing senate polls, and 11 variables:

**cycle** the election year

**state** the state of the poll

**senate\_class** the class of the senate

**start\_date** the start date of the poll

**end\_date** the end odate of the poll

**dem\_poll** the percent of support for the Democrat during the poll

**rep\_poll** the percent of support for the Republican during the poll

**dem\_result** the result percent of support for the Democrat during the election

**rep\_result** the result percent of support for the Republican during the election

**error** the difference between the percent of support of one party during the poll and the result percent of support for the same party during the election

**absolute\_error** the absolute value of the error value

### Source

Emerson College's poll of registered voters

---

avengers	<i>Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building</i>
----------	---

---

### Description

The raw data behind the story "Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building" <https://fivethirtyeight.com/features/avengers-death-comics-age-of-ultron/>.

### Usage

avengers

### Format

A data frame with 173 rows representing characters and 21 variables:

**url** The URL of the comic character on the Marvel Wikia

**name\_alias** The full name or alias of the character

**appearances** The number of comic books that character appeared in as of April 30

**current** Is the member currently active on an avengers affiliated team?

**gender** The recorded gender of the character

**probationary\_intro** Sometimes the character was given probationary status as an Avenger, this is the date that happened

**full\_reserve\_avengers\_intro** The month and year the character was introduced as a full or reserve member of the Avengers

**year** The year the character was introduced as a full or reserve member of the Avengers

**years\_since\_joining** 2015 minus the year

**honorary** The status of the avenger, if they were given "Honorary" Avenger status, if they are simply in the "Academy," or "Full" otherwise

**death1** TRUE if the Avenger died, FALSE if not.

**return1** TRUE if the Avenger returned from their first death, FALSE if they did not, blank if not applicable

**death2** TRUE if the Avenger died a second time after their revival, FALSE if they did not, blank if not applicable

**return2** TRUE if the Avenger returned from their second death, FALSE if they did not, blank if not applicable

**death3** TRUE if the Avenger died a third time after their second revival, FALSE if they did not, blank if not applicable

**return3** TRUE if the Avenger returned from their third death, FALSE if they did not, blank if not applicable

**death4** TRUE if the Avenger died a fourth time after their third revival, FALSE if they did not, blank if not applicable

**return4** TRUE if the Avenger returned from their fourth death, FALSE if they did not, blank if not applicable

**death5** TRUE if the Avenger died a fifth time after their fourth revival, FALSE if they did not, blank if not applicable

**return5** TRUE if the Avenger returned from their fifth death, FALSE if they did not, blank if not applicable

**notes** Descriptions of deaths and resurrections.

### Source

Deaths of Marvel comic book characters between the time they joined the Avengers and April 30, 2015, the week before Secret Wars #1.

---

bachelorette

*Bachelorette / Bachelor*

---

### Description

The raw data behind the stories: "How To Spot A Front-Runner On The 'Bachelor' Or 'Bachelorette'" <https://fivethirtyeight.com/features/the-bachelorette/>, "Rachel's Season Is Fitting Neatly Into 'Bachelorette' History" <https://fivethirtyeight.com/features/rachels-season-is-fitting-neatly-into-bachelorette-history/> and "Rachel Lindsay's 'Bachelorette' Season, In Three Charts" <https://fivethirtyeight.com/features/rachel-lindsays-bachelorette-season-in-three-charts/>.

### Usage

bachelorette

### Format

A data frame with 887 rows representing the Bachelorette and Bachelor contestants and 23 variables:

**show** Bachelor or Bachelorette.

**season** Which season?

**contestant** An identifier for the contestant in a given season.

**elimination\_1** Who was eliminated in week 1.

**elimination\_2** Who was eliminated in week 2.

**elimination\_3** Who was eliminated in week 3.

**elimination\_4** Who was eliminated in week 4.

**elimination\_5** Who was eliminated in week 5.

**elimination\_6** Who was eliminated in week 6.

**elimination\_7** Who was eliminated in week 7.

**elimination\_8** Who was eliminated in week 8.

- elimination\_9** Who was eliminated in week 9.
- elimination\_10** Who was eliminated in week 10.
- dates\_1** Who was on which date in week 1.
- dates\_2** Who was on which date in week 2.
- dates\_3** Who was on which date in week 3.
- dates\_4** Who was on which date in week 4.
- dates\_5** Who was on which date in week 5.
- dates\_6** Who was on which date in week 6.
- dates\_7** Who was on which date in week 7.
- dates\_8** Who was on which date in week 8.
- dates\_9** Who was on which date in week 9.
- dates\_10** Who was on which date in week 10.

### Details

Eliminates connote either an elimination (starts with "E") or a rose (starts with "R"). Eliminations supersede roses. "E" connotes a standard elimination, typically at a rose ceremony. "EQ" means the contestant quits. "EF" means the contestant was fired by production. "ED" connotes a date elimination. "EU" connotes an unscheduled elimination, one that takes place at a time outside of a date or rose ceremony. "R" means the contestant received a rose. "R1" means the contestant got a first impression rose. "D1" means a one-on-one date, "D2" means a 2-on-1, "D3" means a 3-on-1 group date, and so on. Weeks of the show are eliminated by rose ceremonies, and may not line up exactly with episodes.

### Source

[https://bachelor-nation.fandom.com/wiki/Bachelor\\_Nation\\_Wiki](https://bachelor-nation.fandom.com/wiki/Bachelor_Nation_Wiki) and then missing seasons were filled in by ABC and FiveThirtyEight staffers.

---

bad\_drivers

*Dear Mona, Which State Has The Worst Drivers?*

---

### Description

The raw data behind the story "Dear Mona, Which State Has The Worst Drivers?" <https://fivethirtyeight.com/features/which-state-has-the-worst-drivers/>

### Usage

bad\_drivers

**Format**

A data frame with 51 rows representing the 50 states + D.C. and 8 variables:

**state** State

**num\_drivers** Number of drivers involved in fatal collisions per billion miles

**perc\_speeding** Percentage of drivers involved in fatal collisions who were speeding

**perc\_alcohol** Percentage of drivers involved in fatal collisions who were alcohol-impaired

**perc\_not\_distracted** Percentage of drivers involved in fatal collisions who were not distracted

**perc\_no\_previous** Percentage of drivers involved in fatal collisions who had not been involved in any previous accidents

**insurance\_premiums** Car insurance premiums (\$)

**losses** Losses incurred by insurance companies for collisions per insured driver (\$)

**Source**

National Highway Traffic Safety Administration 2012, National Highway Traffic Safety Administration 2009 & 2012, National Association of Insurance Commissioners 2010 & 2011.

---

bechdel

*The Dollar-And-Cents Case Against Hollywood's Exclusion of Women*

---

**Description**

The raw data behind the story "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" <https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion>

**Usage**

bechdel

**Format**

A data frame with 1794 rows representing movies and 15 variables:

**year** Year of release

**imdb** Text to construct IMDB url. Ex: <https://www.imdb.com/title/tt1711425>

**title** Movie test

**test** bechdel test result (detailed, with discrepancies indicated)

**clean\_test** bechdel test result (detailed): ok = passes test, dubious, men = women only talk about men, notalk = women don't talk to each other, nowomen = fewer than two women

**binary** Bechdel Test PASS vs FAIL binary

**budget** Film budget

**domgross** Domestic (US) gross

**intgross** Total International (i.e., worldwide) gross  
**code** Bechdel Code  
**budget\_2013** Budget in 2013 inflation adjusted dollars  
**domgross\_2013** Domestic gross (US) in 2013 inflation adjusted dollars  
**intgross\_2013** Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars  
**period\_code**  
**decade\_code**

### Details

A vignette of an analysis of this dataset using the tidyverse can be found on [CRAN](#) or by running:  
`vignette("bechdel", package = "fivethirtyeightdata")`

### Source

<https://bechdeltest.com/> and <https://www.the-numbers.com/>. The original data can be found at <https://github.com/fivethirtyeight/data/tree/master/bechdel>.

---

biopics

*'Straight Outta Compton' Is The Rare Biopic Not About White Dudes*

---

### Description

The raw data behind the story "'Straight Outta Compton' Is The Rare Biopic Not About White Dudes" <https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about-white-dudes/>  
 An analysis using this data was contributed by Pradeep Adhokshaja as a package vignette at <https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/biopics.html>.

### Usage

biopics

### Format

A data frame with 761 rows representing movies and 14 variables:

**title** Title of the film.  
**site** Text to construct IMDB url. Ex: <https://www.imdb.com/title/tt1711425>  
**country** Country of origin.  
**year\_release** Year of release.  
**box\_office** Gross earnings at U.S. box office.  
**director** Director of film.  
**number\_of\_subjects** The number of subjects featured in the film.  
**subject** The actual name of the featured subject.

**type\_of\_subject** The occupation of subject or reason for recognition.

**race\_known** Indicates whether the subject's race was discernible based on background of self, parent, or grandparent.

**subject\_race** Race of the subject.

**person\_of\_color** Dummy variable that indicates person of color.

**subject\_sex** Sex of subject.

**lead\_actor\_actress** The actor or actress who played the subject.

### Source

IMDB <https://www.imdb.com/>

---

bob\_ross

*A Statistical Analysis of the Work of Bob Ross*

---

### Description

The raw data behind the story "A Statistical Analysis of the Work of Bob Ross" <https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/>. An analysis using this data was contributed by Jonathan Bouchet as a package vignette at [https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/bob\\_ross.html](https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/bob_ross.html).

### Usage

bob\_ross

### Format

A data frame with 403 rows representing episodes and 71 variables:

**episode** Episode code

**season** Season number

**episode\_num** Episode number

**title** Title of episode

**apple\_frame** Present (1) or not (0)

**aurora\_borealis** Present (1) or not (0)

**barn** Present (1) or not (0)

**beach** Present (1) or not (0)

**boat** Present (1) or not (0)

**bridge** Present (1) or not (0)

**building** Present (1) or not (0)

**bushes** Present (1) or not (0)

**cabin** Present (1) or not (0)  
**cactus** Present (1) or not (0)  
**circle\_frame** Present (1) or not (0)  
**cirrus** Present (1) or not (0)  
**cliff** Present (1) or not (0)  
**clouds** Present (1) or not (0)  
**conifer** Present (1) or not (0)  
**cumulus** Present (1) or not (0)  
**deciduous** Present (1) or not (0)  
**diane\_andre** Present (1) or not (0)  
**dock** Present (1) or not (0)  
**double\_oval\_frame** Present (1) or not (0)  
**farm** Present (1) or not (0)  
**fence** Present (1) or not (0)  
**fire** Present (1) or not (0)  
**florida\_frame** Present (1) or not (0)  
**flowers** Present (1) or not (0)  
**fog** Present (1) or not (0)  
**framed** Present (1) or not (0)  
**grass** Present (1) or not (0)  
**guest** Present (1) or not (0)  
**half\_circle\_frame** Present (1) or not (0)  
**half\_oval\_frame** Present (1) or not (0)  
**hills** Present (1) or not (0)  
**lake** Present (1) or not (0)  
**lakes** Present (1) or not (0)  
**lighthouse** Present (1) or not (0)  
**mill** Present (1) or not (0)  
**moon** Present (1) or not (0)  
**mountain** Present (1) or not (0)  
**mountains** Present (1) or not (0)  
**night** Present (1) or not (0)  
**ocean** Present (1) or not (0)  
**oval\_frame** Present (1) or not (0)  
**palm\_trees** Present (1) or not (0)  
**path** Present (1) or not (0)  
**person** Present (1) or not (0)

**portrait** Present (1) or not (0)  
**rectangle\_3d\_frame** Present (1) or not (0)  
**rectangular\_frame** Present (1) or not (0)  
**river** Present (1) or not (0)  
**rocks** Present (1) or not (0)  
**seashell\_frame** Present (1) or not (0)  
**snow** Present (1) or not (0)  
**snowy\_mountain** Present (1) or not (0)  
**split\_frame** Present (1) or not (0)  
**steve\_ross** Present (1) or not (0)  
**structure** Present (1) or not (0)  
**sun** Present (1) or not (0)  
**tomb\_frame** Present (1) or not (0)  
**tree** Present (1) or not (0)  
**trees** Present (1) or not (0)  
**triple\_frame** Present (1) or not (0)  
**waterfall** Present (1) or not (0)  
**waves** Present (1) or not (0)  
**windmill** Present (1) or not (0)  
**window\_frame** Present (1) or not (0)  
**winter** Present (1) or not (0)  
**wood\_framed** Present (1) or not (0)

### Source

See <https://github.com/fivethirtyeight/data/tree/master/bob-ross>

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
bob_ross_tidy <- bob_ross %>%
  pivot_longer(-c(episode, season, episode_num, title),
    names_to = "object", values_to = "present") %>%
  mutate(present = as.logical(present)) %>%
  arrange(episode, object)
```

---

cabinet_turnover	<i>Two Years In, Turnover In Trump's Cabinet Is Still Historically High</i>
------------------	---

---

**Description**

The raw data behind the story "Two Years In, Turnover In Trump's Cabinet Is Still Historically High" <https://fivethirtyeight.com/features/two-years-in-turnover-in-trumps-cabinet-is-still-histori>

**Usage**

cabinet\_turnover

**Format**

A data frame with 312 rows representing cabinet members and 8 variables:

**president** Surname of of sitting President

**position** Cabinet Position

**appointee** Appointee's full name

**start** Date the appointee was sworn in

**end** Date the appointee left office

**length** Length of Tenure, in days

**days** Days into administration that the appointee left office

**combined** Whether or not Cabinet member served in more than one administrations

**Source**

from Federal Government Websites and News Reports

---

candy_rankings	<i>Candy Power Ranking</i>
----------------	----------------------------

---

**Description**

The raw data behind the story "The Ultimate Halloween Candy Power Ranking" <https://fivethirtyeight.com/features/the-ultimate-halloween-candy-power-ranking/>.

**Usage**

candy\_rankings

**Format**

A data frame with 85 rows representing Halloween candy and 13 variables:

**competitorname** The name of the Halloween candy.

**chocolate** Does it contain chocolate?

**fruity** Is it fruit flavored?

**caramel** Is there caramel in the candy?

**peanutyalmondy** Does it contain peanuts, peanut butter or almonds?

**nougat** Does it contain nougat?

**crispedricewafer** Does it contain crisped rice, wafers, or a cookie component?

**hard** Is it a hard candy?

**bar** Is it a candy bar?

**pluribus** Is it one of many candies in a bag or box?

**sugarpercent** The percentile of sugar it falls under within the data set.

**pricepercent** The unit price percentile compared to the rest of the set.

**winpercent** The overall win percentage according to 269,000 matchups.

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking>

**Examples**

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
candy_rankings_tidy <- candy_rankings %>%
  pivot_longer(-c(competitorname, sugarpercent, pricepercent, winpercent),
    names_to = "characteristics", values_to = "present") %>%
  mutate(present = as.logical(present)) %>%
  arrange(competitorname)
```

---

cand\_events\_20150114 *Looking For Clues: Who Is Going To Run For President In 2016?*

---

**Description**

The raw data behind the story "Looking For Clues: Who Is Going To Run For President In 2016?"  
<https://fivethirtyeight.com/features/2016-president-who-is-going-to-run/>.

**Usage**

```
cand_events_20150114
```

**Format**

A data frame with 42 rows representing events attended in Iowa and New Hampshire by potential presidential primary candidates and 8 variables:

**person** Potential presidential candidate  
**party** Political party  
**state** State of event  
**event** Name of event  
**type** Type of event  
**date** Date of event  
**link** Link to event  
**snippet** Snippet of event description

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/potential-candidates>

**See Also**

[cand\\_state\\_20150114](#), [cand\\_events\\_20150130](#), and [cand\\_state\\_20150130](#)

---

cand\_events\_20150130 *Who Will Run For President: Romney Is Out*

---

**Description**

The raw data behind the story "Who Will Run For President: Romney Is Out" <https://fivethirtyeight.com/features/romney-not-running-for-president/>.

**Usage**

```
cand_events_20150130
```

**Format**

A data frame with 74 rows representing events attended by potential presidential primary candidates and 8 variables:

**person** Potential presidential candidate  
**party** Political party  
**state** State of event  
**event** Name of event  
**type** Type of event  
**date** Date of event  
**link** Link to event  
**snippet** Snippet of event description

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/potential-candidates>

**See Also**

[cand\\_state\\_20150130](#), [cand\\_events\\_20150114](#), and [cand\\_state\\_20150114](#)

---

cand\_state\_20150114    *Looking For Clues: Who Is Going To Run For President In 2016?*

---

**Description**

The raw data behind the story "Looking For Clues: Who Is Going To Run For President In 2016?" <https://fivethirtyeight.com/features/2016-president-who-is-going-to-run/>.

**Usage**

cand\_state\_20150114

**Format**

A data frame with 25 rows representing potential presidential primary candidates and 5 variables:

**person** Potential presidential candidate

**party** Political party

**date** Date of event

**latest** Latest statement

**score** Likelihood of running score, 1 = Not running, 5 = Definitely running

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/potential-candidates>

**See Also**

[cand\\_events\\_20150114](#), [cand\\_events\\_20150130](#), and [cand\\_state\\_20150130](#)

---

cand\_state\_20150130     *Who Will Run For President: Romney Is Out*

---

**Description**

The raw data behind the story "Who Will Run For President: Romney Is Out" <https://fivethirtyeight.com/features/romney-not-running-for-president/>.

**Usage**

cand\_state\_20150130

**Format**

A data frame with 27 rows representing potential presidential primary candidates and 5 variables:

**person** Potential presidential candidate

**party** Political party

**date** Date of event

**latest** Latest statement

**score** Likelihood of running score, 1 = Not running, 5 = Definitely running

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/potential-candidates>

**See Also**

[cand\\_events\\_20150130](#), [cand\\_events\\_20150114](#), and [cand\\_state\\_20150114](#)

---

chess\_transfers     *Chess Transfers*

---

**Description**

The raw data behind the story "American Chess Is Great Again" <https://fivethirtyeight.com/features/american-chess-is-great-again/>.

**Usage**

chess\_transfers

**Format**

A data frame with 932 rows representing international player transfers and 5 variables:

**url** The corresponding website on the World Chess Federation page which details the transfers of a given year.

**id** An numeric identifier for the chess player who transferred.

**federation** The current national federation of the chess player

**form\_fed** The national federation from which the chess player has transferred.

**transfer\_date** The date at which the transfer took place.

**Source**

World Chess Federation

---

classic\_rock\_raw\_data *Why Classic Rock Isn't What It Used To Be*

---

**Description**

The raw data behind the story "Why Classic Rock Isn't What It Used To Be" <https://fivethirtyeight.com/features/why-classic-rock-isnt-what-it-used-to-be/>.

**Usage**

```
classic_rock_raw_data
```

**Format**

A data frame with 37,673 rows representing song plays and 8 variables:

**song** Song name

**artist** Artist name

**callsign** Station callsign

**time** Time of song play in seconds elapsed since January 1, 1970

**date\_time** Time of song play in date/time format

**unique\_id** Unique ID for each song play

**combined** Song and artist name combined

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/classic-rock>

**See Also**

[classic\\_rock\\_song\\_list](#)

---

`classic_rock_song_list`*Why Classic Rock Isn't What It Used To Be*

---

**Description**

The raw data behind the story "Why Classic Rock Isn't What It Used To Be" <https://fivethirtyeight.com/features/why-classic-rock-isnt-what-it-used-to-be/>.

**Usage**`classic_rock_song_list`**Format**

A data frame with 2230 rows representing unique songs and 7 variables:

**song** Song name

**artist** Artist name

**release\_year** Release year as listed in SongFacts

**combined** Song and artist name combined

**has\_year** Logical variable of whether release year is included

**playcount** Number of plays across all stations

**playcount\_has\_year** Number of plays across all stations if a year was found

**Source**

SongFacts and <https://github.com/fivethirtyeight/data/tree/master/classic-rock>

**See Also**

[classic\\_rock\\_raw\\_data](#)

---

`college_all_ages`*The Economic Guide To Picking A College Major*

---

**Description**

The raw data behind the story "The Economic Guide To Picking A College Major" <https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/>.

**Usage**`college_all_ages`

**Format**

A data frame with 173 rows representing majors (all ages) and 11 variables:

**major\_code** Major code, FO1DP in ACS PUMS  
**major** Major description  
**major\_category** Category of major from Carnevale et al  
**total** Total number of people with major  
**employed** Number employed (ESR == 1 or 2)  
**employed\_fulltime\_yearround** Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)  
**unemployed** Number unemployed (ESR == 3)  
**unemployment\_rate** Unemployed / (Unemployed + Employed)  
**p25th** 25th percentile of earnings  
**median** Median earnings of full-time, year-round workers  
**p75th** 75th percentile of earnings

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/college-majors/readme.md>.

**See Also**

[college\\_grad\\_students](#), [college\\_recent\\_grads](#)

---

college\_grad\_students *The Economic Guide To Picking A College Major*

---

**Description**

The raw data behind the story "The Economic Guide To Picking A College Major" <https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/>.

**Usage**

```
college_grad_students
```

**Format**

A data frame with 173 rows representing majors (graduate vs nongraduate students) and 22 variables:

**major\_code** Major code, FO1DP in ACS PUMS  
**major** Major description  
**major\_category** Category of major from Carnevale et al

**grad\_total** Total number of people with major

**grad\_sample\_size** Sample size (unweighted) of full-time, year-round ONLY (used for earnings)

**grad\_employed** Number employed (ESR == 1 or 2)

**grad\_employed\_fulltime\_yearround** Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

**grad\_unemployed** Number unemployed (ESR == 3)

**grad\_unemployment\_rate** Unemployed / (Unemployed + Employed)

**grad\_p25th** 25th percentile of earnings

**grad\_median** Median earnings of full-time, year-round workers

**grad\_p75th** 75th percentile of earnings

**nongrad\_total** Total number of people with major

**nongrad\_employed** Number employed (ESR == 1 or 2)

**nongrad\_employed\_fulltime\_yearround** Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

**nongrad\_unemployed** Number unemployed (ESR == 3)

**nongrad\_unemployment\_rate** Unemployed / (Unemployed + Employed)

**nongrad\_p25th** 25th percentile of earnings

**nongrad\_median** Median earnings of full-time, year-round workers

**nongrad\_p75th** 75th percentile of earnings

**grad\_share** grad\_total / (grad\_total + nongrad\_total)

**grad\_premium** (grad\_median-nongrad\_median)/nongrad\_median

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/college-majors/readme.md>.

**See Also**

[college\\_all\\_ages](#), [college\\_recent\\_grads](#)

---

college\_recent\_grads    *The Economic Guide To Picking A College Major*

---

**Description**

The raw data behind the story "The Economic Guide To Picking A College Major" <https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/>.

**Usage**

college\_recent\_grads

**Format**

A data frame with 173 rows representing majors (recent graduates) and 21 variables:

**rank** Rank by median earnings

**major\_code** Major code, FO1DP in ACS PUMS

**major** Major description

**major\_category** Category of major from Carnevale et al

**total** Total number of people with major

**sample\_size** Sample size (unweighted) of full-time, year-round ONLY (used for earnings)

**men** Men with major

**women** Women with major

**sharewomen** Proportion women

**employed** Number employed (ESR == 1 or 2)

**employed\_fulltime** Employed 35 hours or more

**employed\_parttime** Employed less than 35 hours

**employed\_fulltime\_yearround** Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

**unemployed** Number unemployed (ESR == 3)

**unemployment\_rate** Unemployed / (Unemployed + Employed)

**p25th** 25th percentile of earnings

**median** Median earnings of full-time, year-round workers

**p75th** 75th percentile of earnings

**college\_jobs** Number with job requiring a college degree

**non\_college\_jobs** Number with job not requiring a college degree

**low\_wage\_jobs** Number in low-wage service jobs

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/college-majors/readme.md>. Note that women-stem.csv was a subset of the original recent-grads.csv, so no data frame was created.

**See Also**

[college\\_grad\\_students](#), [college\\_all\\_ages](#)

---

comma_survey	<i>Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma</i>
--------------	--

---

## Description

The raw data behind the story "Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma" <https://fivethirtyeight.com/features/elitist-superfluous-or-popular-we-polled-americans>

## Usage

```
comma_survey
```

## Format

A data frame with 1129 rows representing respondents and 13 variables:

**respondent\_id** Respondent ID

**gender** Gender

**age** Age

**household\_income** Household income bracket

**education** Education level

**location** Location (census region)

**more\_grammar\_correct** In your opinion, which sentence is more grammatically correct?

**heard\_oxford\_comma** Prior to reading about it above, had you heard of the serial (or Oxford) comma?

**care\_oxford\_comma** How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?

**write\_following** How would you write the following sentence?

**data\_singular\_plural** When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?

**care\_data** How much, if at all, do you care about the debate over the use of the word "data" as a singular or plural noun?

**care\_proper\_grammar** In your opinion, how important or unimportant is proper use of grammar?

## Source

See <https://github.com/fivethirtyeight/data/tree/master/comma-survey>.

---

congress\_age

*Both Republicans And Democrats Have an Age Problem*

---

## Description

The raw data behind the story "Both Republicans And Democrats Have an Age Problem" <https://fivethirtyeight.com/features/both-republicans-and-democrats-have-an-age-problem/>.

## Usage

congress\_age

## Format

A data frame with 18,635 rows representing members of Congress (House and Senate) and 13 variables:

**congress** Congress number.

**chamber** Chamber of congress: House of Representatives or Senate.

**bioguide** bioguide

**firstname** First name

**middlename** Middle name

**lastname** Last name

**suffix** Suffix

**birthday** Birthday

**state** State abbreviation

**party** Party abbreviation

**incumbent** Boolean variable of whether member was an incumbent.

**termstart** Start date of session.

**age** Age at start of session.

## Source

See <https://github.com/fivethirtyeight/data/tree/master/congress-age>

---

cousin_marriage	<i>How Many Americans Are Married To Their Cousins?</i>
-----------------	---

---

**Description**

The raw data behind the story "How Many Americans Are Married To Their Cousins?" <https://fivethirtyeight.com/features/how-many-americans-are-married-to-their-cousins/>.

**Usage**

```
cousin_marriage
```

**Format**

A data frame with 70 rows representing countries and 2 variables:

**country** Country

**percent** Percent of marriages that are consanguineous

**Source**

```
consang.net
```

---

daily_show_guests	<i>Every Guest Jon Stewart Ever Had On 'The Daily Show'</i>
-------------------	---

---

**Description**

The raw data behind the story "Every Guest Jon Stewart Ever Had On 'The Daily Show'" <https://fivethirtyeight.com/features/every-guest-jon-stewart-ever-had-on-the-daily-show/>.

**Usage**

```
daily_show_guests
```

**Format**

A data frame with 2693 rows representing guests and 5 variables:

**year** The year the episode aired

**google\_knowledge\_occupation** Their occupation or office, according to Google's Knowledge Graph or, if they're not in there, how Stewart introduced them on the program.

**show** Air date of episode. Not unique, as some shows had more than one guest

**group** A larger group designation for the occupation. For instance, us senators, us presidents, and former presidents are all under "politicians"

**raw\_guest\_list** The person or list of people who appeared on the show, according to Wikipedia. The GoogleKnowledge\_Occupation only refers to one of them in a given row.

**Source**

Google Knowledge Graph, The Daily Show clip library, Wikipedia.

---

datasets_master	<i>Master list of all datasets</i>
-----------------	------------------------------------

---

**Description**

All datasets included in both fivethirtyeight and fivethirtyeightdata packages

**Usage**

datasets\_master

**Format**

A data frame with 9 variables:

**Data Frame Name** Name of lazy-loaded data frame

**In fivethirtyeightdata?** Whether the (large) dataset is in the fivethirtyeightdata package

**Article Title** Title as it appears on FiveThirtyEight.com

**URL** Link to article on FiveThirtyEight.com

**Author 1** Main author

**Author 2** Second author (if any)

**Author 3** Third author (if any)

**Date** Date published

**Filed Under** Tag for article

---

democratic_bench	<i>Some Democrats Who Could Step Up If Hillary Isn't Ready For Hillary</i>
------------------	--

---

**Description**

The raw data behind the story "Some Democrats Who Could Step Up If Hillary Isn't Ready For Hillary" <https://fivethirtyeight.com/features/some-democrats-who-could-step-up-if-hillary-isnt-ready>

**Usage**

democratic\_bench

**Format**

A data frame with 67 rows representing members of the Democratic Party and 3 variables:

**candidate** Candidate

**raised\_exp** Amount the candidate was expected to raise

**raised\_act** Amount the candidate actually raised

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/democratic-bench>.

---

 dem\_candidates

*Democratic Primary Candidates 2018*


---

**Description**

The raw data behind the stories: "We Researched Hundreds Of Races. Here's Who Democrats Are Nominating" <https://fivethirtyeight.com/features/democrats-primaries-candidates-demographics/> and "How's The Progressive Wing Doing In Democratic Primaries So Far?" <https://fivethirtyeight.com/features/the-establishment-is-beating-the-progressive-wing-in-democratic-primaries-so-far/>.

**Usage**

dem\_candidates

**Format**

A data frame with 811 rows representing Democratic candidates, and 32 variables:

**candidate** All candidates who received votes in 2018's Democratic primary elections for U.S. Senate, U.S. House and governor in which no incumbent ran. Supplied by Ballotpedia.

**state** The state in which the candidate ran. Supplied by Ballotpedia.

**body** The body of government for which the candidate ran. Supplied by Ballotpedia.

**district\_num** If applicable, congressional district number for which the candidate ran. Supplied by Ballotpedia.

**office\_type** The office for which the candidate ran. Supplied by Ballotpedia.

**race\_type** Whether it was a "regular" or "special" election. Supplied by Ballotpedia.

**race\_primary\_election\_date** The date on which the primary was held. Supplied by Ballotpedia.

**primary\_status** Whether the candidate lost ("Lost") the primary or won/advanced to a runoff ("Advanced"). Supplied by Ballotpedia.

**primary\_runoff\_status** "None" if there was no runoff; "On the Ballot" if the candidate advanced to a runoff but it hasn't been held yet; "Advanced" if the candidate won the runoff; "Lost" if the candidate lost the runoff. Supplied by Ballotpedia.

- general\_status** “On the Ballot” if the candidate won the primary or runoff and has advanced to November; otherwise, “None.” Supplied by Ballotpedia.
- partisan\_lean** The FiveThirtyEight partisan lean of the district or state in which the election was held. Partisan leans are calculated by finding the average difference between how a state or district voted in the past two presidential elections and how the country voted overall, with 2016 results weighted 75 percent and 2012 results weighted 25 percent.
- primary\_percent** The percentage of the vote received by the candidate in his or her primary. In states that hold runoff elections, we looked only at the first round (the regular primary). In states that hold all-party primaries (e.g., California), a candidate’s primary percentage is the percentage of the total Democratic vote they received. Unopposed candidates and candidates nominated by convention (not primary) are given a primary percentage of 100 but were excluded from our analysis involving vote share. Numbers come from official results posted by the secretary of state or local elections authority; if those were unavailable, we used unofficial election results from the New York Times.
- won\_primary** “Yes” if the candidate won his or her primary and has advanced to November; “No” if he or she lost.
- race** “White” if we identified the candidate as non-Hispanic white; “Nonwhite” if we identified the candidate as Hispanic and/or any nonwhite race; blank if we could not identify the candidate’s race or ethnicity. To determine race and ethnicity, we checked each candidate’s website to see if he or she identified as a certain race. If not, we spent no more than two minutes searching online news reports for references to the candidate’s race.
- veteran** If the candidate’s website says that he or she served in the armed forces, we put “Yes.” If the website is silent on the subject (or explicitly says he or she didn’t serve), we put “No.” If the field was left blank, no website was available.
- lgbtq** If the candidate’s website says that he or she is LGBTQ (including indirect references like to a same-sex partner), we put “Yes.” If the website is silent on the subject (or explicitly says he or she is straight), we put “No.” If the field was left blank, no website was available.
- elected\_official** We used Ballotpedia, VoteSmart and news reports to research whether the candidate had ever held elected office before, at any level. We put “Yes” if the candidate has held elected office before and “No” if not.
- self\_funder** We used Federal Election Committee fundraising data (for federal candidates) and state campaign-finance data (for gubernatorial candidates) to look up how much each candidate had invested in his or her own campaign, through either donations or loans. We put “Yes” if the candidate donated or loaned a cumulative \$400,000 or more to his or her own campaign before the primary and “No” for all other candidates.
- stem** If the candidate identifies on his or her website that he or she has a background in the fields of science, technology, engineering or mathematics, we put “Yes.” If not, we put “No.” If the field was left blank, no website was available.
- obama\_alum** We put “Yes” if the candidate mentions working for the Obama administration or campaign on his or her website, or if the candidate shows up on this list of Obama administration members and campaign hands running for office. If not, we put “No.”
- party\_support** “Yes” if the candidate was placed on the DCCC’s Red to Blue list before the primary, was endorsed by the DSCC before the primary, or if the DSCC/DCCC aired pre-primary ads in support of the candidate. (Note: according to the DGA’s press secretary, the DGA does not get involved in primaries.) “No” if the candidate is running against someone for whom

one of the above things is true, or if one of those groups specifically anti-endorsed or spent money to attack the candidate. If those groups simply did not weigh in on the race, we left the cell blank.

**emily\_endorsed** “Yes” if the candidate was endorsed by Emily’s List before the primary. “No” if the candidate is running against an Emily-endorsed candidate or if Emily’s List specifically anti-endorsed or spent money to attack the candidate. If Emily’s List simply did not weigh in on the race, we left the cell blank.

**guns\_sense\_candidate** “Yes” if the candidate received the Gun Sense Candidate Distinction from Moms Demand Action/Everytown for Gun Safety before the primary, according to media reports or the candidate’s website. “No” if the candidate is running against an candidate with the distinction. If Moms Demand Action simply did not weigh in on the race, we left the cell blank.

**biden\_endorsed** “Yes” if the candidate was endorsed by Joe Biden before the primary. “No” if the candidate is running against a Biden-endorsed candidate or if Biden specifically anti-endorsed the candidate. If Biden simply did not weigh in on the race, we left the cell blank.

**warren\_endorsed** “Yes” if the candidate was endorsed by Elizabeth Warren before the primary. “No” if the candidate is running against a Warren-endorsed candidate or if Warren specifically anti-endorsed the candidate. If Warren simply did not weigh in on the race, we left the cell blank.

**sanders\_endorsed** “Yes” if the candidate was endorsed by Bernie Sanders before the primary. “No” if the candidate is running against a Sanders-endorsed candidate or if Sanders specifically anti-endorsed the candidate. If Sanders simply did not weigh in on the race, we left the cell blank.

**our\_revolution\_endorsed** “Yes” if the candidate was endorsed by Our Revolution before the primary, according to the Our Revolution website. “No” if the candidate is running against an Our Revolution-endorsed candidate or if Our Revolution specifically anti-endorsed or spent money to attack the candidate. If Our Revolution simply did not weigh in on the race, we left the cell blank.

**justice\_dems\_endorsed** “Yes” if the candidate was endorsed by Justice Democrats before the primary, according to the Justice Democrats website, candidate website or news reports. “No” if the candidate is running against a Justice Democrats-endorsed candidate or if Justice Democrats specifically anti-endorsed or spent money to attack the candidate. If Justice Democrats simply did not weigh in on the race, we left the cell blank.

**pccc\_endorsed** “Yes” if the candidate was endorsed by the Progressive Change Campaign Committee before the primary, according to the PCCC website, candidate website or news reports. “No” if the candidate is running against a PCCC-endorsed candidate or if the PCCC specifically anti-endorsed or spent money to attack the candidate. If the PCCC simply did not weigh in on the race, we left the cell blank.

**indivisible\_endorsed** “Yes” if the candidate was endorsed by Indivisible before the primary, according to the Indivisible website, candidate website or news reports. “No” if the candidate is running against an Indivisible-endorsed candidate or if Indivisible specifically anti-endorsed or spent money to attack the candidate. If Indivisible simply did not weigh in on the race, we left the cell blank.

**wfp\_endorsed** “Yes” if the candidate was endorsed by the Working Families Party before the primary, according to the WFP website, candidate website or news reports. “No” if the candidate is running against a WFP-endorsed candidate or if the WFP specifically anti-endorsed or spent

money to attack the candidate. If the WFP simply did not weigh in on the race, we left the cell blank.

**vote\_vets\_endorsed** “Yes” if the candidate was endorsed by VoteVets before the primary, according to the VoteVets website, candidate website or news reports. “No” if the candidate is running against a VoteVets-endorsed candidate or if VoteVets specifically anti-endorsed or spent money to attack the candidate. If VoteVets simply did not weigh in on the race, we left the cell blank.

**no\_labels\_support** “Yes” if a No Labels-affiliated group (Citizens for a Strong America Inc., Forward Not Back, Govern or Go Home, United for Progress Inc. or United Together) spent money in support of the candidate in the primary. “No” if the candidate is running against an candidate supported by a No Labels-affiliated group or if a No Labels-affiliated group specifically anti-endorsed or spent money to attack the candidate. If No Labels simply did not weigh in on the race, we left the cell blank.

### Note

This data was also used in "We Looked At Hundreds Of Endorsements. Here's Who Democrats Are Listening To" published on 2008-08-14 <https://fivethirtyeight.com/features/the-establishment-is-beating-t>

### Source

Ballotpedia, New York Times, and candidate websites. See also <https://github.com/fivethirtyeight/data/blob/master/primary-candidates-2018/README.md>

---

drinks

*Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?*

---

### Description

The raw data behind the story "Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?" <https://fivethirtyeight.com/features/dear-mona-followup-where-do-people-drink-the-most->

### Usage

drinks

### Format

A data frame with 193 rows representing countries and 5 variables:

**country** country

**beer\_servings** Servings of beer in average serving sizes per person

**spirit\_servings** Servings of spirits in average serving sizes per person

**wine\_servings** Servings of wine in average serving sizes per person

**total\_litres\_of\_pure\_alcohol** Total litres of pure alcohol per person

**Source**

World Health Organization, Global Information System on Alcohol and Health (GISAH), 2010.

**Examples**

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
drinks_tidy <- drinks %>%
  pivot_longer(cols = ends_with("servings"), names_to = "type", values_to = "servings") %>%
  mutate(
    type = str_sub(type, start=1, end=-10)
  ) %>%
  arrange(country, type)
```

---

 drug\_use

*How Baby Boomers Get High*


---

**Description**

The raw data behind the story "How Baby Boomers Get High" <https://fivethirtyeight.com/features/how-baby-boomers-get-high/>. It covers usage of 13 drugs in the past 12 months across 17 age groups.

**Usage**

drug\_use

**Format**

A data frame with 17 rows representing age groups and 28 variables:

**age** Age group

**n** Number of people surveyed

**alcohol\_use** Percentage who used alcohol

**alcohol\_freq** Median number of times a user used alcohol

**marijuana\_use** Percentage who used marijuana

**marijuana\_freq** Median number of times a user used marijuana

**cocaine\_use** Percentage who used cocaine

**cocaine\_freq** Median number of times a user used cocaine

**crack\_use** Percentage who used crack

**crack\_freq** Median number of times a user used crack

**heroin\_use** Percentage who used heroin

**heroin\_freq** Median number of times a user used heroin

**hallucinogen\_use** Percentage who used hallucinogens  
**hallucinogen\_freq** Median number of times a user used hallucinogens  
**inhalant\_use** Percentage who used inhalants  
**inhalant\_freq** Median number of times a user used inhalants  
**pain\_releiver\_use** Percentage who used pain relievers  
**pain\_releiver\_freq** Median number of times a user used pain relievers  
**oxycontin\_use** Percentage who used oxycontin  
**oxycontin\_freq** Median number of times a user used oxycontin  
**tranquilizer\_use** Percentage who used tranquilizer  
**tranquilizer\_freq** Median number of times a user used tranquilizer  
**stimulant\_use** Percentage who used stimulants  
**stimulant\_freq** Median number of times a user used stimulants  
**meth\_use** Percentage who used meth  
**meth\_freq** Median number of times a user used meth  
**sedative\_use** Percentage who used sedatives  
**sedative\_freq** Median number of times a user used sedatives

### Source

National Survey on Drug Use and Health from the Substance Abuse and Mental Health Data Archive <https://www.icpsr.umich.edu/icpsrweb/content/SAMHDA/index.html>.

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
use <- drug_use %>%
  select(age, n, ends_with("_use")) %>%
  pivot_longer(-c(age, n), names_to = "drug", values_to = "use") %>%
  mutate(drug = str_sub(drug, start=1, end=-5))
freq <- drug_use %>%
  select(age, n, ends_with("_freq")) %>%
  pivot_longer(-c(age, n), names_to = "drug", values_to = "freq") %>%
  mutate(drug = str_sub(drug, start=1, end=-6))
drug_use_tidy <- left_join(x=use, y=freq, by = c("age", "n", "drug")) %>%
  arrange(age)
```

---

`elasticity_by_district`*Political Elasticity Scores*

---

**Description**

This folder contains the data behind the story 'Election Update: The House Districts That Swing The Most (And Least) With The National Mood' <https://fivethirtyeight.com/features/election-update-the-house-districts-that-swing-the-most-and-least-with-the-national-mood/>

**Usage**`elasticity_by_district`**Format**

A dataset with 435 rows representing congressional districts and 2 variables

**district** congressional district

**pvi\_538** pvi

**Note**

The original dataset only has 2 columns: "district" and "elasticity". I separated the "district" columns into two. For example, in row 1 of the dataset, the original "district" = "MI-5", and I separated it into "state" = "Michigan" and "district\_number" = "5". In addition, I used the full names for all states instead of abbreviations.

**Source**

An elasticity score measures how sensitive a state or district it is to changes in the national political environment.

**See Also**

[elasticity\\_by\\_state](#)

---

elasticity\_by\_state     *Political Elasticity Scores*

---

**Description**

This folder contains the data behind the story 'Election Update: The House Districts That Swing The Most (And Least) With The National Mood' <https://fivethirtyeight.com/features/election-update-the-house-districts-that-swing-the-most-and-least-with-the-national-mood/>

**Usage**

elasticity\_by\_state

**Format**

A dataset with 435 rows representing each state and the District of Columbia and 2 variables

**state** state

**pvi\_538** pvi

**Note**

I used the full names for all states instead of abbreviations.

**Source**

An elasticity score measures how sensitive a state or district it is to changes in the national political environment.

**See Also**

[elasticity\\_by\\_district](#)

---

elo\_blatter

*Blatter's Reign At FIFA Hasn't Helped Soccer's Poor*

---

**Description**

The raw data behind the story "Blatter's Reign At FIFA Hasn't Helped Soccer's Poor" <https://fivethirtyeight.com/features/blatters-reign-at-fifa-hasnt-helped-soccers-poor/>.

**Usage**

elo\_blatter

**Format**

A data frame with 191 rows representing countries and 5 variables:

**country** FIFA member country  
**elo98** The team's Elo in 1998  
**elo15** The team's Elo in 2015  
**confederation** Confederation to which country belongs  
**gdp06** The country's purchasing power parity GDP as of 2006  
**popu06** The country's 2006 population  
**gdp\_source** Source for gdp06  
**popu\_source** Source for popu06

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/elo-blatter>.

---

endorsements

*Pols And Polls Say The Same Thing: Jeb Bush Is A Weak Front-Runner*

---

**Description**

The raw data behind the story "Pols And Polls Say The Same Thing: Jeb Bush Is A Weak Front-Runner" <https://fivethirtyeight.com/features/pols-and-polls-say-the-same-thing-jeb-bush-is-a-weak-f>

This data includes something we call "endorsement points," an attempt to quantify the importance of endorsements by weighting each one according to the position held by the endorser: 10 points for each governor, 5 points for each senator and 1 point for each representative

**Usage**

endorsements

**Format**

A data frame with 109 rows representing candidates and 9 variables:

**year** Election year  
**party** Political party  
**candidate** Candidate running in primary  
**endorsement\_points** Weighted endorsements through June 30th of the year before the primary  
**percentage\_endorsement\_points** Percentage of total weighted endorsement points for the candidate's political party through June 30th of the year before the primary  
**money\_raised** Money raised through June 30th of the year before the primary  
**percentage\_of\_money** Percentage of total money raised by the candidate's political party through June 30th of the year before the primary  
**primary\_vote\_percentage** Percentage of votes won in the primary  
**won\_primary** Did the candidate win the primary?

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/endorsements-june-30>

---

endorsements_2020	<i>The 2020 Endorsement Primary - Which Democratic candidates are receiving the most support from prominent members of their party?</i>
-------------------	---

---

**Description**

The raw data behind the story "The 2020 Endorsement Primary - Which Democratic candidates are receiving the most support from prominent members of their party?" <https://projects.fivethirtyeight.com/2020-endorsements/democratic-primary/>.

**Usage**

endorsements\_2020

**Format**

A data frame with 1000 rows representing endorsements and 13 variables:

**date** date of the endorsement  
**position** position of the endorser  
**city** city of the endorser  
**state** state of the endorser  
**endorser** name of the endorser  
**endorsee** name of the endorsee  
**endorser\_party** party of the endorser  
**source** source link of the endorsement  
**order** order of the endorsement  
**category** category of the endorsement  
**body** body of the endorsement  
**district** district  
**points** points the endorsement counts for

**Source**

2020 endorsement tracker. Methodology: <https://fivethirtyeight.com/methodology/how-our-presidential-endorsement-tracker-works/>

fandango

*Be Suspicious Of Online Movie Ratings, Especially Fandango's***Description**

The raw data behind the story "Be Suspicious Of Online Movie Ratings, Especially Fandango's" <https://fivethirtyeight.com/features/fandango-movies-ratings/>. contains every film that has a Rotten Tomatoes rating, a RT User rating, a Metacritic score, a Metacritic User score, and IMDb score, and at least 30 fan reviews on Fandango.

**Usage**

fandango

**Format**

A data frame with 146 rows representing movies and 23 variables:

**film** The film in question

**year** Year of film

**rottentomatoes** The Rotten Tomatoes Tomatometer score for the film

**rottentomatoes\_user** The Rotten Tomatoes user score for the film

**metacritic** The Metacritic critic score for the film

**metacritic\_user** The Metacritic user score for the film

**imdb** The IMDb user score for the film

**fandango\_stars** The number of stars the film had on its Fandango movie page

**fandango\_ratingvalue** The Fandango ratingValue for the film, as pulled from the HTML of each page. This is the actual average score the movie obtained.

**rt\_norm** The Rotten Tomatoes Tomatometer score for the film , normalized to a 0 to 5 point system

**rt\_user\_norm** The Rotten Tomatoes user score for the film , normalized to a 0 to 5 point system

**metacritic\_norm** The Metacritic critic score for the film, normalized to a 0 to 5 point system

**metacritic\_user\_norm** The Metacritic user score for the film, normalized to a 0 to 5 point system

**imdb\_norm** The IMDb user score for the film, normalized to a 0 to 5 point system

**rt\_norm\_round** The Rotten Tomatoes Tomatometer score for the film , normalized to a 0 to 5 point system and rounded to the nearest half-star

**rt\_user\_norm\_round** The Rotten Tomatoes user score for the film , normalized to a 0 to 5 point system and rounded to the nearest half-star

**metacritic\_norm\_round** The Metacritic critic score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star

**metacritic\_user\_norm\_round** The Metacritic user score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star

- imdb\_norm\_round** The IMDb user score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star
- metacritic\_user\_vote\_count** The number of user votes the film had on Metacritic
- imdb\_user\_vote\_count** The number of user votes the film had on IMDb
- fandango\_votes** The number of user votes the film had on Fandango
- fandango\_difference** The difference between the presented Fandango\_Stars and the actual Fandango\_Ratingvalue

### Source

The data from Fandango was pulled on Aug. 24, 2015.

---

fifa_audience	<i>How To Break FIFA</i>
---------------	--------------------------

---

### Description

The raw data behind the story "How To Break FIFA" <https://fivethirtyeight.com/features/how-to-break-fifa/>.

### Usage

fifa\_audience

### Format

A data frame with 3652 rows representing guests and 6 variables:

- country** FIFA member country
- confederation** Confederation to which country belongs
- population\_share** Country's share of global population (percentage)
- tv\_audience\_share** Country's share of global world cup TV Audience (percentage)
- gdp\_weighted\_share** Country's GDP-weighted audience share (percentage)

### Source

See <https://github.com/fivethirtyeight/data/tree/master/fifa>

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
fifa_audience_tidy <- fifa_audience %>%
  pivot_longer(-c(country, confederation),
    names_to = "type", values_to = "share") %>%
  mutate(type = str_sub(type, start=1, end=-7)) %>%
  arrange(country)
```

---

 fight\_songs

*Our Guide To The Exuberant Nonsense of College Fight Songs*


---

**Description**

The data behind the story "Our Guide To The Exuberant Nonsense Of College Fight Songs" <https://projects.fivethirtyeight.com/college-fight-song-lyrics/>.

**Usage**

fight\_songs

**Format**

A data frame with 65 rows representing college fight songs, and 23 variables:

**school** school name  
**conference** school college football conference  
**song\_name** song title  
**writers** song author(s)  
**year** year the song was written; some years are unknown  
**student\_writer** TRUE if song was written by a student, FALSE if not  
**official\_song** TRUE if song is an official fight song according to the university, FALSE if not  
**contest** TRUE if song was chosen as part of a contest, FALSE if not  
**bpm** beats per minute  
**sec\_duration** duration of the song in seconds  
**fight** TRUE if song says 'fight', FALSE if not  
**num\_fights** number of time song says 'fight'  
**victory** TRUE if song says 'victory', FALSE if not  
**win\_won** TRUE if song says 'win' or 'won', FALSE if not  
**victory\_win\_won** TRUE if song says 'victory', 'win', or 'won'  
**rah** TRUE if song says 'rah', FALSE if not  
**nonsense** TRUE if song uses nonsense syllables, FALSE if not  
**colors** TRUE if song mentions school colors, FALSE if not  
**men** TRUE if song refers to a group of men, boys, sons, etc., FALSE if not  
**opponents** TRUE if song mentions opponents, FALSE if not  
**spelling** TRUE if song spells something out, FALSE if not  
**trope\_count** total number of tropes in song  
**spotify\_id** Spotify id for song

**Source**

Spotify <https://www.spotify.com/us/>

---

fivethirtyeight	<i>fivethirtyeight: Data and Code Behind the Stories and Interactives at 'FiveThirtyEight'</i>
-----------------	--

---

### Description

An R package that provides access to the code and data sets published by FiveThirtyEight <https://github.com/fivethirtyeight/data>. Note that while we received guidance from editors at 538, this package is not officially published by 538. You can explore all datasets here: <https://fivethirtyeight-r.netlify.app/articles/fivethirtyeight.html>

### Examples

```
# Example usage:
library(fivethirtyeight)
head(bechdel)

# All information about any data set can be found in the help file:
?bechdel

# To view a list of all data sets:
data(package = "fivethirtyeight")
```

---

flying	<i>41 Percent Of Fliers Think You're Rude If You Recline Your Seat</i>
--------	--

---

### Description

The raw data behind the story "41 Percent Of Fliers Think You're Rude If You Recline Your Seat" <https://fivethirtyeight.com/features/airplane-etiquette-recline-seat/>.

### Usage

```
flying
```

### Format

A data frame with 1040 rows representing respondents and 27 variables:

```
respondent_id RespondentID
gender Gender
age Age
height Height
children_under_18 Do you have any children under 18?
household_income Household income bracket
```

- education** Education Level
- location** Location (census region)
- frequency** How often do you travel by plane?
- recline\_frequency** Do you ever recline your seat when you fly?
- recline\_obligation** Under normal circumstances, does a person who reclines their seat during a flight have any obligation to the person sitting behind them?
- recline\_rude** Is it rude to recline your seat on a plane?
- recline\_eliminate** Given the opportunity, would you eliminate the possibility of reclining seats on planes entirely?
- switch\_seats\_friends** Is it rude to ask someone to switch seats with you in order to be closer to friends?
- switch\_seats\_family** Is it rude to ask someone to switch seats with you in order to be closer to family?
- wake\_up\_bathroom** Is it rude to wake a passenger up if you are trying to go to the bathroom?
- wake\_up\_walk** Is it rude to wake a passenger up if you are trying to walk around?
- baby** In general, is it rude to bring a baby on a plane?
- unruly\_child** In general, is it rude to knowingly bring unruly children on a plane?
- two\_arm\_rests** In a row of three seats, who should get to use the two arm rests?
- middle\_arm\_rest** In a row of two seats, who should get to use the middle arm rest?
- shade** Who should have control over the window shade?
- unsold\_seat** Is it rude to move to an unsold seat on a plane?
- talk\_stranger** Generally speaking, is it rude to say more than a few words to the stranger sitting next to you on a plane?
- get\_up** On a 6 hour flight from NYC to LA, how many times is it acceptable to get up if you're not in an aisle seat?
- electronics** Have you ever used personal electronics during take off or landing in violation of a flight attendant's direction?
- smoked** Have you ever smoked a cigarette in an airplane bathroom when it was against the rules?

## Source

SurveyMonkey survey

---

 food\_world\_cup

*The FiveThirtyEight International Food Association's 2014 World Cup*


---

### Description

The raw data behind the story "The FiveThirtyEight International Food Association's 2014 World Cup" <https://fivethirtyeight.com/features/the-fivethirtyeight-international-food-associations-2014-> For all the countries below, the response to the following question is presented: "Please rate how much you like the traditional cuisine of X"

- 5: I love this country's traditional cuisine. I think it's one of the best in the world.
- 4: I like this country's traditional cuisine. I think it's considerably above average.
- 3: I'm OK with this county's traditional cuisine. I think it's about average.
- 2: I dislike this country's traditional cuisine. I think it's considerably below average.
- 1: I hate this country's traditional cuisine. I think it's one of the worst in the world.
- N/A: I'm unfamiliar with this country's traditional cuisine.

### Usage

food\_world\_cup

### Format

A data frame with 1373 rows representing respondents and 48 variables:

**respondent\_id** Respondent ID

**knowledge** Generally speaking, how would you rate your level of knowledge of cuisines from different parts of the world?

**interest** How much, if at all, are you interested in cuisines from different parts of the world?

**gender** Gender

**age** Age

**household\_income** Household income bracket

**education** Education Level

**location** Location (census region)

**algeria** Cuisine of Algeria

**argentina** Cuisine of Argentina

**australia** Cuisine of Australia

**belgium** Cuisine of Belgium

**bosnia\_and\_herzegovina** Cuisine of Bosnia & Herzegovina

**brazil** Cuisine of Brazil

**cameroon** Cuisine of Cameroon

**chile** Cuisine of Chile  
**china** Cuisine of China  
**colombia** Cuisine of Colombia  
**costa\_rica** Cuisine of Costa Rica  
**croatia** Cuisine of Croatia  
**cuba** Cuisine of Cuba  
**ecuador** Cuisine of Ecuador  
**england** Cuisine of England  
**ethiopia** Cuisine of Ethiopia  
**france** Cuisine of France  
**germany** Cuisine of Germany  
**ghana** Cuisine of Ghana  
**greece** Cuisine of Greece  
**honduras** Cuisine of Honduras  
**india** Cuisine of India  
**iran** Cuisine of Iran  
**ireland** Cuisine of Ireland  
**italy** Cuisine of Italy  
**ivory\_coast** Cuisine of Ivory Coast  
**japan** Cuisine of Japan  
**mexico** Cuisine of Mexico  
**nigeria** Cuisine of Nigeria  
**portugal** Cuisine of Portugal  
**russia** Cuisine of Russia  
**south\_korea** Cuisine of South Korea  
**spain** Cuisine of Spain  
**switzerland** Cuisine of Switzerland  
**thailand** Cuisine of Thailand  
**the\_netherlands** Cuisine of the Netherlands  
**turkey** Cuisine of Turkey  
**united\_states** Cuisine of the United States  
**uruguay** Cuisine of Uruguay  
**vietnam** Cuisine of Vietnam

**See Also**

See <https://github.com/fivethirtyeight/data/tree/master/food-world-cup>

---

forecast\_results\_2018 *How FiveThirtyEight's 2018 Midterm Forecasts Did*

---

### Description

The raw data behind the story 'How FiveThirtyEight's 2018 Midterm Forecasts Did' <https://fivethirtyeight.com/features/how-fivethirtyeights-2018-midterm-forecasts-did/>

### Usage

```
forecast_results_2018
```

### Format

A dataframe with 1518 rows representing forecast results (as of December 3, 2018) and 11 variables:

**cycle** cycle of the election

**branch** branch of the election

**race** election forecast for the gubernatorial race

**forecastdate** the date of the forecast

**version** version of the election forecast

**dem\_win\_prob** the probability of winning for the Democrat

**rep\_win\_prob** the probability of winning for the Republican

**category** the predicted political affiliation of the forecast

**democrat\_won** whether the Democrat won

**republican\_won** whether the Republican won

**uncalled** if a race was uncalled

### Source

FiveThirtyEight's 2018 House Forecast <https://projects.fivethirtyeight.com/2018-midterm-election-forecast-house/>

---

foul_balls	<i>We Watched 906 Foul Balls To Find Out Where The Most Dangerous Ones</i>
------------	--

---

## Description

The raw data behind the story "We Watched 906 Foul Balls To Find Out Where The Most Dangerous Ones" <https://fivethirtyeight.com/features/we-watched-906-foul-balls-to-find-out-where-the-most-dar>

## Usage

foul\_balls

## Format

A data frame with 906 rows representing foul balls and 7 variables:

**matchup** the two teams that played

**game\_date** date of the most foul heavy day at each stadium

**type\_of\_hit** fly, grounder, line drive, popup, batter hits self

**exit\_velocity** recorded velocity of each hit

**predicted\_zone** zone predicted the foul ball would land in gauging angles

**camera\_zone** actual zone the ball landed it confirmed by camera angles

**used\_zone** zone used for analysis

## Details

Information on the Zones from the 538 original article: Zones 1, 2 and 3 are the areas behind home plate and the dugouts. Zones 4 and 5 make up most of the foul territory outside the baselines up until the foul pole. Zones 6 and 7 include the areas beyond the foul poles.

## Source

Baseball Savant <https://baseballsavant.mlb.com/>.

---

generic\_polllist      *Congress Generic Ballot Polls*

---

### Description

The raw data behind the story "Are Democrats Winning The Race For Congress?" <https://projects.fivethirtyeight.com/congress-generic-ballot-polls/>.

### Usage

generic\_polllist

### Format

A data frame with 934 rows representing polls and 21 variables:

**subgroup** No description provided.

**modeldate** No description provided.

**startdate** Start date of the poll.

**enddate** End date of the poll.

**pollster** The organization that conducted the poll (rather than the organization that paid for or sponsored it)

**grade** No description provided.

**samplesize** No description provided.

**population** A = ALL ADULTS, RV = REGISTERED VOTERS, LV = LIKELY VOTERS, V = VOTERS

**weight** No description provided.

**influence** No description provided.

**dem** No description provided.

**rep** No description provided.

**adjusted\_dem** No description provided.

**adjusted\_rep** No description provided.

**multiversions** No description provided.

**tracking** No description provided.

**url** No description provided.

**poll\_id** No description provided.

**question\_id** No description provided.

**createddate** No description provided.

**timestamp** No description provided.

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/congress-generic-ballot/README.md>

**See Also**

[generic\\_topline](#)

---

generic_topline	<i>Congress Generic Ballot Polls</i>
-----------------	--------------------------------------

---

**Description**

The raw data behind the story "Are Democrats Winning The Race For Congress?" <https://projects.fivethirtyeight.com/congress-generic-ballot-polls/>.

**Usage**

```
generic_topline
```

**Format**

A data frame with 751 rows representing polls and 9 variables:

**subgroup** No description provided.  
**modeldate** No description provided.  
**dem\_estimate** No description provided.  
**dem\_hi** No description provided.  
**dem\_lo** No description provided.  
**rep\_estimate** No description provided.  
**rep\_hi** No description provided.  
**rep\_lo** No description provided.  
**timestamp** No description provided.

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/congress-generic-ballot/README.md>

**See Also**

[generic\\_polllist](#)

---

google_trends	<i>The Media Really Started Paying Attention to Puerto Rico When Trump Did</i>
---------------	--

---

### Description

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/> Google Trends Data.

### Usage

```
google_trends
```

### Format

A data frame with 37 rows representing dates and 5 variables:

**date** Date

**hurricane\_harvey\_us** US Google search interest on the specified date for Hurricane Harvey

**hurricane\_irma\_us** US Google search interest on the specified date for Hurricane Irma

**hurricane\_maria\_us** US Google search interest on the specified date for Hurricane Maria

**hurricane\_jose\_us** US Google search interest on the specified date for Hurricane Jose

### Details

Google search interest is measured in search term popularity relative to peak popularity in the given region and time period (with 100 as peak popularity)

### Source

Google Trends <https://trends.google.com/trends/>

### See Also

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#)

---

governor\_national\_forecast  
*2018 Governors Forecast*

---

### Description

The raw data behind the story 'Forecasting the races for governor' <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/governor/>

### Usage

```
governor_national_forecast
```

### Format

A dataframe with 150 rows representing national-level results of the classic, lite, and deluxe gubernatorial forecasts since Oct. 11, 2018. and 11 variables

**forecastdate** date of the forecast

**party** the party of the forecast

**model** the model of the forecast

**win\_probability** the probability of the corresponding party winning

**mean\_seats** the mean of the number of seats

**median\_seats** the median number of seats

**p10\_seats** the top 10 percentile of number of seats

**p90\_seats** the top 90 percentile of number of seats

**margin** unknown

**p10\_margin** the margin of p10\_seats

**p90\_margin** the margin of p90\_seats

### Note

The original dataset included a meaningless column called "state", and all variables under this column was "US". So this column was removed.

### Source

FiveThirtyEight's House, Senate And Governor Models Methodology: <https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/>

### See Also

[governor\\_state\\_forecast](#)

---

governor\_state\_forecast

*2018 Governors Forecast*

---

### Description

The raw data behind the story 'Forecasting the races for governor' <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/governor/>

### Usage

governor\_state\_forecast

### Format

A dataframe with 7743 rows representing state-level results of the classic, lite, and deluxe gubernatorial forecasts since Oct. 11, 2018. and 10 variables

**forecastdate** date of the forecast

**state** state of the forecast

**candidate** name of the candidate

**party** party of the candidate

**incumbent** whether the candidate is incumbent

**model** the model of the forecast

**win\_probability** the probability of the corresponding party winning

**voteshare** the voteshare of the corresponding party

**p10\_voteshare** the top 10 percentile of the voteshare

**p90\_voteshare** the top 00 percentile of the voteshare

### Note

The original dataset included two empty column "district" and "special", which were removed.

### Source

FiveThirtyEight's House, Senate And Governor Models Methodology: <https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/>

### See Also

[governor\\_national\\_forecast](#)

hate\_crimes

*Higher Rates Of Hate Crimes Are Tied To Income Inequality***Description**

The raw data behind the story "Higher Rates Of Hate Crimes Are Tied To Income Inequality"  
<https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/>.

**Usage**

hate\_crimes

**Format**

A data frame with 51 rows representing US states and DC and 13 variables:

**state** State name

**state\_abbrev** State abbreviation

**median\_house\_inc** Median household income, 2016

**share\_unemp\_seas** Share of the population that is unemployed (seasonally adjusted), Sept. 2016

**share\_pop\_metro** Share of the population that lives in metropolitan areas, 2015

**share\_pop\_hs** Share of adults 25 and older with a high-school degree, 2009

**share\_non\_citizen** Share of the population that are not U.S. citizens, 2015

**share\_white\_poverty** Share of white residents who are living in poverty, 2015

**gini\_index** Gini Index, 2015

**share\_non\_white** Share of the population that is not white, 2015

**share\_vote\_trump** Share of 2016 U.S. presidential voters who voted for Donald Trump

**hate\_crimes\_per\_100k\_splc** Hate crimes per 100,000 population, Southern Poverty Law Center, Nov. 9-18, 2016

**avg\_hatecrimes\_per\_100k\_fbi** Average annual hate crimes per 100,000 population, FBI, 2010-2015

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/hate-crimes>

**Examples**

```
library(ggplot2)
ggplot(hate_crimes, aes(x = share_vote_trump, y = hate_crimes_per_100k_splc)) +
  geom_text(aes(label = state_abbrev)) +
  geom_smooth(se = FALSE, method = "lm") +
  labs(x = "Proportion of votes for Donald Trump",
       y = "Hate crimes per 100k during Nov 9-18, 2016 (SPLC)",
       title = "Relationship between Trump support & hate crimes")
```

---

hiphop\_cand\_lyrics     *Hip-Hop Is Turning On Donald Trump*

---

**Description**

The raw data behind the story "Hip-Hop Is Turning On Donald Trump" <https://projects.fivethirtyeight.com/clinton-trump-hip-hop-lyrics/>.

**Usage**

hiphop\_cand\_lyrics

**Format**

A data frame with 377 rows representing hip-hop songs referencing POTUS candidates in 2016 and 8 variables:

**candidate** Candidate referenced  
**song** Song name  
**artist** Artist name  
**sentiment** Positive, negative or neutral  
**theme** Theme of lyric  
**album\_release\_date** Date of album release  
**line** Lyrics  
**url** Genius link

**Source**

Genius <https://genius.com/>

---

hist\_ncaa\_bball\_casts     *The NCAA Bracket: Checking Our Work*

---

**Description**

The raw data behind the story "The NCAA Bracket: Checking Our Work" <https://fivethirtyeight.com/features/the-ncaa-bracket-checking-our-work/>.

**Usage**

hist\_ncaa\_bball\_casts

**Format**

A data frame with 253 rows representing NCAA men's basketball tournament games and 6 variables:

**year**

**round**

**favorite**

**underdog**

**favorite\_prob**

**favorite\_win**

**Source**

See <https://fivethirtyeight.com/features/the-ncaa-bracket-checking-our-work/>

---

hist\_senate\_preds      *How The FiveThirtyEight Senate Forecast Model Works*

---

**Description**

The raw data behind the story "How The FiveThirtyEight Senate Forecast Model Works" <https://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/>.

**Usage**

hist\_senate\_preds

**Format**

A data frame with 207 rows representing US state elections and 5 variables:

**state** Election

**year** Year of election

**candidate** Last name

**forecast\_prob** Probability of winning election per FiveThirtyEight Election Day forecast

**result** 'Win' or 'Loss'

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/forecast-methodology>

---

house\_national\_forecast

*2018 House Forecast*

---

### Description

The raw data behind the story 'Forecasting the race for the House' <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/>

### Usage

```
house_national_forecast
```

### Format

A dataframe with 588 rows representing district-level results of the classic, lite, and deluxe house forecasts since 2018/08/01 and 11 variables.

**forecastdate** date of the forecast

**party** the party of the forecast

**model** the model of the forecast

**win\_probability** the probability of the corresponding party winning

**mean\_seats** the mean of the number of seats

**median\_seats** the median number of seats

**p10\_seats** the top 10 percentile of number of seats

**p90\_seats** the top 90 percentile of number of seats

**margin** unknown

**p10\_margin** the margin of p10\_seats

**p90\_margin** the margin of p90\_seats

### Note

The original dataset included a meaningless column called "state", and all variables under this column was "US". So this column was removed.

### Source

FiveThirtyEight's House, Senate And Governor Models Methodology: <https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/>

### See Also

[house\\_district\\_forecast](#)

---

impeachment\_polls      *Do Americans Support Impeaching Trump?*

---

### Description

Raw data behind this story "Do Americans Support Impeaching Trump?" <https://projects.fivethirtyeight.com/impeachment-polls/>

### Usage

impeachment\_polls

### Format

A data frame with 388 rows of polling data and 24 variables:

**start** Poll start date, the first date responses were collected

**end** Poll end date, the last date responses were collected

**pollster** entity/organization that created poll, collected and published data

**sponsor** sponsor of pollster

**sample\_size** number of respondents for each

**pop** categorical variable with 3 categories: a, rv, lv – value unknown

**tracking** true/false logical – value unknown

**text** poll question

**category** category of poll question with 5 categories: impeach and remove, begin proceedings, begin inquiry, reasons, impeach

**include** yes/no logical – value unknown

**yes** Percent of respondents in sample who answered "Yes" to the poll question

**no** Percent of respondents in sample who answered "No" to the poll question

**unsure** Percent of respondents in sample who did not answer "Yes" or "No" to the poll question

**rep\_sample** number of Republican respondents in sample

**rep\_yes** Percent of Republican respondents who answered "yes"

**rep\_no** Percent of Republican respondents who answered "no"

**dem\_sample** number of Democrat respondents in sample

**dem\_yes** Percent of Democrat respondents who answered "yes"

**dem\_no** Percent of Democrat respondents who answered "no"

**ind\_sample** number of Independent respondents in sample

**ind\_yes** Percent of Independent respondents who answered "yes"

**ind\_no** Percent of Independent respondents who answered "no"

**url** URL links to poll websites

**notes** any notes relating to polls in sample

**Source**

data from <https://github.com/fivethirtyeight/data/tree/master/impeachment-polls>.

---

librarians

*Where Are America's Librarians?*

---

**Description**

The raw data behind the story "Where Are America's Librarians?" <https://fivethirtyeight.com/features/where-are-americas-librarians/>.

**Usage**

```
librarians
```

**Format**

A data frame with 371 rows representing areas in the US and 9 variables:

**prim\_state**

**area\_name**

**tot\_emp**

**emp\_prse**

**jobs\_1000**

**loc\_quotient**

**mor**

**high\_emp**

**low\_emp**

**Source**

Bureau of Labor Statistics

---

love_actually_adj	<i>The Definitive Analysis Of 'Love Actually,' The Greatest Christmas Movie Of Our Time</i>
-------------------	---

---

### Description

The raw data behind the story "The Definitive Analysis Of 'Love Actually,' The Greatest Christmas Movie Of Our Time" <https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-l>

The adjacency matrix of which actors appear in the same scene together.

### Usage

```
love_actually_adj
```

### Format

A data frame with 14 rows representing actors and 15 variables:

**actors**

**bill\_nighy**

**keira\_knightley**

**andrew\_lincoln**

**hugh\_grant**

**colin\_firth**

**alan\_rickman**

**heike\_makatsch**

**laura\_linney**

**emma\_thompson**

**liam\_neeson**

**kris\_marshall**

**abdul\_salis**

**martin\_freeman**

**rowan\_atkinson**

### See Also

[love\\_actually\\_appearance.](#)

---

love\_actually\_appearance

*The Definitive Analysis Of 'Love Actually,' The Greatest Christmas Movie Of Our Time*

---

## Description

The raw data behind the story "The Definitive Analysis Of 'Love Actually,' The Greatest Christmas

Movie Of Our Time" <https://fivethirtyeight.com/features/the-definitive-analysis-of-love-actually-the->

A table of the central actors in "Love Actually" and which scenes they appear in.

## Usage

love\_actually\_appearance

## Format

A data frame with 71 rows representing scenes and 15 variables:

**scenes**

**bill\_nighy**

**keira\_knightley**

**andrew\_lincoln**

**hugh\_grant**

**colin\_firth**

**alan\_rickman**

**heike\_makatsch**

**laura\_linney**

**emma\_thompson**

**liam\_neeson**

**kris\_marshall**

**abdul\_salis**

**martin\_freeman**

**rowan\_atkinson**

## See Also

[love\\_actually\\_adj.](#)

**Examples**

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
library(stringr)
love_actually_appearance_tidy <- love_actually_appearance %>%
  pivot_longer(-scenes, names_to = "actor", values_to = "appears") %>%
  arrange(scenes)
```

---

mad\_men

*"Mad Men" Is Ending. What's Next For The Cast?*


---

**Description**

The raw data behind the story "'Mad Men' Is Ending. What's Next For The Cast?" <https://fivethirtyeight.com/features/mad-men-is-ending-whats-next-for-the-cast/>.

**Usage**

mad\_men

**Format**

A data frame with 248 rows representing performers on TV shows and 15 variables:

**performer** The name of the actor, according to IMDb. This is not a unique identifier - two performers appeared in more than one program

**show** The television show where this actor appeared in more than half the episodes

**show\_start** The year the television show began

**show\_end** The year the television show ended, "PRESENT" if the show remains on the air as of May 10.

**status** Why the actor is no longer on the program: "END" if the show has concluded, "LEFT" if the show remains on the air.

**charend** The year the character left the show. Equal to "Show End" if the performer stayed on until the final season.

**years\_since** 2015 minus CharEnd

**num\_lead** The number of leading roles in films the performer has appeared in since and including "CharEnd", according to OpusData

**num\_support** The number of leading roles in films the performer has appeared in since and including "CharEnd", according to OpusData

**num\_shows** The number of seasons of television of which the performer appeared in at least half the episodes since and including "CharEnd", according to OpusData

**score** #LEAD + #Shows + 0.25\*(#SUPPORT)

**score\_div\_y** "Score" divided by "Years Since"

**lead\_notes** The list of films counted in #LEAD

**support\_notes** The list of films counted in #SUPPORT

**show\_notes** The seasons of shows counted in #Shows

**Source**

IMDB <https://imdb.com>

---

male\_flight\_attend      *Dear Mona, How Many Flight Attendants Are Men?*

---

**Description**

The raw data behind the story "Dear Mona, How Many Flight Attendants Are Men?" <https://fivethirtyeight.com/features/dear-mona-how-many-flight-attendants-are-men/>.

**Usage**

male\_flight\_attend

**Format**

A data frame with 320 rows representing job categories and 2 variables:

**job\_category** Category of job

**percentage\_male** Percentage of workforce that are male

**Source**

IPUMS 2012 <https://usa.ipums.org/usa/>

---

masculinity\_survey      *Masculinity Survey*

---

**Description**

This folder contains the data behind the story: "What Do Men Think It Means To Be A Man?" <https://fivethirtyeight.com/features/what-do-men-think-it-means-to-be-a-man/>

**Usage**

masculinity\_survey

**Format**

A dataset with 189 rows representing answers and 12 variables:

**question** the survey question

**response** the survey response

**overall** the ratio of overall participants who selected this response

**age\_18\_34** the ratio of participants age 18 to 34 who selected this response

**age\_35\_64** the ratio of participants age 35 to 64 who selected this response

**age\_65\_over** the ratio of participants age 65 or over who selected this response

**white\_yes** the ratio of overall white participants who selected this response

**white\_no** the ratio of overall non-white participants who selected this response

**children\_yes** the ratio of participants who have child(ren) who selected this response

**children\_no** the ratio of participants who do not have children who selected this response

**straight\_yes** the ratio of straight participants who selected this response

**straight\_no** the ratio of non-straight participants who selected this response

**Note**

The original ‘masculinity-survey.csv’ contains the results of a survey of 1,615 adult men conducted by SurveyMonkey in partnership with FiveThirtyEight and WNYC Studios from May 10-22, 2018. The modeled error estimate for this survey is plus or minus 2.5 percentage points. The percentages have been weighted for age, race, education, and geography using the Census Bureau’s American Community Survey to reflect the demographic composition of the United States age 18 and over. Crosstabs with less than 100 respondents have been left blank because responses would not be statistically significant. I made heavy editions in Excel to make the dataset easily usable in R.

**Source**

The original survey responses and original datasets can be found here: <https://github.com/fivethirtyeight/data/tree/master/masculinity-survey>

**Examples**

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(stringr)

# Data wrangling
masculinity_tidy <- masculinity_survey %>%
  # Narrow down rows to those pertaining to first question of survey:
  filter(question == 'In general, how masculine or "manly" do you feel?') %>%
  # Eliminate columns not relating to sexual orientation:
  select(-c(age_18_34, age_35_64, age_65_over, white_yes, white_no, children_yes,
            children_no, overall)) %>%
  # Convert data frame to tidy data (long) format:
  pivot_longer(-c(question, response), names_to = "sexuality", values_to = "ratio_by_sexuality")
```

```
# Visualize results
ggplot(data = masculinity_tidy, aes(x = response, y = ratio_by_sexuality, fill = sexuality)) +
  geom_bar(stat="identity", position = 'dodge') +
  labs(x = "Response", y = "Proportion", labs = "Sexuality",
       title = "In general, how masculine or 'manly' do you feel?")
```

---

mediacloud\_hurricanes *The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

## Description

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/> Mediacloud Hurricanes Data.

## Usage

```
mediacloud_hurricanes
```

## Format

A data frame with 38 rows representing dates and 5 variables:

**date** Date

**harvey** The number of sentences in online news which mention Hurricane Harvey on the specified date

**irma** The number of sentences in online news which mention Hurricane Irma

**maria** The number of sentences in online news which mention Hurricane Maria

**jose** The number of sentences in online news which mention Hurricane Jose

## Source

Mediacloud <https://mediacloud.org/>

## See Also

[mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#), [google\\_trends](#)

---

mediacloud\_online\_news

*The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

### Description

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/> Mediacloud Top Online News Data.

### Usage

mediacloud\_online\_news

### Format

A data frame with 49 rows representing media outlets and 2 variables:

**name** Name of media outlet source included in Media Cloud's "U.S. Top Online News" collection

**url** URL of corresponding media outlet source

### Source

Mediacloud <https://mediacloud.org/>

### See Also

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#), [google\\_trends](#)

---

mediacloud\_states

*The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

### Description

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/> Mediacloud States Data.

### Usage

mediacloud\_states

**Format**

A data frame with 51 rows representing dates and 4 variables:

**date** Date

**texas** The number of sentences in online news which mention Texas on the specified date

**puerto\_rico** The number of sentences in online news which mention Puerto Rico

**florida** The number of sentences in online news which mention Florida

**Source**

Mediacloud <https://mediacloud.org/>

**See Also**

[mediacloud\\_hurricanes](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#), [google\\_trends](#)

---

mediacloud_trump	<i>The Media Really Started Paying Attention to Puerto Rico When Trump Did</i>
------------------	--

---

**Description**

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico/>  
Mediacloud Trump Data.

**Usage**

```
mediacloud_trump
```

**Format**

A data frame with 51 rows representing dates and 7 variables:

**date** Date

**puerto\_rico** The number of headlines that mention Puerto Rico on the given date

**puerto\_rico\_and\_trump** The number of headlines that mention Puerto Rico and either President or Trump

**florida** The number of headlines that mention Florida

**florida\_and\_trump** The number of headlines that mention Florida and either President or Trump

**texas** The number of headlines that mention Texas

**texas\_and\_trump** The number of headlines that mention Texas and either President or Trump

**Source**

Mediacloud <https://mediacloud.org/>

**See Also**

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#), [google\\_trends](#)

---

media\_mentions\_2020    *2020 Presidential Candidates Media Mentions*

---

**Description**

The raw data behind the story "Beto O'Rourke Ignored Cable News - And It Ignored Him" <https://fivethirtyeight.com/features/beto-orourke-ignored-cable-news-and-it-ignored-him/>

**Usage**

media\_mentions\_cable

media\_mentions\_online

**Format**

2 dataframes about 2020 presidential candidate media mentions

An object of class `spec_tbl_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 954 rows and 6 columns.

**media\_mentions\_cable**

A data frame with 972 rows representing weeks of cable coverage and 7 variables:

**date** start date for the week of coverage

**name** candidate's name

**matched\_clips** number of 15-second clips in that week that mention the specified candidate

**all\_candidate\_clips** number of 15-second clips in that week that mention any candidates

**total\_clips** total number of 15-second clips that week across the three networks

**pct\_of\_all\_candidate\_clips** percentage of clips in which that specific candidate is mentioned out of all clips mentioning any candidate for that week (`matched_clips / all_candidate_clips`)

**query** query used for the GDELT Television API

**media\_mentions\_online**

A data frame with 954 rows representing weeks and 6 variables:

**date** start date for the week of coverage

**name** candidate's name

**matched\_stories** number of stories in that week that mention the specified candidate

**all\_candidate\_stories** number of stories in that week that mention any candidate

**pct\_of\_all\_candidate\_stories** percentage of stories in which that specific candidate is mentioned out of all stories mentioning any candidate for that week (matched\_stories / all\_candidate\_stories)

**query** query for Media Cloud

**Source**

The GDELT Television API <https://blog.gdeltproject.org/gdelt-2-0-television-api-debuts/>, which processes the data from the TV News Archive <https://archive.org/details/tv>.

Two collections in the Media Cloud <https://mediacloud.org/> database U.S. Top Online News <https://sources.mediacloud.org/#/collections/58722749> and U.S. Top Digital Native News <https://sources.mediacloud.org/#/collections/57078150>

---

mlb\_as\_play\_talent      *The Best MLB All-Star Teams Ever*

---

**Description**

The raw data behind the story "The Best MLB All-Star Teams Ever" <https://fivethirtyeight.com/features/the-best-mlb-all-star-teams-ever/>.

**Usage**

mlb\_as\_play\_talent

**Format**

A data frame with 3930 rows representing Major League Baseball players in given seasons and 15 variables:

**bbref\_id** Player's ID at Baseball-Reference.com

**yearid** The season in question

**gamenum** Order of All-Star Game for the season (in years w/ multiple ASGs; set to 0 when only 1 per year)

**gameid** Game ID at Baseball-Reference.com

**lgid** League of All-Star team

**startingpos** Position (according to baseball convention; 1=pitcher, 2=catcher, etc.) if starter

**off600** Estimate of offensive talent, in runs above league average per 600 plate appearances

- def600** Estimate of fielding talent, in runs above league average per 600 plate appearances
- pitch200** Estimate of pitching talent, in runs above league average per 200 innings pitched
- asg\_pa** Number of plate appearances in the All-Star Game itself
- asg\_ip** Number of innings pitched in the All-Star Game itself
- offper9innasg** Expected offensive runs added above average (from talent) based on PA in ASG, scaled to a 9-inning game
- defper9innasg** Expected defensive runs added above average (from talent) based on PA in ASG, scaled to a 9-inning game
- pitper9innasg** Expected pitching runs added above average (from talent) based on IP in ASG, scaled to a 9-inning game
- totper9innasg** Expected runs added above average (from talent) based on PA/IP in ASG, scaled to a 9-inning game

### Source

<https://www.baseball-reference.com/> , <http://chadwick-bureau.com>, Fangraphs

---

mlb\_as\_team\_talent      *The Best MLB All-Star Teams Ever*

---

### Description

The raw data behind the story "The Best MLB All-Star Teams Ever" <https://fivethirtyeight.com/features/the-best-mlb-all-star-teams-ever/>.

### Usage

mlb\_as\_team\_talent

### Format

A data frame with 172 rows representing Major League Baseball seasons and 16 variables:

- yearid** The season in question
- gamenum** Order of All-Star Game for the season (in years w/ multiple ASGs; set to 0 when only 1 per year)
- gameid** Game ID at Baseball-Reference.com
- lgid** League of All-Star team
- tm\_off\_talent** Total runs of offensive talent above average per game (36 plate appearances)
- tm\_def\_talent** Total runs of fielding talent above average per game (36 plate appearances)
- tm\_pit\_talent** Total runs of pitching talent above average per game (9 innings)
- mlb\_avg\_rpg** MLB average runs scored/game that season
- talent\_rspg** Expected runs scored per game based on talent (MLB R/G + team OFF talent)

**talent\_rapg** Expected runs allowed per game based on talent (MLB R/G - team DEF talent- team PIT talent)

**unadj\_pyth** Unadjusted Pythagorean talent rating;  $PYTH = (RSPG^{1.83}) / (RSPG^{1.83} + RAPG^{1.83})$

**timeline\_adj** Estimate of relative league quality where 2015 MLB = 1.00

**sos** Strength of schedule faced; adjusts an assumed .500 SOS downward based on timeline adjustment

**adj\_pyth** Adjusted Pythagorean record;  $= (SOS * unadj\_Pyth) / ((2 * unadj\_Pyth * SOS) - SOS - unadj\_Pyth + 1)$

**no\_1\_player** Best player according to combo of actual PA/IP and talent

**no\_2\_player** 2nd-best player according to combo of actual PA/IP and talent

### Source

<https://www.baseball-reference.com/> , <http://chadwick-bureau.com>, Fangraphs

---

mueller\_approval\_polls

*Both Parties Think The Mueller Report Was Fair. They Just Completely Disagree On What It Says.*

---

### Description

The raw data behind the story 'Both Parties Think The Mueller Report Was Fair. They Just Completely Disagree On What It Says.' <https://fivethirtyeight.com/features/both-parties-think-the-mueller-report/>

### Usage

mueller\_approval\_polls

### Format

A dataset with 65 rows representing every job approval poll of Robert Mueller that we could find from when Mueller was appointed as special council on May 17, 2017 through May 3, 2019 and 12 variables

**start** the start date of the poll

**end** the end date of the poll

**pollster** the name of the pollster

**sample\_size** the size of the poll sample

**population** unknown

**text** the text of the poll question

**approve** the number of approval in the poll

**disapprove** the number of disapproval in the poll

**unsure** the number of unsure in the poll

**approve\_(republican)** the number of approval from Republican

**approve\_(democrat)** the number of approval from Democrat

**url** the url of the poll

**Source**

Polls, Washington Post / ABC and Washington Post / Schar School Polls

---

murder\_2015\_final      *A Handful Of Cities Are Driving 2016's Rise In Murder*

---

**Description**

The raw data behind the story "A Handful Of Cities Are Driving 2016's Rise In Murder" <https://fivethirtyeight.com/features/a-handful-of-cities-are-driving-2016s-rise-in-murders/>.

**Usage**

murder\_2015\_final

**Format**

A data frame with 83 rows representing large US cities and 5 variables:

**city** Name of city

**state** Name of state

**murders\_2014** Total murders in 2014

**murders\_2015** Total murders in 2015

**change** 2015 - 2014

**Source**

Unknown

---

murder\_2016\_prelim      *A Handful Of Cities Are Driving 2016's Rise In Murder*

---

**Description**

The raw data behind the story "A Handful Of Cities Are Driving 2016's Rise In Murder" <https://fivethirtyeight.com/features/a-handful-of-cities-are-driving-2016s-rise-in-murders/>.

**Usage**

murder\_2016\_prelim

**Format**

A data frame with 79 rows representing large US cities and 7 variables:

**city** Name of city

**state** Name of state

**murders\_2015** Number of murders in 2015

**murders\_2016** Number of murder in 2016 (as of as\_of date)

**change** 2016 - 2015

**source** Source of data

**as\_of** 2016 murders up to this date

**Source**

Listed as source variable in dataset

---

nba\_draft\_2015

*Projecting The Top 50 Players In The 2015 NBA Draft Class*

---

**Description**

The raw data behind the story "Projecting The Top 50 Players In The 2015 NBA Draft Class"

<https://fivethirtyeight.com/features/projecting-the-top-50-players-in-the-2015-nba-draft-class/>.

An analysis using this data was contributed by G. Elliott Morris as a package vignette at <https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/NBA.html>.

**Usage**

nba\_draft\_2015

**Format**

A data frame with 1090 rows representing National Basketball Association players/prospects and 9 variables:

**player** Player name

**position** The player's position going into the draft

**id** The player's identification code

**draft\_year** The year the player was eligible for the NBA draft

**projected\_spm** The model's projected statistical plus/minus over years 2-5 of the player's NBA career

**superstar** Probability of becoming a superstar player (1 per draft, SPM  $\geq$  +3.3)

**starter** Probability of becoming a starting-caliber player (10 per draft, SPM  $\geq$  +0.5)

**role\_player** Probability of becoming a role player (25 per draft, SPM  $\geq$  -1.4)

**bust** Probability of becoming a bust (everyone else, SPM  $<$  -1.4)

**Source**

See <https://fivethirtyeight.com/features/projecting-the-top-50-players-in-the-2015-nba-draft-class/>

---

nba\_drayment

*A Better Way to Evaluate NBA Defense*

---

**Description**

The raw data behind the story "A Better Way to Evaluate NBA Defense" <https://fivethirtyeight.com/features/a-better-way-to-evaluate-nba-defense/>.

**Usage**

nba\_drayment

**Format**

A data frame with 3009 rows representing DRAYMOND ratings (Defensive Rating Accounting for Yielding Minimal Openness by Nearest Defender) for every player since the 2013-14 season with 4 variables:

**season** The second year of the season; for example, 2018-2019 season would be listed as 2019

**player** Name of the player

**possessions** Number of possessions a player during the season

**drayment** Defensive Rating Accounting for Yielding Minimal Openness by Nearest Defender

**Source**

see <https://github.com/fivethirtyeight/data/tree/master/nba-drayment>

---

nba\_elo

*NBA Elo Ratings*

---

**Description**

The raw data behind all nba predictions, including the story "The Complete History of the NBA" <https://projects.fivethirtyeight.com/complete-history-of-the-nba>

**Usage**

nba\_elo\_latest

**Format**

An object of class `spec_tbl_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 1230 rows and 24 columns.

**nba\_elo\_latest**

A data frame with 1230 rows representing game played during the most current season of the NBA, and 24 variables:

**date** Date

**season** the season in which the game was played

**neutral** True if the game was played on neutral territory, False if not

**playoff** True if the game was played in a playoff, False if not

**team1** name of first team

**team2** name of second team

**elo1\_pre** Team 1 Elo rating before game

**elo2\_pre** Team 2 Elo rating before game

**elo\_prob1** Team 1's probability of winning based on Elo rating

**elo\_prob2** Team 2's probability of winning based on Elo rating

**elo1\_post** Team 1 Elo rating after the game

**elo2\_post** Team 2 Elo rating after the game

**score1** the score of team 1

**score2** the score of team 2

---

nba\_tattoos

*Accurately Counting NBA Tattoos Isn't Easy, Even If You're Up Close*

---

**Description**

The raw data behind the story "Accurately Counting NBA Tattoos Isn't Easy, Even If You're Up Close" <https://fivethirtyeight.com/features/accurately-counting-nba-tattoos-isnt-easy-even-if-youre-up-close/>

**Usage**

nba\_tattoos

**Format**

A data frame with 636 rows representing National Basketball Association players and 2 variables:

**player\_name** Name of player

**tattoos** TRUE corresponds to player having tattoos, FALSE corresponds to not

**Source**

Ethan Swan <https://nbatattoos.tumblr.com/>

---

ncaa\_w\_bball\_tourney *The Rise And Fall Of Women's NCAA Tournament Dynasties*

---

### Description

The raw data behind the story 'The Rise And Fall Of Women's NCAA Tournament Dynasties'  
<https://fivethirtyeight.com/features/louisiana-tech-was-the-uconn-of-the-80s/>

### Usage

```
ncaa_w_bball_tourney
```

### Format

A dataset with 2092 rows representing every team that has participated in the NCAA Division I Women's Basketball Tournament since it began in 1982 and 19 variables

**year** the year of the game which the team participated in

**school** the school of the participating team

**seed** The '(OR)' seeding designation in 1983 notes the eight teams that played an opening-round game to become the No. 8 seed in each region.

**conference** the conference record of the team (if available)

**conf\_w** number of winning in conference record

**conf\_l** number of losses in conference record

**conf\_percent** percent of winning in conference record

**reg\_w** number of winning in regular-season record

**reg\_l** number of losses in regular-season record

**reg\_percent** percent of winning in regular-season record

**how\_qual** Whether the school qualified with an automatic bid (by winning its conference or conference tournament) or an at-large bid.

**first\_home\_game** Whether the school played its first-round tournament games on its home court.

**tourney\_w** number of winning in tournament record

**tourney\_l** number of losses in tournament record

**tourney\_finish** The round of the final game for each team. OR=opening-round loss (1983 only); 1st=first-round loss; 2nd=second-round loss; RSF=loss in the Sweet 16; RF=loss in the Elite Eight; NSF=loss in the national semifinals; N2nd=national runner-up; Champ=national champions

**full\_w** number of winning in full record

**full\_l** number of losses in full record

**full\_percent** percent of winning in full record

### Source

NCAA

---

nfltix\_div\_avgprice     *Who Goes To Meaningless NFL Games And Why?*

---

**Description**

The raw data behind the story "Who Goes To Meaningless NFL Games And Why?" <https://fivethirtyeight.com/features/who-goes-to-meaningless-nfl-games-and-why/>.

**Usage**

nfltix\_div\_avgprice

**Format**

A data frame with 108 rows representing National Football League games and 3 variables:

**event** NFL divisional game info

**division** NFL division

**avg\_tix\_price** Average ticket price

**Source**

StubHub

---

nfltix\_usa\_avg     *Who Goes To Meaningless NFL Games And Why?*

---

**Description**

The raw data behind the story "Who Goes To Meaningless NFL Games And Why?" <https://fivethirtyeight.com/features/who-goes-to-meaningless-nfl-games-and-why/>.

**Usage**

nfltix\_usa\_avg

**Format**

A data frame with 32 rows representing National Football League teams and 2 variables:

**team** Name of NFL team

**avg\_tix\_price** Average ticket price

**Source**

StubHub

---

nflwr_aging_curve	<i>The Football Hall Of Fame Has A Receiver Problem</i>
-------------------	---

---

**Description**

The raw data behind the story "The Football Hall Of Fame Has A Receiver Problem" <https://fivethirtyeight.com/features/the-football-hall-of-fame-has-a-receiver-problem/>.

**Usage**

nflwr\_aging\_curve

**Format**

A data frame with 24 rows representing National Football League wide receiver ages and 3 variables:

**age\_from** Beginning age

**age\_to** Ending age

**trypg\_change** Change in TRY per game from one age-year to next

**Source**

Unknown

---

nflwr_hist	<i>The Football Hall Of Fame Has A Receiver Problem</i>
------------	---

---

**Description**

The raw data behind the story "The Football Hall Of Fame Has A Receiver Problem" <https://fivethirtyeight.com/features/the-football-hall-of-fame-has-a-receiver-problem/>.

**Usage**

nflwr\_hist

**Format**

A data frame with 6496 rows representing National Football League wide receivers and 6 variables:

**pfr\_player\_id** Player identification code at <https://www.pro-football-reference.com/>

**player\_name** The player's name

**career\_try** Career True Receiving Yards

**career\_ranypa** Adjusted Net Yards Per Attempt (relative to average) of player's career teams, weighted by TRY w/ each team

**career\_wowy** The amount by which career\_ranypa exceeds what would be expected from his QBs' (age-adjusted) performance without the receiver

**bcs\_rating** The number of yards per game by which a player would outgain an average receiver on the same team, after adjusting for teammate quality and age

### Source

See <https://fivethirtyeight.com/features/the-football-hall-of-fame-has-a-receiver-problem/>

---

nfl\_fandom\_google

*How Every NFL Team's Fans Lean Politically*

---

### Description

The raw data behind the story "How Every NFL Team's Fans Lean Politically" <https://fivethirtyeight.com/features/how-every-nfl-teams-fans-lean-politically/>: Google Trends Data.

### Usage

nfl\_fandom\_google

### Format

a data frame with 207 rows representing designated market areas and 9 variables:

**dma** Designated Market Area

**nfl** The percentage of search traffic in the media market region related to the NFL over the past 5 years

**nba** The percentage of search traffic in the region related to the NBA over the past 5 years

**mlb** The percentage of search traffic in the region related to the MLB over the past 5 years

**nascar** The percentage of search traffic in the region related to NASCAR over the past 5 years

**cbb** The percentage of search traffic in the region related to the CBB over the past 5 years

**cfb** The percentage of search traffic in the region related to the CFB over the past 5 years

**trump\_2016\_vote** The percentage of voters in the region who voted for Trump in the 2016 Presidential Election

### Source

Google Trends <https://trends.google.com/trends/>.

### See Also

[nfl\\_fandom\\_surveymonkey](#)

**Examples**

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
nfl_fandom_google_tidy <- nfl_fandom_google %>%
  pivot_longer(-c("dma", "trump_2016_vote"),
    names_to = "sport", values_to = "search_traffic") %>%
  arrange(dma)
```

---

nfl\_fandom\_surveymonkey

*How Every NFL Team's Fans Lean Politically*


---

**Description**

The raw data behind the story "How Every NFL Team's Fans Lean Politically" <https://fivethirtyeight.com/features/how-every-nfl-teams-fans-lean-politically/>: SurveyMonkey Data.

**Usage**

```
nfl_fandom_surveymonkey
```

**Format**

a data frame with 33 rows representing teams and 25 variables:

**team** NFL team

**total\_respondents** Total number of poll respondents who ranked the given team in their top 3 favorites

**asian\_dem** Number of Asian, democrat poll respondents who ranked the given team in their top 3 favorites

**black\_dem** Number of Black, democrat poll respondents who ranked the given team in their top 3 favorites

**hispanic\_dem** Number of Hispanic, democrat poll respondents who ranked the given team in their top 3 favorites

**other\_dem** Number of democrat poll respondents who identified their race as "other" (not Asian, Black, Hispanic, or White) and ranked the given team in their top 3 favorites

**white\_dem** Number of White, democrat poll respondents who ranked the given team in their top 3 favorites

**total\_dem** Total number of democrat poll respondents who ranked the given team in their top 3 favorites

**asian\_ind** Number of Asian, independent poll respondents who ranked the given team in their top 3 favorites

**black\_ind** Number of Black, independent poll respondents who ranked the given team in their top 3 favorites

**hispanic\_ind** Number of Hispanic, independent poll respondents who ranked the given team in their top 3 favorites

**other\_ind** Number of independent poll respondents who identified their race as "other" (not Asian, Black, Hispanic, or White) and ranked the given team in their top 3 favorites

**white\_ind** Number of White, independent poll respondents who ranked the given team in their top 3 favorites

**total\_ind** Total number of independent poll respondents who ranked the given team in their top 3 favorites

**asian\_gop** Number of Asian, republican poll respondents who ranked the given team in their top 3 favorites

**black\_gop** Number of Black, republican poll respondents who ranked the given team in their top 3 favorites

**hispanic\_gop** Number of Hispanic, republican poll respondents who ranked the given team in their top 3 favorites

**other\_gop** Number of republican poll respondents who identified their race as "other" (not Asian, Black, Hispanic, or White) and ranked the given team in their top 3 favorites

**white\_gop** Number of White, republican poll respondents who ranked the given team in their top 3 favorites

**total\_gop** Total number of republican poll respondents who ranked the given team in their top 3 favorites

**gop\_percent** Percent of fans (who ranked the team in their top 3 favorite NFL teams) who are republicans

**dem\_percent** Percent of fans who are democrats

**ind\_percent** Percent of fans who are independent

**white\_percent** Percent of fans who are White

**nonwhite\_percent** Percent of fans who are not White

### Source

See [https://github.com/fivethirtyeight/data/blob/master/nfl-fandom/NFL\\_fandom\\_data-surveymonkey.csv](https://github.com/fivethirtyeight/data/blob/master/nfl-fandom/NFL_fandom_data-surveymonkey.csv)

### See Also

[nfl\\_fandom\\_google](#)

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
nfl_fandom_surveymonkey_tidy <- nfl_fandom_surveymonkey %>%
  pivot_longer(-c("team", "total_respondents", "gop_percent", "dem_percent",
                 "ind_percent", "white_percent", "nonwhite_percent"),
              names_to = "race_party", values_to = "percent") %>%
  arrange(team)
```

---

nfl_fav_team	<i>The Rams Are Dead To Me, So I Answered 3,352 Questions To Find A New NFL Team</i>
--------------	--

---

### Description

The raw data behind the story "The Rams Are Dead To Me, So I Answered 3,352 Questions To Find A New NFL Team" <https://fivethirtyeight.com/features/the-rams-are-dead-to-me-so-i-answered-3352-questions-to-find-a-new-nfl-team/>

### Usage

nfl\_fav\_team

### Format

A data frame with 32 rows representing National Football League teams and 17 variables:

**team** Name of NFL team

**fan\_relations** Fan relations - Courtesy by players, coaches and front offices toward fans, and how well a team uses technology to reach them

**ownership** Ownership - Honesty; loyalty to core players and the community

**players** Players - Effort on the field, likability off it

**future\_wins** Future wins - Projected wins over next 5 seasons

**bandwagon** Bandwagon Factor - Are the team's next 5 years likely to be better than their previous 5?

**tradition** Tradition - Championships/division titles/wins in team's entire history

**bang\_buck** Bang for the buck - Wins per fan dollars spent

**behavior** Behavior - Suspensions by players on team since 2007, with extra weight to transgressions vs. women

**nyc\_prox** Proximity to New York City

**stlouis\_prox** Proximity to St. Louis

**afford** Affordability - Price of tickets, parking and concessions

**small\_market** Small Market - Size of market in terms of population, where smaller is better

**stadium\_exp** Stadium experience - Quality of venue; fan-friendliness of environment; frequency of game-day promotions

**coaching** Coaching - Strength of on-field leadership

**uniform** Uniform - Stylishness of uniform design, according to Uni Watch's Paul Lukas

**big\_market** Big Market - Size of market in terms of population, where bigger is better

### Source

<https://www.allourideas.org/nflteampickingsample>

---

nfl_suspensions	<i>The NFL's Uneven History Of Punishing Domestic Violence</i>
-----------------	--

---

**Description**

The raw data behind the story "The NFL's Uneven History Of Punishing Domestic Violence" <https://fivethirtyeight.com/features/nfl-domestic-violence-policy-suspensions/>.

**Usage**

nfl\_suspensions

**Format**

A data frame with 269 rows representing National Football League players and 7 variables:

**name** first initial.last name

**team** team at time of suspension

**games** number of games suspended (one regular season = 16 games)

**category** personal conduct, substance abuse, performance enhancing drugs or in-game violence

**description** description of suspension

**year** year of suspension

**source** news source

**Source**

[https://en.wikipedia.org/wiki/List\\_of\\_players\\_and\\_coaches\\_suspended\\_by\\_the\\_NFL](https://en.wikipedia.org/wiki/List_of_players_and_coaches_suspended_by_the_NFL), <https://www.sportrac.com/fines-tracker/nfl/suspensions/>

---

nutrition_pvalues	<i>You Can't Trust What You Read About Nutrition</i>
-------------------	--

---

**Description**

The raw data behind the story "You Can't Trust What You Read About Nutrition" <https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/>.

**Usage**

nutrition\_pvalues

**Format**

A data frame with 27716 rows representing Regression fits for p-hacking and 3 variables:

**food** Name of food (response/dependent variable)

**characteristic** Name of characteristic (predictor/independent variable)

**p\_values** P-value from regression fit

**Source**

See <https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/>

---

partisan\_lean\_district

*FiveThirtyEight's Partisan Lean*

---

**Description**

This directory contains the data for FiveThirtyEight's partisan lean, which is used in our [House] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house> [Senate] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate> and [Governor] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/governor/> forecasts.

**Usage**

partisan\_lean\_district

**Format**

A dataset with 435 rows representing votes and 4 variables

**state** the state of the vote

**district\_number** the district\_number of the vote

**pvi\_party** the party of the vote

**pvi\_amount** the Cook Partisan Voting Index of the vote

**Note**

The original dataset only has 2 columns: "district" and "pvi\_538". I separated each of the 2 columns into two. For example, in row 1 of the dataset, the original "district" = "AK-1", and I separated it into "state" = "Arkansas" and "district\_number" = "1"; the original "pvi\_538" = "R+15.21", and I separated it into "pvi\_party" = "R" and "pvi\_amount" = "15.21". In addition, I used the full names for all states instead of abbreviations.

**Source**

Partisan lean is the average difference between how a state or district votes and how the country votes overall, with 2016 presidential election results weighted 50 percent, 2012 presidential election results weighted 25 percent and results from elections for the state legislature weighted 25 percent.

**See Also**

[partisan\\_lean\\_state](#)

---

partisan\_lean\_state     *FiveThirtyEight's Partisan Lean*

---

**Description**

This directory contains the data for FiveThirtyEight's partisan lean, which is used in our [House] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house> [Senate] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate> and [Governor] <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/governor/> forecasts.

**Usage**

partisan\_lean\_state

**Format**

A dataset with 50 rows representing states and 3 variables

**state** the state

**pvi\_party** the party of the vote

**pvi\_amount** the Cook Partisan Voting Index of the vote

**Note**

The original dataset only has 2 columns: "state" and "pvi\_538". I separated the "pvi\_538" columns into two. For example, in row 1 of the dataset, the original "pvi\_538" = "R+27", and I separated it into "pvi\_party" = "R" and "pvi\_amount" = "27".

**Source**

Partisan lean is the average difference between how a state or district votes and how the country votes overall, with 2016 presidential election results weighted 50 percent, 2012 presidential election results weighted 25 percent and results from elections for the state legislature weighted 25 percent.

**See Also**

[partisan\\_lean\\_district](#)

---

police_deaths	<i>The Dallas Shooting Was Among The Deadliest For Police In U.S. History</i>
---------------	---

---

**Description**

The raw data behind the story "The Dallas Shooting Was Among The Deadliest For Police In U.S. History" <https://fivethirtyeight.com/features/the-dallas-shooting-was-among-the-deadliest-for-police>

**Usage**

police\_deaths

**Format**

A data frame with 22800 rows representing Police officers/dogs who lost their lives and 7 variables:

**person** Name of person/canine who died

**cause\_of\_death** Cause of death

**date** Date of event

**year** Year of event

**canine** TRUE if canine, FALSE if human

**dept\_name** Name of police department

**state** State of police department

**Source**

Officer Down Memorial Page <https://www.odmp.org/>

---

police_killings	<i>Where Police Have Killed Americans In 2015</i>
-----------------	---

---

**Description**

The raw data behind the story "Where Police Have Killed Americans In 2015" <https://fivethirtyeight.com/features/where-police-have-killed-americans-in-2015/>.

**Usage**

police\_killings

**Format**

A data frame with 467 rows representing People who died from interactions with police and 34 variables:

**name** Name of deceased  
**age** Age of deceased  
**gender** Gender of deceased  
**raceethnicity** Race/ethnicity of deceased  
**month** Month of killing  
**day** Day of incident  
**year** Year of incident  
**streetaddress** Address/intersection where incident occurred  
**city** City where incident occurred  
**state** State where incident occurred  
**latitude** Latitude, geocoded from address  
**longitude** Longitude, geocoded from address  
**state\_fp** State FIPS code  
**county\_fp** County FIPS code  
**tract\_ce** Tract ID code  
**geo\_id** Combined tract ID code  
**county\_id** Combined county ID code  
**namelsad** Tract description  
**lawenforcementagency** Agency involved in incident  
**cause** Cause of death  
**armed** How/whether deceased was armed  
**pop** Tract population  
**share\_white** Share of pop that is non-Hispanic white  
**share\_black** Share of pop that is black (alone, not in combination)  
**share\_hispanic** Share of pop that is Hispanic/Latino (any race)  
**p\_income** Tract-level median personal income  
**h\_income** Tract-level median household income  
**county\_income** County-level median household income  
**comp\_income** 'h\_income' / 'county\_income'  
**county\_bucket** Household income, quintile within county  
**nat\_bucket** Household income, quintile nationally  
**pov** Tract-level poverty rate (official)  
**urate** Tract-level unemployment rate  
**college** Share of 25+ pop with BA or higher

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/police-killings>

---

`police_locals`*Most Police Don't Live In The Cities They Serve*

---

### Description

The raw data behind the story "Most Police Don't Live In The Cities They Serve" <https://fivethirtyeight.com/features/most-police-dont-live-in-the-cities-they-serve/>.

### Usage

```
police_locals
```

### Format

A data frame with 75 rows representing cities and 8 variables:

**city** U.S. city

**force\_size** Number of police officers serving that city

**all** Percentage of the total police force that lives in the city

**white** Percentage of white (non-Hispanic) police officers who live in the city

**non\_white** Percentage of non-white police officers who live in the city

**black** Percentage of black police officers who live in the city

**hispanic** Percentage of Hispanic police officers who live in the city

**asian** Percentage of Asian police officers who live in the city

### Details

The dataset includes the cities with the 75 largest police forces, with the exception of Honolulu for which data is not available. All calculations are based on data from the U.S. Census.

The Census Bureau numbers are potentially going to differ from other counts for three reasons:

1. The census category for police officers also includes sheriffs, transit police and others who might not be under the same jurisdiction as a city's police department proper. The census category won't include private security officers.
2. The census data is estimated from 2006 to 2010; police forces may have changed in size since then.
3. There is always a margin of error in census numbers; they are estimates, not complete counts.

Note: Missing values means that there are fewer than 100 police officers of that race serving that city.

### Source

See <https://github.com/fivethirtyeight/data/tree/master/police-locals>

## Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
police_locals_tidy <- police_locals %>%
  pivot_longer(all:asian, names_to = "race", values_to = "perc_in")
```

---

pres_2016_trail	<i>The Last 10 Weeks Of 2016 Campaign Stops In One Handy Gif</i>
-----------------	--

---

## Description

The raw data behind the story "The Last 10 Weeks Of 2016 Campaign Stops In One Handy Gif"  
<https://fivethirtyeight.com/features/the-last-10-weeks-of-2016-campaign-stops-in-one-handy-gif/>.

## Usage

```
pres_2016_trail
```

## Format

A data frame with 177 rows representing 2016 Republican and Democratic candidate campaign trail stops and 5 variables:

**candidate** Clinton or Trump

**date** The date of the event

**location** The location of the event

**lat** Latitude of the event location

**lng** Longitude of the event location

## Source

<https://hillaryspeeches.com/>, <https://www.conservativedailynews.com/>

---

pres_commencement	<i>Sitting Presidents Give Way More Commencement Speeches Than They Used To</i>
-------------------	---

---

**Description**

The raw data behind the story "Sitting Presidents Give Way More Commencement Speeches Than They Used To" <https://fivethirtyeight.com/features/sitting-presidents-give-way-more-commencement-speeches-they-used-to/>

**Usage**

pres\_commencement

**Format**

A data frame with 154 rows representing speeches and 8 variables:

**pres** Number of president (33 is Harry Truman, the 33rd president; 44 is Barack Obama, the 44th president)

**pres\_name** Name of president

**title** Description of commencement speech

**date** Date speech was delivered

**city** City where speech was delivered

**state** State where speech was delivered

**building** Name of building in which speech was delivered

**room** Room in which speech was delivered

**Source**

American Presidency Project, Gerhard Peters and John T. Woolley <https://www.presidency.ucsb.edu>

---

pulitzer	<i>Do Pulitzers Help Newspapers Keep Readers?</i>
----------	---

---

**Description**

The raw data behind the story "Do Pulitzers Help Newspapers Keep Readers?" <https://fivethirtyeight.com/features/do-pulitzers-help-newspapers-keep-readers/>.

**Usage**

pulitzer

**Format**

A data frame with 50 rows representing newspapers and 7 variables:

**newspaper** Newspaper

**circ2004** Daily Circulation in 2004

**circ2013** Daily Circulation in 2013

**pctchg\_circ** Percent change in Daily Circulation from 2004 to 2013

**num\_finals1990\_2003** Number of Pulitzer Prize winners and finalists from 1990 to 2003

**num\_finals2004\_2014** Number of Pulitzer Prize winners and finalists from 2004 to 2014

**num\_finals1990\_2014** Number of Pulitzer Prize winners and finalists from 1990 to 2014

**Source**

See <https://fivethirtyeight.com/features/do-pulitzers-help-newspapers-keep-readers/>

---

riddler\_castles

*Can You Rule Riddler Nation?*

---

**Description**

The raw data behind the story "Can You Rule Riddler Nation?" <https://fivethirtyeight.com/features/can-you-rule-riddler-nation/>. Analysis of the submitted solutions can be found at: <https://fivethirtyeight.com/features/can-you-save-the-drowning-swimmer/>

**Usage**

riddler\_castles

**Format**

A data frame with 1387 rows representing submissions and 11 variables:

**castle1** Number of troops out of 100 send to castle 1

**castle2** Number of troops out of 100 send to castle 2

**castle3** Number of troops out of 100 send to castle 3

**castle4** Number of troops out of 100 send to castle 4

**castle5** Number of troops out of 100 send to castle 5

**castle6** Number of troops out of 100 send to castle 6

**castle7** Number of troops out of 100 send to castle 7

**castle8** Number of troops out of 100 send to castle 8

**castle9** Number of troops out of 100 send to castle 9

**castle10** Number of troops out of 100 send to castle 10

**reason** Why did you choose your troop deployment?

## Source

See <https://github.com/fivethirtyeight/data/tree/master/riddler-castles>

## See Also

[riddler\\_castles2](#)

## Examples

```
# To convert data frame to tidy data (long) format, run
library(dplyr)
library(tidyr)
library(stringr)
riddler_castles_tidy<-riddler_castles %>%
  pivot_longer(castle1:castle10, names_to = "castle" , values_to = "soldiers") %>%
  mutate(castle = as.numeric(str_replace(castle, "castle","")))
```

---

riddler\_castles2

*The Battle For Riddler Nation, Round 2*

---

## Description

The raw data behind the story "The Battle For Riddler Nation, Round 2" <https://fivethirtyeight.com/features/the-battle-for-riddler-nation-round-2/>. Analysis of the submitted solutions can be found at: <https://fivethirtyeight.com/features/how-much-should-you-bid-for-that-painting/>

## Usage

```
riddler_castles2
```

## Format

A data frame with 932 rows representing submissions and 11 variables:

**castle1** Number of troops out of 100 send to castle 1  
**castle2** Number of troops out of 100 send to castle 2  
**castle3** Number of troops out of 100 send to castle 3  
**castle4** Number of troops out of 100 send to castle 4  
**castle5** Number of troops out of 100 send to castle 5  
**castle6** Number of troops out of 100 send to castle 6  
**castle7** Number of troops out of 100 send to castle 7  
**castle8** Number of troops out of 100 send to castle 8  
**castle9** Number of troops out of 100 send to castle 9  
**castle10** Number of troops out of 100 send to castle 10  
**reason** Why did you choose your troop deployment?

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/riddler-castles>

**See Also**

[riddler\\_castles](#)

**Examples**

```
# To convert data frame to tidy data (long) format, run
library(dplyr)
library(tidyr)
library(stringr)
riddler_castles_tidy<-riddler_castles2 %>%
  pivot_longer(castle1:castle10, names_to = "castle" , values_to = "soldiers") %>%
  mutate(castle = as.numeric(str_replace(castle, "castle","")))
```

---

riddler\_pick\_lowest     *Pick A Number, Any Number*

---

**Description**

The raw data behind the story "Pick A Number, Any Number" <https://fivethirtyeight.com/features/pick-a-number-any-number/>

**Usage**

```
riddler_pick_lowest
```

**Format**

A data frame with 3660 rows representing dates and 1 variable:

**your\_number** Guessed number

**show\_your\_work** People showing their work

---

russia\_investigation *Russia Investigation*

---

### Description

This folder contains data behind the story 'Is The Russia Investigation Really Another Watergate?' <https://projects.fivethirtyeight.com/russia-investigation/>

### Usage

russia\_investigation

### Format

A dataset with 194 rows representing every special investigation since the Watergate probe began in 1973 and 13 variables

**investigation** Unique id for each investigation

**investigation\_start** Start date of the investigation

**investigation\_end** End date of the investigation

**investigation\_days** Length, in days, of the investigation. Days will be negative if the charge occurred before the investigation began.

**name** Name of the person charged (if applicable). Will be blank if there were no charges.

**indictment\_days** Length, in days, from the start of the investigation to the date the person was charged (if applicable). Days will be negative if the charge occurred before the investigation began.

**type** Result of charge (if applicable)

**cp\_date** Date the person plead guilty or was convicted (if applicable)

**cp\_days** Length, in days, from the start of the investigation to the date the person plead guilty or was convicted (if applicable)

**overturned** Whether or not the relevant person's conviction was overturned

**pardoned** Whether or not the relevant person's charge was pardoned

**american** Whether or not the relevant person's charge was a U.S. resident

**president** President at the center of the investigation

### Source

Information for this story is drawn from an original data set of special counsel, independent counsel and special prosecutor investigations from 1973 to 2019. The data set was created by consulting historical sources, including final reports generated by independent counsels, special counsels and special prosecutors; reports in Congressional Quarterly; and contemporaneous news stories. Secondary historical sources were also consulted, including a 2006 Congressional Research Service report about independent counsel investigations and a history of the Watergate investigation by Stanley Kutler. Data about pardons was obtained from the Office of the Pardon Attorney. Indicted

organizations were excluded from our analysis. The data set, which is available on Github, includes the names of all people charged as part of these investigations, as well as the outcome of their cases and the dates of major actions in their cases.

2006 Congressional Research Service report: [https://digital.library.unt.edu/ark:/67531/metadc815038/m2/1/high\\_res\\_d/98-19\\_2006Jun08.pdf](https://digital.library.unt.edu/ark:/67531/metadc815038/m2/1/high_res_d/98-19_2006Jun08.pdf)

dataset in GitHub: <https://github.com/fivethirtyeight/data/tree/master/russia-investigation>

---

sandy\_311

*The (Very) Long Tail Of Hurricane Recovery*

---

### Description

The raw data behind the story "The (Very) Long Tail Of Hurricane Recovery" <https://projects.fivethirtyeight.com/sandy-311/>

### Usage

sandy\_311

### Format

A data frame with 1783 rows representing dates and 25 variables:

**date** Date

**nyc\_311** No description provided.

**acs** The number of emergency hotline (311) calls made to the Administration for Children's Services related to Hurricane Sandy on the given date

**bpsi** The number of emergency hotline (311) calls made to Building Protection Systems, Inc related to Hurricane Sandy

**cau** The number of emergency hotline (311) calls made to the Community Affairs Unit related to Hurricane Sandy

**chall** The number of emergency hotline (311) calls made to the City Hall related to Hurricane Sandy

**dep** The number of emergency hotline (311) calls made to the Department of Environmental Protection related to Hurricane Sandy

**dob** The number of emergency hotline (311) calls made to the Department of Buildings related to Hurricane Sandy

**doe** The number of emergency hotline (311) calls made to the Department of Education related to Hurricane Sandy

**dof** The number of emergency hotline (311) calls made to the Department of Finance related to Hurricane Sandy

**dohmh** The number of emergency hotline (311) calls made to the Department of Health and Mental Hygiene related to Hurricane Sandy

- dpr** The number of emergency hotline (311) calls made to the Department of Parks and Recreation related to Hurricane Sandy
- fema** The number of emergency hotline (311) calls made to the Federal Emergency Management Agency related to Hurricane Sandy
- hpd** The number of emergency hotline (311) calls made to the Department of Housing Preservation and Development related to Hurricane Sandy
- hra** The number of emergency hotline (311) calls made to the Human Resources Administration related to Hurricane Sandy
- mfanyc** The number of emergency hotline (311) calls made to the Mayor's Fund to Advance NYC related to Hurricane Sandy
- mose** The number of emergency hotline (311) calls made to the Mayor's Office of Special Enforcement related to Hurricane Sandy
- nycem** The number of emergency hotline (311) calls made to Emergency Management related to Hurricane Sandy
- nycha** The number of emergency hotline (311) calls made to the New York City Housing Authority related to Hurricane Sandy
- nyc\_service** The number of emergency hotline (311) calls made to NYC Service related to Hurricane Sandy
- nypd** The number of emergency hotline (311) calls made to the New York Police Department related to Hurricane Sandy
- nysdol** The number of emergency hotline (311) calls made to the NYC Department of Labor related to Hurricane Sandy
- sbs** The number of emergency hotline (311) calls made to Small Business Services related to Hurricane Sandy
- nys\_emergency\_mg** The number of emergency hotline (311) calls made to NYS Emergency Management related to Hurricane Sandy
- total** The total number of emergency hotline (311) calls made related to Hurricane Sandy

## Source

Data from NYC Open Data <https://data.cityofnewyork.us/City-Government/311-Call-Center-Inquiry/tdd6-3ysr>, Agency acronyms from the Data Dictionary. See also <https://github.com/fivethirtyeight/data/tree/master/sandy-311-calls>

## Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
sandy_311_tidy <- sandy_311 %>%
  pivot_longer(-c("date", "total"), names_to = "agency", values_to = "num_calls") %>%
  arrange(date) %>%
  select(date, agency, num_calls, total) %>%
  rename(total_calls = total) %>%
  mutate(agency = as.factor(agency))
```

---

san_andreas	<i>The Rock Isn't Alone: Lots Of People Are Worried About 'The Big One'</i>
-------------	---

---

### Description

The raw data behind the story "The Rock Isn't Alone: Lots Of People Are Worried About 'The Big One'" <https://fivethirtyeight.com/features/the-rock-isnt-alone-lots-of-people-are-worried-about-the>

### Usage

san\_andreas

### Format

A data frame with 1013 rows representing respondents and 11 variables:

**worry\_general** In general, how worried are you about earthquakes?

**worry\_bigone** How worried are you about the "Big One," a massive, catastrophic earthquake?

**will\_occur** Do you think the "Big One" will occur in your lifetime?

**experience** Have you ever experienced an earthquake?

**prepared** Have you or anyone in your household taken any precautions for an earthquake (packed an earthquake survival kit, prepared an evacuation plan, etc.)?

**fam\_san\_andreas** How familiar are you with the San Andreas Fault line?

**fam\_yellowstone** How familiar are you with the Yellowstone Supervolcano?

**age** Age

**female** Gender

**hhold\_income** How much total combined money did all members of your HOUSEHOLD earn last year?

**region** US Region

### Source

See <https://github.com/fivethirtyeight/data/tree/master/san-andreas>

---

senate\_national\_forecast

*Senate Forecast 2018*

---

## Description

This file contains links to the data behind FiveThirtyEight's 'Senate forecasts' <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate/>

## Usage

senate\_national\_forecast

## Format

A dataframe with 450 rows representing national-level results of the classic, lite, and deluxe Senate forecasts since Aug. 1, 2018 and 11 variables

**forecastdate** date of the forecast

**party** the party of the forecast

**model** the model of the forecast

**win\_probability** the probability of the corresponding party winning

**mean\_seats** the mean of the number of seats

**median\_seats** the median number of seats

**p10\_seats** the top 10 percentile of number of seats

**p90\_seats** the top 90 percentile of number of seats

**margin** unknown

**p10\_margin** the margin of p10\_seats

**p90\_margin** the margin of p90\_seats

## Note

The original dataset included a meaningless column called "state", and all variables under this column was "US". So this column was removed.

## Source

FiveThirtyEight's House, Senate And Governor Models Methodology: <https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/>

## See Also

[senate\\_seat\\_forecast](#)

---

senate_polls	<i>Early Senate Polls Have Plenty to Tell Us About November</i>
--------------	---

---

**Description**

The raw data behind the story "Early Senate Polls Have Plenty to Tell Us About November" <https://fivethirtyeight.com/features/early-senate-polls-have-plenty-to-tell-us-about-november/>.

**Usage**

```
senate_polls
```

**Format**

A data frame with 107 rows representing a poll and 4 variables:

**year** Year

**election\_result** Final poll margin

**presidential\_approval** Early presidential approval rating

**poll\_average** Early poll margin

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/early-senate-polls>

---

senate_seat_forecast	<i>Senate Forecast 2018</i>
----------------------	-----------------------------

---

**Description**

This file contains links to the data behind FiveThirtyEight's 'Senate forecasts' <https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate/>

**Usage**

```
senate_seat_forecast
```

**Format**

A dataframe with 28353 rows representing seat-level results of the classic, lite, and deluxe Senate forecasts since Aug. 1, 2018 and 12 variables

**forecastdate** date of the forecast

**state** state of the forecast

**class** class of the forecast

**special** unknown

**candidate** name of the candidate

**party** party of the candidate

**incumbent** whether the candidate is incumbent

**model** the model of the forecast

**win\_probability** the probability of the corresponding party winning

**voteshare** the voteshare of the corresponding party

**p10\_voteshare** the top 10 percentile of the voteshare

**p90\_voteshare** the top 00 percentile of the voteshare

**Source**

FiveThirtyEight's House, Senate And Governor Models Methodology: <https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/>

**See Also**

[senate\\_national\\_forecast](#)

---

spi\_global\_rankings     *Current SPI ratings and rankings for men's club teams*

---

**Description**

The raw data behind the stories "Club Soccer Predictions" <https://projects.fivethirtyeight.com/soccer-predictions/> and "Global Club Soccer Rankings" <https://projects.fivethirtyeight.com/soccer-predictions/global-club-rankings/>.

**Usage**

spi\_global\_rankings

**Format**

A data frame with 453 rows representing soccer rankings and 7 variables:

**name** The name of the soccer club.

**league** The name of the league to which the club belongs.

**rank** A club's current global ranking.

**prev\_rank** A club's previous global ranking

**off** Offensive rating for a given team (the higher the value the stronger the team's offense).

**def** Defensive rating for a given team (the lower the value the stronger the team's defense).

**spi** A club's SPI score.

**Source**

See <https://github.com/fivethirtyeight/data/blob/master/soccer-spi/README.md>

**See Also**

[spi\\_matches](#)

---

state\_info

*Information on each state*

---

**Description**

State name, abbreviation, US Census designated division & region.

**Usage**

state\_info

**Format**

A data frame with 51 rows representing airlines and 4 variables:

**state** State name

**state\_abbrev** State abbreviation

**division** US Census designated division. Values for division are nested within region

**region** US Census designated region

**Source**

US Census Bureau [https://en.wikipedia.org/wiki/List\\_of\\_regions\\_of\\_the\\_United\\_States#Interstate\\_regions](https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States#Interstate_regions).

## Examples

```
library(dplyr)
# Number of states in each division
state_info %>%
  count(division)
# Number of states in each region
state_info %>%
  count(region)
```

---

state\_of\_the\_state      *What America's Governors Are Talking About*

---

## Description

The raw data behind the story "What America's Governors Are Talking About" <https://fivethirtyeight.com/features/what-americas-governors-are-talking-about/>

## Usage

```
state_index
state_words
```

## Format

2 data frames about the 50 U.S Governors' Speeches

An object of class `spec_tbl_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 2223 rows and 9 columns.

## state\_index

A data frame with 50 rows representing the 50 U.S. states and 5 variables:

**state** the state

**governor** the name of the state's governor

**party** the party of the state's governor

**filename** the filename of the speech in the speeches folder at <https://github.com/rudeboybert/fivethirtyeight/tree/master/data-raw/state-of-the-state/speeches>

**url** a link to an official/media source for the speech

**state\_words**

A data frame with 2,223 rows representing phrases and 9 variables:

**phrase** one-, two-, and three-word phrases spoken repeatedly  
**category** thematic categories for the phrases  
**d\_speeches** number of Democratic speeches containing the phrase  
**r\_speeches** number of Republican speeches containing the phrase  
**total** total number of speeches containing the phrase  
**percent\_of\_d\_speeches** percent of the 23 Democratic speeches containing the phrase  
**percent\_of\_r\_speeches** percent of the 27 Republican speeches containing the phrase  
**chi2** the chi-square test statistic for statistical significance  
**pval** p-value for chi<sup>2</sup> test

**Source**

The chi-square test statistic [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.chi2.html#sklearn.feature\\_selection.chi2](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2)

---

steak\_survey

*How Americans Like Their Steak*

---

**Description**

The raw data behind the story "How Americans Like Their Steak" <https://fivethirtyeight.com/features/how-americans-like-their-steak/>.

**Usage**

steak\_survey

**Format**

A data frame with 550 rows representing respondents and 15 variables:

**respondent\_id** Respondent ID  
**lottery\_a** not sure  
**smoke** Is respondent a smoker?  
**alcohol** Is respondent a drinker?  
**gamble** Is respondent a gambler?  
**skydiving** Is respondent a skydiver?  
**speed** not sure  
**cheated** not sure  
**steak** not sure

**steak\_prep** Preferred steak preparation

**female** Is respondent female?

**age** Age

**hhold\_income** Household income

**educ** Education level

**region** Region of US

### Source

See <https://fivethirtyeight.com/features/how-americans-like-their-steak/>

---

tarantino

*A Complete Catalog Of Every Time Someone Cursed Or Bled Out In A Quentin Tarantino Movie*

---

### Description

The raw data behind the story "A Complete Catalog Of Every Time Someone Cursed Or Bled Out In

A Quentin Tarantino Movie" <https://fivethirtyeight.com/features/complete-catalog-curses-deaths-quentin->

An analysis using this data was contributed by Olivia Barrows, Jojo Miller, and Jayla Nakayama as a package vignette at [https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/tarantino\\_swears.html](https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/tarantino_swears.html).

### Usage

tarantino

### Format

A data frame with 1894 rows representing curse/death instances and 4 variables:

**movie** Film title

**profane** Whether the event was a profane word (TRUE) or a death (FALSE)

**word** The specific profane word, if the event was a word

**minutes\_in** The number of minutes into the film the event occurred

### Source

See <https://github.com/fivethirtyeight/data/tree/master/tarantino>

---

tennis\_events\_time      *Why Some Tennis Matches Take Forever*

---

**Description**

The raw data behind the story "Why Some Tennis Matches Take Forever" <https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/>.

**Usage**

tennis\_events\_time

**Format**

A data frame with 205 rows representing tournaments and 5 variables:

**tournament** Name of event

**surface** Court surface used at the event

**sec\_added** Seconds added per point for this event on this surface in years shown, from regression model controlling for players, year and other factors

**year\_start** Start year for data used from this tournament in regression

**year\_end** End year for data used from this tournament in regression

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/tennis-time>

**See Also**

[tennis\\_players\\_time](#) and [tennis\\_serve\\_time](#)

---

tennis\_players\_time      *Why Some Tennis Matches Take Forever*

---

**Description**

The raw data behind the story "Why Some Tennis Matches Take Forever" <https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/>.

**Usage**

tennis\_players\_time

**Format**

A data frame with 218 rows representing players and 2 variables:

**player** Player Name

**sec\_added** Weighted average of seconds added per point as loser and winner of matches, 1991-2015, from regression model controlling for tournament, surface, year and other factors

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/tennis-time>

**See Also**

[tennis\\_events\\_time](#) and [tennis\\_serve\\_time](#)

---

tennis_serve_time	<i>Why Some Tennis Matches Take Forever</i>
-------------------	---

---

**Description**

The raw data behind the story "Why Some Tennis Matches Take Forever" <https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/>.

**Usage**

```
tennis_serve_time
```

**Format**

A data frame with 120 rows representing serves and 7 variables:

**server** Name of player serving at 2015 French Open

**sec\_between** Time in seconds between end of marked point and next serve, timed by stopwatch app

**opponent** Opponent, receiving serve

**game\_score** Score in the current game during the timed interval between points

**set** Set number, out of five

**game** Score in games within the set

**date** Date

**Source**

See <https://github.com/fivethirtyeight/data/tree/master/tennis-time>

**See Also**

[tennis\\_events\\_time](#) and [tennis\\_players\\_time](#)

---

tenth_circuit	<i>For A Trump Nominee, Neil Gorsuch's Record Is Surprisingly Moderate On Immigration</i>
---------------	---

---

### Description

The raw data behind the story "For A Trump Nominee, Neil Gorsuch's Record Is Surprisingly Moderate On Immigration" <https://fivethirtyeight.com/features/for-a-trump-nominee-neil-gorsuchs-record-is>

### Usage

```
tenth_circuit
```

### Format

A data frame with 954 rows representing cases and 13 variables:

**title** Name of the case

**date** Date of decision

**federalreporter\_cit** Case citation, as listed in the Federal Reporter Series

**westlaw\_cit** Case citation, Westlaw format

**issue** Issue number, in cases divided into multiple issues

**weight** Weight per issue (total weight per case equals one)

**judge1** Name of first judge

**judge2** Name of second judge

**judge3** Name of third judge

**vote1\_liberal** Vote of first judge. 1 = liberal, 0 = conservative.

**vote2\_liberal** Vote of second judge. 1 = liberal, 0 = conservative.

**vote3\_liberal** Vote of third judge. 1 = liberal, 0 = conservative.

**category** Category of case, immigration or discrimination

### Note

In immigration cases, partial relief to immigration petitioner is coded as liberal because the petitioner typically seeks just one core remedy (e.g., withholding of removal, adjustment of status, or asylum); in discrimination cases, partial relief is coded as multiple issues because the plaintiff often seeks separate remedies under multiple claims (e.g., disparate treatment, retaliation, etc.) and different sources of law.

### Source

See <https://github.com/fivethirtyeight/data/tree/master/tenth-circuit>

---

trumpworld\_issues      *What the World Thinks of Trump*

---

### Description

The raw data behind the story "What the World Thinks of Trump" <https://fivethirtyeight.com/features/what-the-world-thinks-of-trump/>: Trump World Issues Dataset

### Usage

```
trumpworld_issues
```

### Format

A data frame with 185 rows representing countries and 6 variables:

**country** The country whose population is being polled

**net\_approval** The difference in the number of respondents from the given country who approve and who disapprove of the issue (Trump proposal) in question (approve-disapprove)

**approve** The number of respondents from the given country who approve of the issue (Trump proposal)

**disapprove** The number of respondents who disapprove of the issue

**dk\_refused** undefined

**issue** The specific trump policy proposal being posed. Specifically: 1: Withdraw support for international climate change agreements 2: Build a wall on the border between the U. S. and Mexico 3: Withdraw U.S. support from the Iran nuclear weapons agreement 4: Withdraw U.S. support for major trade agreements 5: Introduce tighter restrictions on those entering the U.S. from some majority-Muslim countries

### Source

Pew Research Center <https://www.pewresearch.org/fact-tank/2017/07/17/9-charts-on-how-the-world-sees-t>

### See Also

[trumpworld\\_polls](#)

---

trumpworld\_polls

*What the World Thinks of Trump*


---

### Description

The raw data behind the story "What the World Thinks of Trump" <https://fivethirtyeight.com/features/what-the-world-thinks-of-trump/>: Trump World Polls Dataset.

### Usage

```
trumpworld_polls
```

### Format

A data frame with 32 rows representing years and 40 variables:

**year** Year the poll was conducted

**avg** The average percentage people who answered the poll question positively (support the president or have a favorable view of the U.S.)

**canada** The percentage of people from Canada who answered the poll question positively

**france** The percentage of people from France who answered the poll question positively

**germany** The percentage of people from Germany who answered the poll question positively

**greece** The percentage of people from Greece who answered the poll question positively

**hungary** The percentage of people from Hungary who answered the poll question positively

**italy** The percentage of people from Italy who answered the poll question positively

**netherlands** The percentage of people from Netherlands who answered the poll question positively

**poland** The percentage of people from Poland who answered the poll question positively

**spain** The percentage of people from Spain who answered the poll question positively

**sweden** The percentage of people from Sweden who answered the poll question positively

**uk** The percentage of people from the U.K. who answered the poll question positively

**ruussia** The percentage of people from Russia who answered the poll question positively

**australia** The percentage of people from Australia who answered the poll question positively

**india** The percentage of people from India who answered the poll question positively

**indonesia** The percentage of people from Indonesia who answered the poll question positively

**japan** The percentage of people from Japan who answered the poll question positively

**philippines** The percentage of people from the Philippines who answered the poll question positively

**south\_korea** The percentage of people from South Korea who answered the poll question positively

**vietnam** The percentage of people from Vietnam who answered the poll question positively

**israel** The percentage of people from Israel who answered the poll question positively

**jordan** The percentage of people from Jordan who answered the poll question positively

**lebanon** The percentage of people from Lebanon who answered the poll question positively

**tunisia** The percentage of people from Tunisia who answered the poll question positively

**turkey** The percentage of people from Turkey who answered the poll question positively

**ghana** The percentage of people from Ghana who answered the poll question positively

**kenya** The percentage of people from Kenya who answered the poll question positively

**nigeria** The percentage of people from Nigeria who answered the poll question positively

**senegal** The percentage of people from Senegal who answered the poll question positively

**south\_africa** The percentage of people from South Africa who answered the poll question positively

**tanzania** The percentage of people from Tanzania who answered the poll question positively

**argentina** The percentage of people from Argentina who answered the poll question positively

**brazil** The percentage of people from Brazil who answered the poll question positively

**chile** The percentage of people from Chile who answered the poll question positively

**colombia** The percentage of people from Colombia who answered the poll question positively

**mexico** The percentage of people from Mexico who answered the poll question positively

**peru** The percentage of people from Peru who answered the poll question positively

**venezuela** The percentage of people from Venezuela who answered the poll question positively

**question** The item being polled. Specifically, whether respondents: 1) Have a favorable view of the U.S. or 2) Trust the U.S. President when it comes to foreign affairs

### Source

Pew Research Center <https://www.pewresearch.org/fact-tank/2017/07/17/9-charts-on-how-the-world-sees-t>

### See Also

[trumpworld\\_issues](#)

### Examples

```
# To convert data frame to tidy data (long) format, run:
library(dplyr)
library(tidyr)
trumpworld_polls_tidy <- trumpworld_polls %>%
  pivot_longer(-c("year", "avg", "question"),
    names_to = "country", values_to = "percent_positive")
```

---

trump\_approval\_poll    *How Popular is Donald Trump*

---

### Description

The raw data behind the story: "How Popular is Donald Trump" <https://projects.fivethirtyeight.com/trump-approval-ratings/>: Approval Poll Dataset

### Usage

trump\_approval\_poll

### Format

A data frame with 3051 rows representing individual polls and 20 variables:

**subgroup** The subgroup the poll falls into as defined by the type of people being polled (all polls, voters, adults)

**start\_date** The date the polling began

**end\_date** The date the polling concluded

**pollster** The polling group which produced the poll

**grade** The grade for President Trump that the respondents' approval ratings correspond to

**sample\_size** The sample size of the poll

**population** The type of people being polled (a for adults, lv for likely voters, rv for registered voters)

**weight** The weight fivethirtyeight gives the poll when determining approval ratings based on historical accuracy of the pollster

**approve** The percentage of respondents who approve of the president

**disapprove** The percentage of respondents who disapprove of the president

**adjusted\_approve** The percentage of respondents who approve of the president adjusted for systematic tendencies of the polling firm

**adjusted\_disapprove** The percentage of respondents who disapprove of the president adjusted for systematic tendencies of the polling firm

**multiversions** True if there are multiple versions of the poll, False if there are not

**tracking** TRUE if the poll was tracked, FALSE if not

**url** Poll result URL

**poll\_id** Poll ID number

**question\_id** ID number for the question being polled

**created\_date** Date the poll was created

**timestamp** Date and time the poll was compiled

### Details

Variables "model\_date", "influence", and "president" were deleted because each observation contained the same value for these variables: January 5, 2018; 0; and Donald Trump respectively.

### Source

[https://projects.fivethirtyeight.com/trump-approval-data/approval\\_polllist.csv](https://projects.fivethirtyeight.com/trump-approval-data/approval_polllist.csv) and [https://projects.fivethirtyeight.com/trump-approval-data/approval\\_topline.csv](https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv)

### See Also

[trump\\_approval\\_trend](#)

---

trump\_approval\_trend *How Popular is Donald Trump*

---

### Description

The raw data behind the story: "How Popular is Donald Trump" <https://projects.fivethirtyeight.com/trump-approval-ratings/>: Approval Trend Dataset.

### Usage

```
trump_approval_trend
```

### Format

A data frame with 1044 rows representing poll trends and 11 variables:

**subgroup** The subgroup the poll falls into as defined by the type of people being polled (all polls, voters, adults)

**modeldate** The date the model was created

**approve\_estimate** Estimated approval ratings

**approve\_high** Higher bound of the estimated approval percentage

**approve\_low** Lower bound of the estimated approval percentage

**disapprove\_estimate** Estimated disapproval percentage

**disapprove\_high** Higher bound of the estimated disapproval percentage

**disapprove\_low** Lower bound of the estimated disapproval percentage

**timestamp** Date and time the model was compiled

### Details

The Variable "president" was removed because all values were "Donald Trump"

**Source**

[https://projects.fivethirtyeight.com/trump-approval-data/approval\\_topline.csv](https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv)

**See Also**

[trump\\_approval\\_poll](#)

---

trump\_lawsuits

*Trump Lawsuits*

---

**Description**

This folder contains the data behind the stories: 'What Trump's Legal Battles Tell Us About Presidential Power' <https://fivethirtyeight.com/features/what-trumps-legal-battles-tell-us-about-presidential-power/>; 'Why It Might Be Impossible To Overturn A Presidential Pardon' <https://fivethirtyeight.com/features/why-it-might-be-impossible-to-overturn-a-presidential-pardon/>; 'Will The Supreme Court Fast-Track Cases Involving Trump?' <https://fivethirtyeight.com/features/will-the-supreme-court-fast-track-cases-involving-trump/>; 'Why One of Trump's Biggest Legal Threats Is New York's Attorney General' <https://fivethirtyeight.com/features/why-one-of-trumps-biggest-legal-threats-is-new-yorks-attorney-general/>; 'Should Judges Pay Attention To Trump's Tweets?' <https://fivethirtyeight.com/features/should-judges-pay-attention-to-trumps-tweets/>; 'Trump Is Losing The Legal Fight Against Sanctuary Cities, But It May Still Pay Off Politically' <https://fivethirtyeight.com/features/trump-is-losing-the-legal-fight-against-sanctuary-cities-but-it-may-still-pay-off-politically/>; 'Will Trump's Latest Lawsuits Keep Congress From Investigating Future Presidents?' <https://fivethirtyeight.com/features/will-trumps-latest-lawsuits-keep-congress-from-investigating-future-presidents/>

**Usage**

trump\_lawsuits

**Format**

A dataset with 57 rows representing lawsuits and 16 variables

**docket\_number** Current docket number

**date\_filed** Date lawsuit was originally filed

**case\_name** Case name (current)

**plaintiff** Names of plaintiffs (if more than five, "et al" for plaintiffs who are not in case name)

**defendant** Names of defendants (if more than five, "et al" for defendants who are not in case name)

**current\_location** Court the lawsuit is currently in front of

**previous\_location** Other courts the case has appeared before

**jurisdiction** Where the case is being heard | 1 = Federal; 2 = State

**judge** Names of the judges the case is currently before

**nature** PACER code for nature of lawsuit (Not relevant for criminal cases) <https://pacer.uscourts.gov/help/faqs/what-nature-suit-code>

**trump\_category** Whether the case is related to action before Trump was president, his personal conduct as president, or a policy action as president | 1 = Case directed at pre-presidency action; 2 = Case directed at personal action of Trump as president; 3 = Case directed at policy action of Trump as president

**capacity** The capacity in which Trump is implicated | 1 = Case directed at Trump personally; 2 = Case directed at action of Trump administration; 3 = Trump as plaintiff; 4 = Trump administration as plaintiff; 5 = Case directed at Trump associate; 6 = Other

**type** Criminal vs. civil | 1 = Criminal; 2 = Civil

**issue** Key topic area raised in the case (i.e. emoluments, First Amendment, DACA, etc). Categories created based on key policy topic area or legal issue. Calls are subjective and based on reporting and may change.

**docket\_orig** Original docket number, if case has been appealed or changed jurisdiction

**status** Whether the case, or the part of the case connected to Trump, is ongoing. | 1 = Case is ongoing; 2 = Case or part of case connected to Trump is closed

### Source

Approval Polls

---

trump\_news

*How Trump Hacked The Media*

---

### Description

The raw data behind the story "How Trump Hacked The Media" <https://fivethirtyeight.com/features/how-donald-trump-hacked-the-media/>.

### Usage

trump\_news

### Format

A data frame with 286 rows representing lead stories and 3 variables:

**date** Date of lead story about Donald Trump.

**major\_cat** Story classification

**detail**

### Source

Memeorandum <https://www.memeorandum.com/>.

---

`trump_twitter`*The World's Favorite Donald Trump Tweets*

---

**Description**

The raw data behind the story "The World's Favorite Donald Trump Tweets" <https://fivethirtyeight.com/features/the-worlds-favorite-donald-trump-tweets/>. Tweets posted on twitter by Donald Trump (@realDonaldTrump). An analysis using this data was contributed by Adam Spannbauer as a package vignette at [https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/trump\\_twitter.html](https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/trump_twitter.html).

**Usage**`trump_twitter`**Format**

A data frame with 448 rows representing tweets and 3 variables:

`id``created_at``text`**Source**

Twitter <https://twitter.com/realdonaldtrump>

---

`tv_hurricanes`*The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

**Description**

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/>. TV Hurricanes Data.

**Usage**`tv_hurricanes`

**Format**

A data frame with 37 rows representing dates and 5 variables:

**date** Date

**harvey** The percent of sentences in TV news that mention Hurricane Harvey on the given date

**irma** The percent of sentences in TV news that mention Hurricane Irma

**maria** The percent of sentences in TV news that mention Hurricane Maria

**jose** The percent of sentences in TV news that mention Hurricane Irma

**Source**

Internet TV News Archive <https://archive.org/details/tv> and Television Explorer

**See Also**

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes\\_by\\_network](#), [tv\\_states](#), [google\\_trends](#)

---

tv\_hurricanes\_by\_network

*The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

**Description**

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico-when-trump-did/> TV Hurricanes by Network Data.

**Usage**

tv\_hurricanes\_by\_network

**Format**

A data frame with 84 rows representing dates and 6 variables:

**date** Date

**query** The hurricane in question

**bbc\_news** The percent of sentences on the BBC News TV channel on the given date that mention the hurricane in question

**cnn** The percent of sentences on CNN News that mention the hurricane in question

**fox\_news** The percent of sentences on Fox News that mention the hurricane in question

**msnbc** The percent of sentences on MSNBC News that mention the hurricane in question

**Source**

Internet TV News Archive <https://archive.org/details/tv> and Television Explorer

**See Also**

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_states](#), [google\\_trends](#)

---

tv\_states

*The Media Really Started Paying Attention to Puerto Rico When Trump Did*

---

**Description**

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" <https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-puerto-rico/> TV States Data.

**Usage**

tv\_states

**Format**

A data frame with 52 rows representing dates and 4 variables:

**date** Date

**florida** The percent of sentences in TV News on the given day that mention Florida

**texas** The percent of sentences in TV News on the given day that mention Texas

**puerto\_rico** The percent of sentences in TV News on the given day that mention Puerto Rico

**Source**

Internet TV News Archive <https://archive.org/details/tv> and Television Explorer

**See Also**

[mediacloud\\_hurricanes](#), [mediacloud\\_states](#), [mediacloud\\_online\\_news](#), [mediacloud\\_trump](#), [tv\\_hurricanes](#), [tv\\_hurricanes\\_by\\_network](#), [google\\_trends](#)

---

undefeated	<i>Mayweather Is Defined By The Zero Next To His Name</i>
------------	---

---

**Description**

The raw data behind: "Mayweather Is Defined By The Zero Next To His Name" <https://fivethirtyeight.com/features/mayweather-is-defined-by-the-zero-next-to-his-name/>

**Usage**

```
undefeated
```

**Format**

A data frame with 2125 rows representing boxing matches and 4 variables:

**name** Name of boxer

**url** URL with the boxer's record

**date** Date of the match

**wins** Number of cumulative wins for the boxer including the match at the specified date

**Source**

Box Rec

---

unisex_names	<i>The Most Common Unisex Names In America: Is Yours One Of Them?</i>
--------------	---

---

**Description**

The raw data behind the story "The Most Common Unisex Names In America: Is Yours One Of Them?" <https://fivethirtyeight.com/features/there-are-922-unisex-names-in-america-is-yours-one-of->

**Usage**

```
unisex_names
```

**Format**

A data frame with 919 rows representing names and 5 variables:

**name** First names from the Social Security Administration

**total** Total number of living Americans with the name

**male\_share** Percentage of people with the name who are male

**female\_share** Percentage of people with the name who are female

**gap** Gap between male\_share and female\_share

**Source**

Social Security Administration <https://www.ssa.gov/oact/babynames/limits.html>. See <https://github.com/fivethirtyeight/data/tree/master/unisex-names>.

---

US_births_1994_2003	<i>Some People Are Too Superstitious To Have A Baby On Friday The 13th</i>
---------------------	--

---

**Description**

The raw data behind the story "Some People Are Too Superstitious To Have A Baby On Friday The 13th" <https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-frid>

**Usage**

US\_births\_1994\_2003

**Format**

A data frame with 3652 rows representing dates and 6 variables:

**year** Year

**month** Month

**date\_of\_month** Day

**date** POSIX date

**day\_of\_week** Abbreviation of day of week

**births** Number of births

**Source**

Centers for Disease Control and Prevention's National Center for Health Statistics

**See Also**

[US\\_births\\_2000\\_2014](#)

---

US\_births\_2000\_2014    *Some People Are Too Superstitious To Have A Baby On Friday The 13th*

---

**Description**

The raw data behind the story "Some People Are Too Superstitious To Have A Baby On Friday The 13th" <https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-frid>

**Usage**

US\_births\_2000\_2014

**Format**

A data frame with 5479 rows representing dates and 6 variables:

**year** Year

**month** Month

**date\_of\_month** Day

**date** POSIX date

**day\_of\_week** Abbreviation of day of week

**births** Number of births

**Source**

Social Security Administration

**See Also**

[US\\_births\\_1994\\_2003](#).

---

weather\_check    *Where People Go To Check The Weather*

---

**Description**

The raw data behind the story "Where People Go To Check The Weather" <https://fivethirtyeight.com/features/weather-forecast-news-app-habits/>.

**Usage**

weather\_check

**Format**

A data frame with 928 rows representing respondents and 9 variables:

**respondent\_id** Respondent ID

**ck\_weather** Do you typically check a daily weather report?

**weather\_source** How do you typically check the weather?

**weather\_source\_site** If they responded "A specific website or app" when asked how they typically check the weather, they were asked to write-in the app or website they used.

**ck\_weather\_watch** If you had a smartwatch (like the soon to be released Apple Watch), how likely or unlikely would you be to check the weather on that device?

**age** Age

**female** Gender

**hhold\_income** How much total combined money did all members of your HOUSEHOLD earn last year?

**region** US Region

**Source**

The source of the data is a Survey Monkey Audience poll commissioned by FiveThirtyEight and conducted from April 6 to April 10, 2015. See <https://github.com/fivethirtyeight/data/tree/master/weather-check>

---

wwc\_2019

*2019 Women's World Cup Predictions*

---

**Description**

The raw data behind the story "2019 Women's World Cup Predictions" <https://projects.fivethirtyeight.com/2019-womens-world-cup-predictions/>

**Usage**

wwc\_2019\_forecasts

wwc\_2019\_matches

**Format**

2 dataframes about the 2019 Women's World Cup matches and teams

An object of class `spec_tbl_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 52 rows and 18 columns.

**wwc\_2019\_forecasts**

A data frame with 192 rows representing 2019 Women's World Cup team match-by-match projections, and 21 variables:

**date** Date match was played

**team** Team

**group** Assigned group for the group stage

**spi** Soccer power index

**global\_o** SPI offensive rating

**global\_d** SPI defensive rating

**sim\_wins** Simulated number of wins

**sim\_ties** Simulated number of ties

**sim\_losses** Simulated number of losses

**sim\_goal\_diff** Simulated difference between goals\_scored and goals\_against

**goals\_scored** The number of goals that a team is expected to score against an average team on a neutral field

**goals\_against** The number of goals that a team is expected to concede against an average team on a neutral field

**group\_1** Chance of winning group stage game 1

**group\_2** Chance of winning group stage game 2

**group\_3** Chance of winning group stage game 3

**group\_4** Chance of winning group stage game 4

**make\_round\_of\_16** Chance of playing in the round of 16

**make\_quarters** Chance of playing in the quarter-finals

**make\_semis** Chance of playing in the semi-finals

**make\_final** Chance of playing in the finals

**win\_league** Chance of winning the tournament

**wwc\_2019\_matches**

2019 Women's World Cup Predictions A data frame with 52 rows representing Women's World Cup matches, and 18 variables:

**date** Date match was played

**team1** Team 1

**team2** Team 2

**spi1** Soccer power index of team 1

**spi2** Soccer power index of team 2

**prob1** Probability that team 1 will win match

**prob2** Probability that team 2 will win match

**prob\_tie** Probability that the teams will tie the match

**proj\_score1** Projected number of goals scored by team 1

**proj\_score2** Projected number of goals scored by team 2

**score1** Actual number of goals scored by team 1

**score2** Actual number of goals scored by team 2

**xg1** Shot-based expected goals for team 1

**xg2** Shot-based expected goals for team 2

**nsxg1** Non-shot expected goals for team 1

**nsxg2** Non-shot expected goals for team 2

**adj\_score1** Goals scored by team 1 accounting for the conditions under which each goal was scored

**adj\_score2** Goals scored by team 2 accounting for the conditions under which each goal was scored

### Source

[https://projects.fivethirtyeight.com/soccer-api/international/2019/wwc\\_forecasts.csv](https://projects.fivethirtyeight.com/soccer-api/international/2019/wwc_forecasts.csv)

[https://projects.fivethirtyeight.com/soccer-api/international/2019/wwc\\_matches.csv](https://projects.fivethirtyeight.com/soccer-api/international/2019/wwc_matches.csv)

# Index

## \* datasets

ahca\_polls, 5  
airline\_safety, 6  
antiquities\_act, 7  
august\_senate\_polls, 8  
avengers, 9  
bachelorette, 10  
bad\_drivers, 11  
bechdel, 12  
biopics, 13  
bob\_ross, 14  
cabinet\_turnover, 17  
cand\_events\_20150114, 18  
cand\_events\_20150130, 19  
cand\_state\_20150114, 20  
cand\_state\_20150130, 21  
candy\_rankings, 17  
chess\_transfers, 21  
classic\_rock\_raw\_data, 22  
classic\_rock\_song\_list, 23  
college\_all\_ages, 23  
college\_grad\_students, 24  
college\_recent\_grads, 25  
comma\_survey, 27  
congress\_age, 28  
cousin\_marriage, 29  
daily\_show\_guests, 29  
datasets\_master, 30  
dem\_candidates, 31  
democratic\_bench, 30  
drinks, 34  
drug\_use, 35  
elasticity\_by\_district, 37  
elasticity\_by\_state, 38  
elo\_blatter, 38  
endorsements, 39  
endorsements\_2020, 40  
fandango, 41  
fifa\_audience, 42  
fight\_songs, 43  
flying, 44  
food\_world\_cup, 46  
forecast\_results\_2018, 48  
foul\_balls, 49  
generic\_polllist, 50  
generic\_topline, 51  
google\_trends, 52  
governor\_national\_forecast, 53  
governor\_state\_forecast, 54  
hate\_crimes, 55  
hiphop\_cand\_lyrics, 56  
hist\_ncaa\_bball\_casts, 56  
hist\_senate\_preds, 57  
house\_national\_forecast, 58  
impeachment\_polls, 59  
librarians, 60  
love\_actually\_adj, 61  
love\_actually\_appearance, 62  
mad\_men, 63  
male\_flight\_attend, 64  
masculinity\_survey, 64  
media\_mentions\_2020, 69  
mediacloud\_hurricanes, 66  
mediacloud\_online\_news, 67  
mediacloud\_states, 67  
mediacloud\_trump, 68  
mlb\_as\_play\_talent, 70  
mlb\_as\_team\_talent, 71  
mueller\_approval\_polls, 72  
murder\_2015\_final, 73  
murder\_2016\_prelim, 73  
nba\_draft\_2015, 74  
nba\_draymond, 75  
nba\_elo, 75  
nba\_tattoos, 76  
ncaa\_w\_bball\_tourney, 77  
nfl\_fandom\_google, 80  
nfl\_fandom\_surveymonkey, 81

- nfl\_fav\_team, 83
- nfl\_suspensions, 84
- nfltix\_div\_avgprice, 78
- nfltix\_usa\_avg, 78
- nflwr\_aging\_curve, 79
- nflwr\_hist, 79
- nutrition\_pvalues, 84
- partisan\_lean\_district, 85
- partisan\_lean\_state, 86
- police\_deaths, 87
- police\_killings, 87
- police\_locals, 89
- pres\_2016\_trail, 90
- pres\_commencement, 91
- pulitzer, 91
- riddler\_castles, 92
- riddler\_castles2, 93
- riddler\_pick\_lowest, 94
- russia\_investigation, 95
- san\_andreas, 98
- sandy\_311, 96
- senate\_national\_forecast, 99
- senate\_polls, 100
- senate\_seat\_forecast, 100
- spi\_global\_rankings, 101
- state\_info, 102
- state\_of\_the\_state, 103
- steak\_survey, 104
- tarantino, 105
- tennis\_events\_time, 106
- tennis\_players\_time, 106
- tennis\_serve\_time, 107
- tenth\_circuit, 108
- trump\_approval\_poll, 112
- trump\_approval\_trend, 113
- trump\_lawsuits, 114
- trump\_news, 115
- trump\_twitter, 116
- trumpworld\_issues, 109
- trumpworld\_polls, 110
- tv\_hurricanes, 116
- tv\_hurricanes\_by\_network, 117
- tv\_states, 118
- undefeated, 119
- unisex\_names, 119
- US\_births\_1994\_2003, 120
- US\_births\_2000\_2014, 121
- weather\_check, 121
- wwc\_2019, 122
- ahca\_polls, 5
- airline\_safety, 6
- antiquities\_act, 7
- august\_senate\_polls, 8
- avengers, 9
- bachelorette, 10
- bad\_drivers, 11
- bechdel, 12
- biopics, 13
- bob\_ross, 14
- cabinet\_turnover, 17
- cand\_events\_20150114, 18, 20, 21
- cand\_events\_20150130, 19, 19, 20, 21
- cand\_state\_20150114, 19, 20, 20, 21
- cand\_state\_20150130, 19, 20, 21
- candy\_rankings, 17
- chess\_transfers, 21
- classic\_rock\_raw\_data, 22, 23
- classic\_rock\_song\_list, 22, 23
- college\_all\_ages, 23, 25, 26
- college\_grad\_students, 24, 24, 26
- college\_recent\_grads, 24, 25, 25
- comma\_survey, 27
- congress\_age, 28
- cousin\_marriage, 29
- daily\_show\_guests, 29
- datasets\_master, 30
- dem\_candidates, 31
- democratic\_bench, 30
- drinks, 34
- drug\_use, 35
- elasticity\_by\_district, 37, 38
- elasticity\_by\_state, 37, 38
- elo\_blatter, 38
- endorsements, 39
- endorsements\_2020, 40
- fandango, 41
- fifa\_audience, 42
- fight\_songs, 43
- fivethirtyeight, 44
- flying, 44
- food\_world\_cup, 46
- forecast\_results\_2018, 48

- foul\_balls, 49
- generic\_polllist, 50, 51
- generic\_topline, 51, 51
- google\_trends, 52, 66–69, 117, 118
- governor\_national\_forecast, 53, 54
- governor\_state\_forecast, 53, 54
- hate\_crimes, 55
- hiphop\_cand\_lyrics, 56
- hist\_ncaa\_bball\_casts, 56
- hist\_senate\_preds, 57
- house\_district\_forecast, 58
- house\_national\_forecast, 58
- impeachment\_polls, 59
- librarians, 60
- love\_actually\_adj, 61, 62
- love\_actually\_appearance, 61, 62
- mad\_men, 63
- male\_flight\_attend, 64
- masculinity\_survey, 64
- media\_mentions\_2020, 69
- media\_mentions\_cable  
(media\_mentions\_2020), 69
- media\_mentions\_online  
(media\_mentions\_2020), 69
- mediacloud\_hurricanes, 52, 66, 67–69, 117, 118
- mediacloud\_online\_news, 52, 66, 67, 68, 69, 117, 118
- mediacloud\_states, 52, 66, 67, 67, 69, 117, 118
- mediacloud\_trump, 52, 66–68, 68, 117, 118
- mlb\_as\_play\_talent, 70
- mlb\_as\_team\_talent, 71
- mueller\_approval\_polls, 72
- murder\_2015\_final, 73
- murder\_2016\_prelim, 73
- nba\_draft\_2015, 74
- nba\_draymond, 75
- nba\_elo, 75
- nba\_elo\_latest (nba\_elo), 75
- nba\_tattoos, 76
- ncaa\_w\_bball\_tourney, 77
- nfl\_fandom\_google, 80, 82
- nfl\_fandom\_surveymonkey, 80, 81
- nfl\_fav\_team, 83
- nfl\_suspensions, 84
- nfltix\_div\_avgprice, 78
- nfltix\_usa\_avg, 78
- nflwr\_aging\_curve, 79
- nflwr\_hist, 79
- nutrition\_pvalues, 84
- partisan\_lean\_district, 85, 86
- partisan\_lean\_state, 86, 86
- police\_deaths, 87
- police\_killings, 87
- police\_locals, 89
- pres\_2016\_trail, 90
- pres\_commencement, 91
- pulitzer, 91
- riddler\_castles, 92, 94
- riddler\_castles2, 93, 93
- riddler\_pick\_lowest, 94
- russia\_investigation, 95
- san\_andreas, 98
- sandy\_311, 96
- senate\_national\_forecast, 99, 101
- senate\_polls, 100
- senate\_seat\_forecast, 99, 100
- spi\_global\_rankings, 101
- spi\_matches, 102
- state\_index (state\_of\_the\_state), 103
- state\_info, 102
- state\_of\_the\_state, 103
- state\_words (state\_of\_the\_state), 103
- steak\_survey, 104
- tarantino, 105
- tennis\_events\_time, 106, 107
- tennis\_players\_time, 106, 106, 107
- tennis\_serve\_time, 106, 107, 107
- tenth\_circuit, 108
- trump\_approval\_poll, 112, 114
- trump\_approval\_trend, 113, 113
- trump\_lawsuits, 114
- trump\_news, 115
- trump\_twitter, 116
- trumpworld\_issues, 109, 111
- trumpworld\_polls, 109, 110
- tv\_hurricanes, 52, 66–69, 116, 118
- tv\_hurricanes\_by\_network, 52, 66–69, 117, 117, 118

tv\_states, [52](#), [66–69](#), [117](#), [118](#), [118](#)

undefeated, [119](#)

unisex\_names, [119](#)

US\_births\_1994\_2003, [120](#), [121](#)

US\_births\_2000\_2014, [120](#), [121](#)

weather\_check, [121](#)

wwc\_2019, [122](#)

wwc\_2019\_forecasts (wwc\_2019), [122](#)

wwc\_2019\_matches (wwc\_2019), [122](#)