Package ‘fivethirtyeight’

October 7, 2021

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.ys.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] (https://orcid.org/0000-0001-7824-306X), Chester Ismay [aut] (https://orcid.org/0000-0003-2820-2547), Jennifer Chunn [aut], Meredith Manley [ctb] (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] (<https://orcid.org/0000-0002-3279-0516>),
Rana Gahwagy [ctb] (<https://orcid.org/0000-0002-6331-2840>),
Natalia Iannucci [ctb] (<https://orcid.org/0000-0001-5663-1731>),
Marium Tapal [ctb] (<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

R topics documented:

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
R topics documented:

- 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
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>nba_draft_2015</td>
<td>74</td>
</tr>
<tr>
<td>nba_draymond</td>
<td>75</td>
</tr>
<tr>
<td>nba_elo</td>
<td>75</td>
</tr>
<tr>
<td>nba_tattoos</td>
<td>76</td>
</tr>
<tr>
<td>ncaa_w_bbball_tourney</td>
<td>77</td>
</tr>
<tr>
<td>nfltx_div_avgprice</td>
<td>78</td>
</tr>
<tr>
<td>nfltx_usa_avg</td>
<td>78</td>
</tr>
<tr>
<td>nflwr_aging_curve</td>
<td>79</td>
</tr>
<tr>
<td>nflwr_hist</td>
<td>79</td>
</tr>
<tr>
<td>nfl_fandom_google</td>
<td>80</td>
</tr>
<tr>
<td>nfl_fandom_surveymonkey</td>
<td>81</td>
</tr>
<tr>
<td>nfl_fav_team</td>
<td>83</td>
</tr>
<tr>
<td>nfl_suspensions</td>
<td>84</td>
</tr>
<tr>
<td>nutrition_pvalues</td>
<td>84</td>
</tr>
<tr>
<td>partisan_lean_district</td>
<td>85</td>
</tr>
<tr>
<td>partisan_lean_state</td>
<td>86</td>
</tr>
<tr>
<td>police_deaths</td>
<td>87</td>
</tr>
<tr>
<td>police_killings</td>
<td>87</td>
</tr>
<tr>
<td>police_locals</td>
<td>89</td>
</tr>
<tr>
<td>pres_2016_trail</td>
<td>90</td>
</tr>
<tr>
<td>pres_commencement</td>
<td>91</td>
</tr>
<tr>
<td>pulitzer</td>
<td>91</td>
</tr>
<tr>
<td>riddler_castles</td>
<td>92</td>
</tr>
<tr>
<td>riddler_castles2</td>
<td>93</td>
</tr>
<tr>
<td>riddler_pick_lowest</td>
<td>94</td>
</tr>
<tr>
<td>russia_investigation</td>
<td>95</td>
</tr>
<tr>
<td>sandy_311</td>
<td>96</td>
</tr>
<tr>
<td>san_andreas</td>
<td>98</td>
</tr>
<tr>
<td>senate_national_forecast</td>
<td>99</td>
</tr>
<tr>
<td>senate_polls</td>
<td>100</td>
</tr>
<tr>
<td>senate_seat_forecast</td>
<td>100</td>
</tr>
<tr>
<td>spi_global_rankings</td>
<td>101</td>
</tr>
<tr>
<td>state_info</td>
<td>102</td>
</tr>
<tr>
<td>state_of_the_state</td>
<td>103</td>
</tr>
<tr>
<td>steak_survey</td>
<td>104</td>
</tr>
<tr>
<td>tarantino</td>
<td>105</td>
</tr>
<tr>
<td>tennis_events_time</td>
<td>106</td>
</tr>
<tr>
<td>tennis_players_time</td>
<td>106</td>
</tr>
<tr>
<td>tennis_serve_time</td>
<td>107</td>
</tr>
<tr>
<td>tenth_circuit</td>
<td>108</td>
</tr>
<tr>
<td>trumpworld_issues</td>
<td>109</td>
</tr>
<tr>
<td>trumpworld_polls</td>
<td>110</td>
</tr>
<tr>
<td>trump_approval_poll</td>
<td>112</td>
</tr>
<tr>
<td>trump_approval_trend</td>
<td>113</td>
</tr>
<tr>
<td>trump_lawsuits</td>
<td>114</td>
</tr>
<tr>
<td>trump_news</td>
<td>115</td>
</tr>
<tr>
<td>trump_twitter</td>
<td>116</td>
</tr>
<tr>
<td>tv_hurricanes</td>
<td>116</td>
</tr>
</tbody>
</table>
**Description**


**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](https://github.com/fivethirtyeight/data/blob/master/ahca-polls/README.md)

**Examples**

```r
# 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")
```
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-crashes-in-the-past/](https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had-crashes-in-the-past/).

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


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**

---

*Trump Might Be The First President To Scrap A National Monument*

**Description**


**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
- **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/](https://www.npca.org/) and National Parks Service Archeology Program [https://www.nps.gov/history/archeology/sites/antiquities/MonumentsList.htm](https://www.nps.gov/history/archeology/sites/antiquities/MonumentsList.htm)
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 date 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**  

**Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building**

---

**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.

<table>
<thead>
<tr>
<th>bachelorette</th>
<th>Bachelor / Bachelorette</th>
</tr>
</thead>
</table>

Description


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 received 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 and then missing seasons were filled in by ABC and FiveThirtyEight staffers.

bad_driversDear 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


---

**bechdel**

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

Description


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
biopics

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


---

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/](https://www.imdb.com/)

---

**Description**


**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

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-historically-high/].

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

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 crispy 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

```r
# 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?"

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](https://github.com/fivethirtyeight/data/tree/master/potential-candidates)

**See Also**

cand_state_20150114, cand_events_20150130, and cand_state_20150130

---

**Description**


**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
cand_state_20150114

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](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**

c 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](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 Also

- college_grad_students, college_recent_grads

Description


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
college_recent_grads

- **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

Description


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 (WK == 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](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](#)
Description


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

### congress_age

table: congress_age

<table>
<thead>
<tr>
<th>congress_number</th>
<th>chamber</th>
<th>bioguide</th>
<th>firstname</th>
<th>middlename</th>
<th>lastname</th>
<th>suffix</th>
<th>birthday</th>
<th>state</th>
<th>party</th>
<th>incumbent</th>
<th>termstart</th>
<th>age</th>
</tr>
</thead>
</table>

**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

How Many Americans Are Married To Their Cousins?

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
Every Guest Jon Stewart Ever Had On 'The Daily Show'

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

Master list of all datasets

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

Some Democrats Who Could Step Up If Hillary Isn’t Ready For Hillary

Description


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**


---

**dem_candidates**  
Democratic Primary Candidates 2018

**Description**


**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.
dem_candidates

**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 a 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 specifi-
cally 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 pri-
mary, 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 Jus-
tice 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 Com-
nitee 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 specifi-
cally 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, ac-
cording 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 pri-
mary, 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-the-progressive-wing-in-democratic-primaries-so-far/](https://fivethirtyeight.com/features/the-establishment-is-beating-the-progressive-wing-in-democratic-primaries-so-far/)

**Source**


---

**drinks**

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

**Description**


**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
drug_use

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)

Description


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
drug_use

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)
Description


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**

**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
```
endorsements

**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**


<table>
<thead>
<tr>
<th>endorsements</th>
<th>Pols And Polls Say The Same Thing: Jeb Bush Is A Weak Front-Runner</th>
</tr>
</thead>
</table>

**Description**


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
- **moneyRaised**  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_voted_for**  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  The 2020 Endorsement Primary - Which Democratic candidates are receiving the most support from prominent members of their party?

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

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.

---

**Description**


---

**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](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)
**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/]
Description

An R package that provides access to the code and data sets published by FiveThirtyEight [https://github.com/fivethirtyeight/data](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](https://fivethirtyeight-r.netlify.app/articles/fivethirtyeight.html)

Examples

```r
# 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")
```

Usage

```r
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
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-world-cup/](https://fivethirtyeight.com/features/the-fivethirtyeight-international-food-associations-2014-world-cup/). 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 country’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
See Also

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


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

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-dangerous-ones/.

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/.
Description


Usage

generic_polllist

Format

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

- `subgroup`: No description provided.
- `modedate`: 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  Congress Generic Ballot Polls

Description

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
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
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"

**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_popMetro** 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**

```r
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**


**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/](https://genius.com/)

---

**hist_ncaa_bball_casts**  
*The NCAA Bracket: Checking Our Work*

**Description**


**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


---

Hist Senate Preds

How The FiveThirtyEight Senate Forecast Model Works

Description


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](https://github.com/fivethirtyeight/data/tree/master/forecast-methodology)
house_national_forecast

2018 House Forecast

Description


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


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/](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


---

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
## Description

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

## Usage

```r
love_actually_adj
```

## Format

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

<table>
<thead>
<tr>
<th>actors</th>
<th>bill_nighy</th>
<th>keira_knightley</th>
<th>andrew_lincoln</th>
<th>hugh_grant</th>
<th>colin_firth</th>
<th>alan_rickman</th>
<th>heike_makatsch</th>
<th>laura_linney</th>
<th>emma_thompson</th>
<th>liam_neeson</th>
<th>kris_marshall</th>
<th>abdul_salis</th>
<th>martin_freeman</th>
<th>rowan_atkinson</th>
</tr>
</thead>
<tbody>
<tr>
<td>love_actually_adj</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See Also

`love_actually.Appearance`
love_actually_appearance

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

Description


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)

Description


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


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

```r
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?")

---

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

Description


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/](https://mediacloud.org/)

See Also

mediacloud_states, mediacloud_online_news, mediacloud_trump, tv_hurricanes, tv_hurricanes_by_network, tv_states, google_trends
**Description**


**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/](https://mediacloud.org/)

**See Also**

*mediacloud_hurricanes, mediacloud_states, mediacloud_trump, tv_hurricanes, tv_hurricanes_by_network, tv_states, google_trends*

---

**Description**


**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/](https://mediacloud.org/)

**See Also**

mediacloud_hurricanes, mediacloud_online_news, mediacloud_trump, tv_hurricanes, tv_hurricanes_by_network, tv_states, google_trends

---

**Description**


**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/](https://mediacloud.org/)

See Also

[mediacloud_hurricanes, mediacloud_states, mediacloud_online_news, tv_hurricanes, tv_hurricanes_by_network, tv_states, google_trends](https://mediacloud.org/)

---

**media_mentions_2020**

2020 Presidential Candidates Media Mentions

**Description**


**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


---

mlb_as_play_talent  The Best MLB All-Star Teams Ever

Description


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

Description

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
gid: 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**


---

**Description**

The raw data behind the story 'Both Parties Think The Mueller Report Was Fair. They Just Completely Disagree On What It Says.'

**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 counsel 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


Description


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

Description


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 murders 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"  
An analysis using this data was contributed by G. Elliott Morris as a package vignette at [https://fivethirtyeightdata.github.io/fivethirtyeightdata/articles/NBA.html](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 >= +3.3)
- **starter**: Probability of becoming a starting-caliber player (10 per draft, SPM >= +0.5)
- **role_player**: Probability of becoming a role player (25 per draft, SPM >= -1.4)
- **bust**: Probability of becoming a bust (everyone else, SPM < -1.4)
**nba_draymond**

**Source**

**Description**

**Usage**
nba_draymond

**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
- **draymond** Defensive Rating Accounting for Yielding Minimal Openness by Nearest Defender

**Source**
see https://github.com/fivethirtyeight/data/tree/master/nba-draymond

---

**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

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/]
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
Who Goes To Meaningless NFL Games And Why?

Description

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

---

Who Goes To Meaningless NFL Games And Why?

Description

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
The Football Hall Of Fame Has A Receiver Problem

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

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

The Football Hall Of Fame Has A Receiver Problem

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

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_woovy  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

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

```r
# 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)
```

Description


Usage

```r
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 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)
The Rams Are Dead To Me, So I Answered 3,352 Questions To Find A New NFL Team

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-team/.

Usage

Usage

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  The NFL’s Uneven History Of Punishing Domestic Violence

Description


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


nutrition_pvalues  You Can’t Trust What You Read About Nutrition

Description


Usage

nutrition_pvalues
**partisan_lean_district**

**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**


---

**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

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
The Dallas Shooting Was Among The Deadliest For Police In U.S. History

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-in-u-s-history/.

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/

Where Police Have Killed Americans In 2015

Description


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](https://github.com/fivethirtyeight/data/tree/master/police-killings)
Description


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](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")

Description

The raw data behind the story "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

| pres_commencement | Sitting Presidents Give Way More Commencement Speeches Than They Used To |

**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-than-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

| pulitzer | Do Pulitzers Help Newspapers Keep Readers? |

**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


---

**riddler_castles**  
*Can You Rule Riddler Nation?*

Description


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?
riddler_castles2

Source

See https://github.com/fivethirtyeight/data/tree/master/riddler-castles

See Also

riddler_castles2

Examples

```r
# 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?
riddler_pick_lowest

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
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.


dataset in GitHub: https://github.com/fivethirtyeight/data/tree/master/russia-investigation

<table>
<thead>
<tr>
<th>sandy_311</th>
<th>The (Very) Long Tail Of Hurricane Recovery</th>
</tr>
</thead>
</table>

**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
**sandy_311**

- **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
- **nyccha** 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](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](https://github.com/fivethirtyeight/data/tree/master/sandy-311-calls)

**Examples**

```r
# 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") %>%
arange(date) %>%
select(date, agency, num_calls, total) %>%
rename(total_calls = total) %>%
mutate(agency = as.factor(agency))
```
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-big-one/].

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]
Description

This file contains links to the data behind FiveThirtyEight’s 'Senate forecasts' [https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate/](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


See Also

senate_seat_forecast
senate_polls  Early Senate Polls Have Plenty to Tell Us About November

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  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_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**


**See Also**

- `senate_national_forecast`
State Information

**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](https://github.com/fivethirtyeight/data/blob/master/soccer-spi/README.md)

**See Also**
- `spi_matches`

---

**State Information**

**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**
Examples

library(dplyr)
# Number of states in each division
state_info %>%
  count(division)
# Number of states in each region
state_info %>%
  count(region)

Description

The raw data behind the story "What America’s Governors Are Talking About" [https://fivethirtyeight.com/features/what-americas-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
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^2 test

Source


---

steak_survey

How Americans Like Their Steak

Description


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/](https://fivethirtyeight.com/features/how-americans-like-their-steak/)

---

**Description**


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](https://github.com/fivethirtyeight/data/tree/master/tarantino)
tennis_events_time

Why Some Tennis Matches Take Forever

Description

Usage
tenis_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
tenis_players_time and tennis_serve_time

---

tenis_players_time

Why Some Tennis Matches Take Forever

Description

Usage
tenis_players_time
tennis_serve_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

---

Description

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
For A Trump Nominee, Neil Gorsuch’s Record Is Surprisingly Moderate On Immigration

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-surprisingly-moderate-on-immigration/].

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]
Description


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


See Also

trumpworld_polls
Description


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
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
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
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
russia 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


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/](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 approve 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


See Also

trump_approval_trend

---

Description

The raw data behind the story: "How Popular is Donald Trump" https://projects.fivethirtyeight.com/trump-approval-ratings/: Approval Trend Dataset.

Usage

trumpApprovalTrend

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)
- **model_date** 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

See Also

trump_approval_poll

trump_lawsuits

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
docket_number PACER code for nature of lawsuit (Not relevant for criminal cases) https://pacer.uscourts.gov/help/faqs/what-nature-suit-code

Description

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


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/](https://www.memeorandum.com/).
**trump_twitter**

The World’s Favorite Donald Trump Tweets

**Description**


**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](https://twitter.com/realdonaldtrump)

---

**tv_hurricanes**

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

**Description**


**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

Description


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-

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
**Mayweather Is Defined By The Zero Next To His Name**

**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/](https://fivethirtyeight.com/features/mayweather-is-defined-by-the-zero-next-to-his-name/)

**Usage**

**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

---

**The Most Common Unisex Names In America: Is Yours One Of Them?**

**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-them/](https://fivethirtyeight.com/features/there-are-922-unisex-names-in-america-is-yours-one-of-them/)

**Usage**

**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**


**US_births_1994_2003**

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-friday-the-13th/](https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/).

**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**

### 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-friday-the-13th/](https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/).

### 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


---

**weather_check**

### 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


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_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
  bobRoss, 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_gbball_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
  mediachoice_hurricanes, 66
  mediachoice_online_news, 67
  mediachoice_states, 67
  mediachoice_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_gbball_tourney, 77
  nfl_fandom_google, 80
  nfl_fandom_survey_monkey, 81
INDEX

foul_balls, 49

generic_polllist, 50, 51
generic_topline, 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, 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_elgo, 75

nba_elgo_latest (nba_elgo), 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

nfltx_div_avgprice, 78

nfltx_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