Package 'fivethirtyeight'

October 7, 2018

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 0.4.0

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 6.0.1

Suggests fivethirtyeight, tidyverse, lubridate, stringr, magrittr, knitr, rmarkdown, broom, scales, tidytext, ggthemes, hunspell, grid, fmsb, wordcloud, gridExtra, corrplot, ggraph, igraph, highcharter, janitor

VignetteBuilder knitr

NeedsCompilation no

Author Albert Y. Kim [aut, cre], Chester Ismay [aut], Jennifer Chunn [aut], Meredith Manley [ctb], Maggie Shea [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]

Repository CRAN

Date/Publication 2018-02-11 17:34:04 UTC

R topics documented:

`ahca_polls` *American Health Care Act Polls*

Description

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

Usage

```
ahca_polls
```

Format

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

start Start date of the poll.

end End date of the poll.

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

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

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

url The website associated with the polling question.

text The polling question asked at the pollster.

Source

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

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
library(stringr)
ahca_polls_tidy <- ahca_polls %>%
  gather(opinion, count, -c(start, end, pollster, text, url))
```

`airline_safety` *Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?*

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-h

Usage

```
airline_safety
```

Format

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

airline airline

incl_reg_subsidiaries indicates that regional subsidiaries are included

avail_seat_km_per_week available seat kilometers flown every week

incidents_85_99 Total number of incidents, 1985-1999

fatal_accidents_85_99 Total number of fatal accidents, 1985-1999

fatalities_85_99 Total number of fatalities, 1985-1999

incidents_00_14 Total number of incidents, 2000-2014

fatal_accidents_00_14 Total number of fatal accidents, 2000-2014

fatalities_00_14 Total number of fatalities, 2000-2014

Source

Aviation Safety Network http://aviation-safety.net.

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
library(stringr)
airline_safety_tidy <- airline_safety %>%
  gather(type, count, -c(airline, incl_reg_subsidiaries, avail_seat_km_per_week)) %>%
  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

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

Usage

```
antiquities_act
```

Format

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

current_name Current name of piece of land designated under the Antiquities Act

states State(s) or territory where land is located

original_name If included, original name of piece of land designated under the Antiquities Act

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

action Type of action taken on land

date Date of action

year Year of action

pres_or_congress President or congress that issued action

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

Source

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

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.

`bachelorette` *Bachelorette / Bachelor*

Description

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

Usage

```
bachelorette
```

Format

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

show Bachelor or Bachelorette.

season Which season?

contestant An identifier for the contestant in a given season.

elimination_1 Who was eliminated in week 1.

elimination_2 Who was eliminated in week 2.

elimination_3 Who was eliminated in week 3.

elimination_4 Who was eliminated in week 4.

elimination_5 Who was eliminated in week 5.

elimination_6 Who was eliminated in week 6.

elimination_7 Who was eliminated in week 7.

elimination_8 Who was eliminated in week 8.

elimination_9 Who was eliminated in week 9.

elimination_10 Who was eliminated in week 10.

dates_1 Who was on which date in week 1.

dates_2 Who was on which date in week 2.

dates_3 Who was on which date in week 3.

dates_4 Who was on which date in week 4.

dates_5 Who was on which date in week 5.

dates_6 Who was on which date in week 6.

dates_7 Who was on which date in week 7.

dates_8 Who was on which date in week 8.

dates_9 Who was on which date in week 9.

dates_10 Who was on which date in week 10.

Details

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

Source

http://bachelor-nation.wikia.com/wiki/Bachelor_Nation_Wikia and then missing seasons were filled in by ABC and FiveThirtyEight staffers.

---

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

Description

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

Usage

```
bad_drivers
```

Format

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

state State

num_drivers Number of drivers involved in fatal collisions per billion miles

perc_speeding Percentage of drivers involved in fatal collisions who were speeding

perc_alcohol Percentage of drivers involved in fatal collisions who were alcohol-impaired

perc_not_distracted Percentage of drivers involved in fatal collisions who were not distracted

perc_no_previous Percentage of drivers involved in fatal collisions who had not been involved in any previous accidents

insurance_premiums Car insurance premiums ($)

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

Source

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

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

Description

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

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: http://www.imdb.com/title/tt1711425

title Movie test

test bechdel test result (detailed, with discrepancies indicated)

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

binary Bechdel Test PASS vs FAIL binary

budget Film budget

domgross Domestic (US) gross

intgross Total International (i.e., worldwide) gross

code Bechdel Code

budget_2013 Budget in 2013 inflation adjusted dollars

domgross_2013 Domestic gross (US) in 2013 inflation adjusted dollars

intgross_2013 Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars

period_code

decade_code

Details

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

Source

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

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

Description

The raw data behind the story "'Straight Outta Compton' Is The Rare Biopic Not About White Dudes" https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about- An analysis using this data was contributed by Pradeep Adhokshaja as a package vignette at http://fivethirtyeight-r.netlify.com/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: http://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 http://www.imdb.com/

bob_ross *A Statistical Analysis of the Work of Bob Ross*

Description

The raw data behind the story "A Statistical Analysis of the Work of Bob Ross" https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/. An analysis using this data was contributed by Jonathan Bouchet as a package vignette at http://fivethirtyeight-r.netlify.com/articles/bob_ross.html.

Usage

```
bob_ross
```

Format

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

episode Episode code

season Season number

episode_num Episode number

title Title of episode

apple_frame Present (1) or not (0)

aurora_borealis Present (1) or not (0)

barn Present (1) or not (0)

beach Present (1) or not (0)

boat Present (1) or not (0)

bridge Present (1) or not (0)

building Present (1) or not (0)

bushes Present (1) or not (0)

cabin Present (1) or not (0)

cactus Present (1) or not (0)

circle_frame Present (1) or not (0)

cirrus Present (1) or not (0)

cliff Present (1) or not (0)

clouds Present (1) or not (0)

conifer Present (1) or not (0)

cumulus Present (1) or not (0)

deciduous Present (1) or not (0)

diane_andre Present (1) or not (0)

dock Present (1) or not (0)

double_oval_frame Present (1) or not (0)

farm Present (1) or not (0)

fence Present (1) or not (0)

fire Present (1) or not (0)
florida_frame Present (1) or not (0)
flowers Present (1) or not (0)
fog Present (1) or not (0)
framed Present (1) or not (0)
grass Present (1) or not (0)
guest Present (1) or not (0)
half_circle_frame Present (1) or not (0)
half_oval_frame Present (1) or not (0)
hills Present (1) or not (0)
lake Present (1) or not (0)
lakes Present (1) or not (0)
lighthouse Present (1) or not (0)
mill Present (1) or not (0)
moon Present (1) or not (0)
mountain Present (1) or not (0)
mountains Present (1) or not (0)
night Present (1) or not (0)
ocean Present (1) or not (0)
oval_frame Present (1) or not (0)
palm_trees Present (1) or not (0)
path Present (1) or not (0)
person Present (1) or not (0)
portrait Present (1) or not (0)
rectangle_3d_frame Present (1) or not (0)
rectangular_frame Present (1) or not (0)
river Present (1) or not (0)
rocks Present (1) or not (0)
seashell_frame Present (1) or not (0)
snow Present (1) or not (0)
snowy_mountain Present (1) or not (0)
split_frame Present (1) or not (0)
steve_ross Present (1) or not (0)
structure Present (1) or not (0)
sun Present (1) or not (0)
tomb_frame Present (1) or not (0)
tree Present (1) or not (0)
trees Present (1) or not (0)
triple_frame Present (1) or not (0)
waterfall Present (1) or not (0)
waves Present (1) or not (0)
windmill Present (1) or not (0)
window_frame Present (1) or not (0)
winter Present (1) or not (0)
wood_framed Present (1) or not (0)

Source

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

Examples

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

---

candy_rankings *Candy Power Ranking*

Description

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

Usage

```
candy_rankings
```

Format

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

competitorname The name of the Halloween candy.

chocolate Does it contain chocolate?

fruity Is it fruit flavored?

caramel Is there caramel in the candy?

peanutyalmondy Does it contain peanuts, peanut butter or almonds?

nougat Does it contain nougat?

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

hard Is it a hard candy?

bar Is it a candy bar?

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

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

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

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

Source

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

Examples

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

---

cand_events_20150114 *Looking For Clues: Who Is Going To Run For President In 2016?*

Description

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

Usage

```
cand_events_20150114
```

Format

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

person Potential presidential candidate

party Political party

state State of event

event Name of event

type Type of event

date Date of event

link Link to event

snippet Snippet of event description

Source

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

See Also

cand_state_20150114, cand_events_20150130, and cand_state_20150130

cand_events_20150130 *Who Will Run For President: Romney Is Out*

Description

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

Usage

```
cand_events_20150130
```

Format

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

person Potential presidential candidate

party Political party

state State of event

event Name of event

type Type of event

date Date of event

link Link to event

snippet Snippet of event description

Source

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

See Also

cand_state_20150130, cand_events_20150114, and cand_state_20150114

cand_state_20150114 *Looking For Clues: Who Is Going To Run For President In 2016?*

Description

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

Usage

```
cand_state_20150114
```

Format

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

person Potential presidential candidate

party Political party

date Date of event

latest Latest statement

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

Source

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

See Also

cand_events_20150114, cand_events_20150130, and cand_state_20150130

---

cand_state_20150130 *Who Will Run For President: Romney Is Out*

Description

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

Usage

```
cand_state_20150130
```

Format

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

person Potential presidential candidate

party Political party

date Date of event

latest Latest statement

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

Source

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

See Also

cand_events_20150130, cand_events_20150114, and cand_state_20150114

chess_transfers *Chess Transfers*

Description

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

Usage

```
chess_transfers
```

Format

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

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

id An numeric identifier for the chess player who transferred.

federation The current national federation of the chess player

form_fed The national federation from which the chess player has transferred.

transfer_date The date at which the transfer took place.

Source

World Chess Federation

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

Description

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

Usage

```
classic_rock_raw_data
```

Format

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

song Song name

artist Artist name

callsign Station callsign

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

date_time Time of song play in date/time format

unique_id Unique ID for each song play

combined Song and artist name combined

Source

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

See Also

classic_rock_song_list

---

classic_rock_song_list

Why Classic Rock Isn't What It Used To Be

Description

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

Usage

```
classic_rock_song_list
```

Format

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

song Song name

artist Artist name

release_year Release year as listed in SongFacts

combined Song and artist name combined

has_year Logical variable of whether release year is included

playcount Number of plays across all stations

playcount_has_year Number of plays across all stations if a year was found

Source

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

See Also

classic_rock_raw_data

college_all_ages *The Economic Guide To Picking A College Major*

Description

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

Usage

```
college_all_ages
```

Format

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

major_code Major code, FO1DP in ACS PUMS

major Major description

major_category Category of major from Carnevale et al

total Total number of people with major

employed Number employed (ESR == 1 or 2)

employed_fulltime_yearround Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

unemployed Number unemployed (ESR == 3)

unemployment_rate Unemployed / (Unemployed + Employed)

p25th 25th percentile of earnings

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

p75th 75th percentile of earnings

Source

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

See Also

college_grad_students, college_recent_grads

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

Description

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

Usage

```
college_grad_students
```

Format

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

major_code Major code, FO1DP in ACS PUMS

major Major description

major_category Category of major from Carnevale et al

grad_total Total number of people with major

grad_sample_size Sample size (unweighted) of full-time, year-round ONLY (used for earnings)

grad_employed Number employed (ESR == 1 or 2)

grad_employed_fulltime_yearround Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

grad_unemployed Number unemployed (ESR == 3)

grad_unemployment_rate Unemployed / (Unemployed + Employed)

grad_p25th 25th percentile of earnings

grad_median Median earnings of full-time, year-round workers

grad_p75th 75th percentile of earnings

nongrad_total Total number of people with major

nongrad_employed Number employed (ESR == 1 or 2)

nongrad_employed_fulltime_yearround Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

nongrad_unemployed Number unemployed (ESR == 3)

nongrad_unemployment_rate Unemployed / (Unemployed + Employed)

nongrad_p25th 25th percentile of earnings

nongrad_median Median earnings of full-time, year-round workers

nongrad_p75th 75th percentile of earnings

grad_share grad_total / (grad_total + nongrad_total)

grad_premium (grad_median-nongrad_median)/nongrad_median

Source

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

See Also

college_all_ages, college_recent_grads

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

Description

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

Usage

```
college_recent_grads
```

Format

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

rank Rank by median earnings

major_code Major code, FO1DP in ACS PUMS

major Major description

major_category Category of major from Carnevale et al

total Total number of people with major

sample_size Sample size (unweighted) of full-time, year-round ONLY (used for earnings)

men Men with major

women Women with major

sharewomen Proportion women

employed Number employed (ESR == 1 or 2)

employed_fulltime Employed 35 hours or more

employed_parttime Employed less than 35 hours

employed_fulltime_yearround Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)

unemployed Number unemployed (ESR == 3)

unemployment_rate Unemployed / (Unemployed + Employed)

p25th 25th percentile of earnings

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

p75th 75th percentile of earnings

college_jobs Number with job requiring a college degree

non_college_jobs Number with job not requiring a college degree

low_wage_jobs Number in low-wage service jobs

Source

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

See Also

`college_grad_students`, `college_all_ages`

`comic_characters` *Comic Books Are Still Made By Men, For Men And About Men*

Description

The raw data behind the story "Comic Books Are Still Made By Men, For Men And About Men" https://fivethirtyeight.com/features/women-in-comic-books/. An analysis using this data was contributed by Jonathan Bouchet as a package vignette at http://fivethirtyeight-r.netlify.com/articles/comics_gender.html.

Usage

```
comic_characters
```

Format

A data frame with 23272 rows representing characters and 16 variables:

publisher Comic publisher: DC Comics or Marvel

page_id The unique identifier for that characters page within the wikia

name The name of the character

urlslug The unique url within the wikia that takes you to the character

id The identity status of the character (Secret Identity, Public identity, [on marvel only: No Dual Identity])

align If the character is Good, Bad or Neutral

eye Eye color of the character

hair Hair color of the character

sex Sex of the character (e.g. Male, Female, etc.)

gsm If the character is a gender or sexual minority (e.g. Homosexual characters, bisexual characters)

alive If the character is alive or deceased

appearances The number of appearances of the character in comic books (as of Sep. 2, 2014. Number will become increasingly out of date as time goes on.)

first_appearance The month and year of the character's first appearance in a comic book, if available

month The month of the character's first appearance in a comic book, if available

year The year of the character's first appearance in a comic book, if available

date The date of the character's first appearance in a comic book, if available

Source

DC Wikia http://dc.wikia.com/wiki/Main_Page and Marvel Wikia http://marvel.wikia.com/Main_Page. Characters were scraped on August 24, 2014. Appearance counts were scraped on September 2, 2014. The month and year of the first issue each character appeared in was pulled on October 6, 2014.

`comma_survey` *Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma*

Description

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

Usage

```
comma_survey
```

Format

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

respondent_id Respondent ID

gender Gender

age Age

household_income Household income bracket

education Education level

location Location (census region)

more_grammar_correct In your opinion, which sentence is more grammatically correct?

heard_oxford_comma Prior to reading about it above, had you heard of the serial (or Oxford) comma?

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

write_following How would you write the following sentence?

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

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

care_proper_grammar In your opinion, how important or unimportant is proper use of grammar?

Source

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

congress_age *Both Republicans And Democrats Have an Age Problem*

Description

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

Usage

```
congress_age
```

Format

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

congress Congress number.

chamber Chamber of congress: House of Representatives or Senate.

bioguide bioguide

firstname First name

middlename Middle name

lastname Last name

suffix Suffix

birthday Birthday

state State abbreviation

party Party abbreviation

incumbent Boolean variable of whether member was an incumbent.

termstart Start date of session.

age Age at start of session.

Source

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

cousin_marriage *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

http://www.consang.net/index.php/Main_Page

---

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.

democratic_bench — *Some Democrats Who Could Step Up If Hillary Isn't Ready For Hillary*

Description

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

Usage

```
democratic_bench
```

Format

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

candidate Candidate

raised_exp Amount the candidate was expected to raise

raised_act Amount the candidate actually raised

Source

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

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

Description

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

Usage

```
drinks
```

Format

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

country country

beer_servings Servings of beer in average serving sizes per person

spirit_servings Servings of spirits in average serving sizes per person

wine_servings Servings of wine in average serving sizes per person

total_litres_of_pure_alcohol Total litres of pure alcohol per person

Source

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

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
library(stringr)
drinks_tidy <- drinks %>%
  gather(type, servings, -c(country, total_litres_of_pure_alcohol)) %>%
  mutate(
    type = str_sub(type, start=1, end=-10)
  ) %>%
  arrange(country, type)
```

---

drug_use *How Baby Boomers Get High*

Description

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

Usage

```
drug_use
```

Format

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

age Age group

n Number of people surveyed

alcohol_use Percentage who used alcohol

alcohol_freq Median number of times a user used alcohol

marijuana_use Percentage who used marijuana

marijuana_freq Median number of times a user used marijuana

cocaine_use Percentage who used cocaine

cocaine_freq Median number of times a user used cocaine

crack_use Percentage who used crack

crack_freq Median number of times a user used crack

heroin_use Percentage who used heroin

heroin_freq Median number of times a user used heroin

hallucinogen_use Percentage who used hallucinogens

hallucinogen_freq Median number of times a user used hallucinogens

inhalant_use Percentage who used inhalants

inhalant_freq Median number of times a user used inhalants

pain_releiver_use Percentage who used pain relievers

pain_releiver_freq Median number of times a user used pain relievers

oxycontin_use Percentage who used oxycontin

oxycontin_freq Median number of times a user used oxycontin

tranquilizer_use Percentage who used tranquilizer

tranquilizer_freq Median number of times a user used tranquilizer

stimulant_use Percentage who used stimulants

stimulant_freq Median number of times a user used stimulants

meth_use Percentage who used meth

meth_freq Median number of times a user used meth

sedative_use Percentage who used sedatives

sedative_freq Median number of times a user used sedatives

Source

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

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
library(stringr)
use <- drug_use %>%
  select(age, n, ends_with("_use")) %>%
  gather(drug, use, -c(age, n)) %>%
  mutate(drug = str_sub(drug, start=1, end=-5))
freq <- drug_use %>%
  select(age, n, ends_with("_freq")) %>%
  gather(drug, freq, -c(age, n)) %>%
  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)
```

---

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

Description

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

Usage

```
elo_blatter
```

Format

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

country FIFA member country

elo98 The team's Elo in 1998

elo15 The team's Elo in 2015

confederation Confederation to which country belongs

gdp06 The country's purchasing power parity GDP as of 2006

popu06 The country's 2006 population

gdp_source Source for gdp06

popu_source Source for popu06

Source

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

---

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

Description

The raw data behind the story "Pols And Polls Say The Same Thing: Jeb Bush Is A Weak Front-Runner" https://fivethirtyeight.com/features/pols-and-polls-say-the-same-thing-jeb-bush-is-a-wea This data includes something we call "endorsement points," an attempt to quantify the importance of endorsements by weighting each one according to the position held by the endorser: 10 points for each governor, 5 points for each senator and 1 point for each representative

Usage

```
endorsements
```

Format

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

year Election year

party Political party

candidate Candidate running in primary

endorsement_points Weighted endorsements through June 30th of the year before the primary

percentage_endorsement_points Percentage of total weighted endorsement points for the candidate's political party through June 30th of the year before the primary

money_raised Money raised through June 30th of the year before the primary

percentage_of_money Percentage of total money raised by the candidate's political party through June 30th of the year before the primary

primary_vote_percentage Percentage of votes won in the primary

won_primary Did the candidate win the primary?

Source

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

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

Description

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

Usage

```
fandango
```

Format

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

film The film in question

year Year of film

rottentomatoes The Rotten Tomatoes Tomatometer score for the film

rottentomatoes_user The Rotten Tomatoes user score for the film

metacritic The Metacritic critic score for the film

metacritic_user The Metacritic user score for the film

imdb The IMDb user score for the film

fandango_stars The number of stars the film had on its Fandango movie page

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

rt_norm The Rotten Tomatoes Tomatometer score for the film , normalized to a 0 to 5 point system

rt_user_norm The Rotten Tomatoes user score for the film , normalized to a 0 to 5 point system

metacritic_norm The Metacritic critic score for the film, normalized to a 0 to 5 point system

metacritic_user_nom The Metacritic user score for the film, normalized to a 0 to 5 point system

imdb_norm The IMDb user score for the film, normalized to a 0 to 5 point system

rt_norm_round The Rotten Tomatoes Tomatometer score for the film , normalized to a 0 to 5 point system and rounded to the nearest half-star

rt_user_norm_round The Rotten Tomatoes user score for the film , normalized to a 0 to 5 point system and rounded to the nearest half-star

metacritic_norm_round The Metacritic critic score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star

metacritic_user_norm_round The Metacritic user score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star

imdb_norm_round The IMDb user score for the film, normalized to a 0 to 5 point system and rounded to the nearest half-star

metacritic_user_vote_count The number of user votes the film had on Metacritic

imdb_user_vote_count The number of user votes the film had on IMDb

fandango_votes The number of user votes the film had on Fandango

fandango_difference The difference between the presented Fandango_Stars and the actual Fandango_Ratingvalue

Source

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

`fifa_audience` *How To Break FIFA*

Description

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

Usage

```
fifa_audience
```

Format

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

country FIFA member country

confederation Confederation to which country belongs

population_share Country's share of global population (percentage)

tv_audience_share Country's share of global world cup TV Audience (percentage)

gdp_weighted_share Country's GDP-weighted audience share (percentage)

Source

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

Examples

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

fivethirtyeight *fivethirtyeight: Data and Code Behind the Stories and Interactives at 'FiveThirtyEight'*

Description

An R library that provides access to the code and data sets published by FiveThirtyEight https://github.com/fivethirtyeight/data. Note that while we received guidance from editors at 538, this package is not officially published by 538. Contribute to this package at https://github.com/rudeboybert/fivethirtyeight.

Examples

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

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

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

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

# Some data sets include vignettes with an example analysis:
vignette("bechdel", package = "fivethirtyeight")

# To browse all vignettes:
browseVignettes(package = "fivethirtyeight")
```

flying *41 Percent Of Fliers Think You're Rude If You Recline Your Seat*

Description

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

Usage

```
flying
```

Format

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

respondent_id RespondentID

gender Gender

age Age

height Height

children_under_18 Do you have any children under 18?

household_income Household income bracket

education Education Level

location Location (census region)

frequency How often do you travel by plane?

recline_frequency Do you ever recline your seat when you fly?

recline_obligation Under normal circumstances, does a person who reclines their seat during a flight have any obligation to the person sitting behind them?

recline_rude Is it rude to recline your seat on a plane?

recline_eliminate Given the opportunity, would you eliminate the possibility of reclining seats on planes entirely?

switch_seats_friends Is it rude to ask someone to switch seats with you in order to be closer to friends?

switch_seats_family Is it rude to ask someone to switch seats with you in order to be closer to family?

wake_up_bathroom Is it rude to wake a passenger up if you are trying to go to the bathroom?

wake_up_walk Is it rude to wake a passenger up if you are trying to walk around?

baby In general, is it rude to bring a baby on a plane?

unruly_child In general, is it rude to knowingly bring unruly children on a plane?

two_arm_rests In a row of three seats, who should get to use the two arm rests?

middle_arm_rest In a row of two seats, who should get to use the middle arm rest?

shade Who should have control over the window shade?

unsold_seat Is it rude to move to an unsold seat on a plane?

talk_stranger Generally speaking, is it rude to say more than a few words to the stranger sitting next to you on a plane?

get_up On a 6 hour flight from NYC to LA, how many times is it acceptable to get up if you're not in an aisle seat?

electronics Have you ever used personal electronics during take off or landing in violation of a flight attendant's direction?

smoked Have you ever smoked a cigarette in an airplane bathroom when it was against the rules?

Source

SurveyMonkey survey

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

Description

The raw data behind the story "The FiveThirtyEight International Food Association's 2014 World Cup" https://fivethirtyeight.com/features/the-fivethirtyeight-international-food-associations-20
For all the countries below, the response to the following question is presented: "Please rate how much you like the traditional cuisine of X"

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

Usage

```
food_world_cup
```

Format

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

respondent_id Respondent ID

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

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

gender Gender

age Age

household_income Household income bracket

education Education Level

location Location (census region)

algeria Cuisine of Algeria

argentina Cuisine of Argentina

australia Cuisine of Australia

belgium Cuisine of Belgium

bosnia_and_herzegovina Cuisine of Bosnia & Herzegovina

brazil Cuisine of Brazil

cameroon Cuisine of Cameroon

chile Cuisine of Chile

china Cuisine of China

colombia Cuisine of Colombia

costa_rica Cuisine of Costa Rica

croatia Cuisine of Croatia

cuba Cuisine of Cuba

ecuador Cuisine of Ecuador

england Cuisine of England

ethiopia Cuisine of Ethiopia

france Cuisine of France

germany Cuisine of Germany

ghana Cuisine of Ghana

greece Cuisine of Greece

honduras Cuisine of Honduras

india Cuisine of India

iran Cuisine of Iran

ireland Cuisine of Ireland

italy Cuisine of Italy

ivory_coast Cuisine of Ivory Coast

japan Cuisine of Japan

mexico Cuisine of Mexico

nigeria Cuisine of Nigeria

portugal Cuisine of Portugal

russia Cuisine of Russia

south_korea Cuisine of South Korea

spain Cuisine of Spain

switzerland Cuisine of Switzerland

thailand Cuisine of Thailand

the_netherlands Cuisine of the Netherlands

turkey Cuisine of Turkey

united_states Cuisine of the United States

uruguay Cuisine of Uruguay

vietnam Cuisine of Vietnam

See Also

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

`generic_polllist` *Congress Generic Ballot Polls*

Description

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

Usage

```
generic_polllist
```

Format

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

subgroup No description provided.

modeldate No description provided.

startdate Start date of the poll.

enddate End date of the poll.

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

grade No description provided.

samplesize No description provided.

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

weight No description provided.

influence No description provided.

dem No description provided.

rep No description provided.

adjusted_dem No description provided.

adjusted_rep No description provided.

multiversions No description provided.

tracking No description provided.

url No description provided.

poll_id No description provided.

question_id No description provided.

createddate No description provided.

timestamp No description provided.

Source

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

See Also

generic_topline

generic_topline *Congress Generic Ballot Polls*

Description

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

Usage

```
generic_topline
```

Format

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

subgroup No description provided.

modeldate No description provided.

dem_estimate No description provided.

dem_hi No description provided.

dem_lo No description provided.

rep_estimate No description provided.

rep_hi No description provided.

rep_lo No description provided.

timestamp No description provided.

Source

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

See Also

generic_polllist

google_trends *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-pue
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

---

goose *The Save Ruined Relief Pitching. The Goose Egg Can Fix It.*

Description

The raw data behind the story "The Save Ruined Relief Pitching. The Goose Egg Can Fix It." https://fivethirtyeight.com/features/goose-egg-new-save-stat-relief-pitchers/.

Usage

```
goose
```

Format

A data frame with 30,533 rows representing pitchers and 12 variables:

name Pitcher name
year Start year of season
team Retrosheet team code
league NL or AL
goose_eggs Goose eggs
broken_eggs Broken eggs
mehs Mehs
league_average_gpct League-average goose percentage
ppf Pitcher park factor
replacement_gpct Replacement-level goose percentage
gwar Goose Wins Above Replacement
key_retro Retrosheet unique player identifier

Source

Retrosheet http://www.retrosheet.org/

`hate_crimes` *Higher Rates Of Hate Crimes Are Tied To Income Inequality*

Description

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

Usage

```
hate_crimes
```

Format

A data frame with 51 rows representing US states and DC and 12 variables:

state State name

median_house_inc Median household income, 2016

share_unemp_seas Share of the population that is unemployed (seasonally adjusted), Sept. 2016

share_pop_metro Share of the population that lives in metropolitan areas, 2015

share_pop_hs Share of adults 25 and older with a high-school degree, 2009

share_non_citizen Share of the population that are not U.S. citizens, 2015

share_white_poverty Share of white residents who are living in poverty, 2015

gini_index Gini Index, 2015

share_non_white Share of the population that is not white, 2015

share_vote_trump Share of 2016 U.S. presidential voters who voted for Donald Trump

hate_crimes_per_100k_splc Hate crimes per 100,000 population, Southern Poverty Law Center, Nov. 9-18, 2016

avg_hatecrimes_per_100k_fbi Average annual hate crimes per 100,000 population, FBI, 2010-2015

Source

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

hiphop_cand_lyrics *Hip-Hop Is Turning On Donald Trump*

Description

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

Usage

```
hiphop_cand_lyrics
```

Format

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

candidate Candidate referenced

song Song name

artist Artist name

sentiment Positive, negative or neutral

theme Theme of lyric

album_release_date Date of album release

line Lyrics

url Genius link

Source

Genius http://genius.com/

hist_ncaa_bball_casts *The NCAA Bracket: Checking Our Work*

Description

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

Usage

```
hist_ncaa_bball_casts
```

Format

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

year

round

favorite

underdog

favorite_prob

favorite_win

Source

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

`hist_senate_preds` *How The FiveThirtyEight Senate Forecast Model Works*

Description

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

Usage

```
hist_senate_preds
```

Format

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

state Election

year Year of election

candidate Last name

forecast_prob Probability of winning election per FiveThirtyEight Election Day forecast

result 'Win' or 'Loss'

Source

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

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 http://www.bls.gov/oes/current/oes254021.htm#(1)

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

Description

The raw data behind the story "The Definitive Analysis Of 'Love Actually,' The Greatest Christmas Movie Of Our Time" https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have
The adjacency matrix of which actors appear in the same scene together.

Usage

```
love_actually_adj
```

Format

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

actors

bill_nighy

keira_knightley

andrew_lincoln

hugh_grant

colin_firth

alan_rickman

heike_makatsch

laura_linney

emma_thompson

liam_neeson

kris_marshall

abdul_salis

martin_freeman

rowan_atkinson

See Also

love_actually_appearance.

`love_actually_appearance`

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

---

Description

The raw data behind the story "The Definitive Analysis Of 'Love Actually,' The Greatest Christmas Movie Of Our Time" https://fivethirtyeight.com/features/the-definitive-analysis-of-love-actually-
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(tidyverse)
library(stringr)
love_actually_appearance_tidy <- love_actually_appearance %>%
  gather(actor, appears, -c(scenes)) %>%
  arrange(scenes)
```

---

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

Description

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

Usage

```
mad_men
```

Format

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

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

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

show_start The year the television show began

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

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

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

years_since 2015 minus CharEnd

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

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

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

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

score_div_y "Score" divided by "Years Since"

lead_notes The list of films counted in #LEAD

support_notes The list of films counted in #SUPPORT

show_notes The seasons of shows counted in #Shows

Source

IMDB http://imdb.com

`male_flight_attend` *Dear Mona, How Many Flight Attendants Are Men?*

Description

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

Usage

```
male_flight_attend
```

Format

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

job_category Category of job

percentage_male Percentage of workforce that are male

Source

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

mayweather_mcgregor_tweets

Mayweather Vs McGregor Tweets

Description

The raw data behind the story "The Mayweather-McGregor Fight As Told Through Emojis" https://fivethirtyeight.com/?post_type=fte_features&p=161615.

Usage

```
mayweather_mcgregor_tweets
```

Format

Because of R package size restrictions, only a preview of the first 10 rows of this dataset is included; to obtain the entire dataset (12118 rows) see Examples below. A data frame with 10 rows representing tweets and 7 variables:

created_at Time and date at which the tweet associated with the Mayweather vs. McGregor fight was sent.

emojis Whether or not emojis were used in the tweet about the fight.

id A numerical identifier for each individual tweet about the fight.

link The link to the tweet about the fight on Twitter.

retweeted Whether or not the tweet about the fight was retweeted.

screen_name The screen name under which the tweet about the fight was posted.

text The text contained in the tweet about the fight.

Source

This data contains 12,118 tweets that contain one or more emojis and match one or more of the following hashtags: #MayMac, #MayweatherMcGregor, #MayweatherVMcGregor, #MayweatherVsMcGregor, #McGregor and #Mayweather. Data was collected on August 27, 2017 between 12:05 a.m. and 1:15 a.m. EDT using the Twitter streaming API. https://github.com/fivethirtyeight/data/tree/master/mayweather-mcgregor

Examples

```
# To obtain the entire dataset, run the code inside the following if statement:
if(FALSE){
  library(tidyverse)
  url <-
   "https://raw.githubusercontent.com/fivethirtyeight/data/master/mayweather-mcgregor/tweets.csv"
  mayweather_mcgregor_tweets <- read_csv(url) %>%
    mutate(
      emojis = as.logical(emojis),
      retweeted = as.logical(retweeted),
      id = as.character(id)
    )
}
```

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

Description

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

Usage

```
mediacloud_hurricanes
```

Format

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

date Date

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

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

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

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

Source

Mediacloud https://mediacloud.org/

See Also

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

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

Description

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

Usage

```
mediacloud_online_news
```

Format

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

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

url URL of corresponding media outlet source

Source

Mediacloud https://mediacloud.org/

See Also

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

---

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

Description

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

Usage

```
mediacloud_states
```

Format

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

date Date

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

puerto_rico The number of sentences in online news which mention Puerto Rico

florida The number of sentences in online news which mention Florida

Source

Mediacloud https://mediacloud.org/

See Also

mediacloud_hurricanes, mediacloud_online_news, mediacloud_trump, tv_hurricanes, tv_hurricanes_by_net tv_states, google_trends

mediacloud_trump *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-pue Mediacloud Trump Data.

Usage

```
mediacloud_trump
```

Format

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

date Date

puerto_rico The number of headlines that mention Puerto Rico on the given date

puerto_rico_and_trump The number of headlines that mention Puerto Rico and either President or Trump

florida The number of headlines that mention Florida

florida_and_trump The number of headlines that mention Florida and either President or Trump

texas The number of headlines that mention Texas

texas_and_trump The number of headlines that mention Texas and either President or Trump

Source

Mediacloud https://mediacloud.org/

See Also

mediacloud_hurricanes, mediacloud_states, mediacloud_online_news,tv_hurricanes, tv_hurricanes_by_net tv_states, google_trends

mlb_as_play_talent *The Best MLB All-Star Teams Ever*

Description

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

Usage

```
mlb_as_play_talent
```

Format

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

bbref_id Player's ID at Baseball-Reference.com

yearid The season in question

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

gameid Game ID at Baseball-Reference.com

lgid League of All-Star team

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

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

def600 Estimate of fielding talent, in runs above league average per 600 plate appearances

pitch200 Estimate of pitching talent, in runs above league average per 200 innings pitched

asg_pa Number of plate appearances in the All-Star Game itself

asg_ip Number of innings pitched in the All-Star Game itself

offper9innasg Expected offensive runs added above average (from talent) based on PA in ASG, scaled to a 9-inning game

defper9innasg Expected defensive runs added above average (from talent) based on PA in ASG, scaled to a 9-inning game

pitper9innasg Expected pitching runs added above average (from talent) based on IP in ASG, scaled to a 9-inning game

totper9innasg Expected runs added above average (from talent) based on PA/IP in ASG, scaled to a 9-inning game

Source

http://baseball-reference.com , http://chadwick-bureau.com, Fangraphs

---

mlb_as_team_talent *The Best MLB All-Star Teams Ever*

Description

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

Usage

```
mlb_as_team_talent
```

Format

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

yearid The season in question

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

gameid Game ID at Baseball-Reference.com

lgid League of All-Star team

tm_off_talent Total runs of offensive talent above average per game (36 plate appearances)

tm_def_talent Total runs of fielding talent above average per game (36 plate appearances)

tm_pit_talent Total runs of pitching talent above average per game (9 innings)

mlb_avg_rpg MLB average runs scored/game that season

talent_rspg Expected runs scored per game based on talent (MLB R/G + team OFF talent)

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

unadj_pyth Unadjusted pythagorean talent rating; PYTH =(RSPG^1.83)/(RSPG^1.83+RAPG^1.83)

timeline_adj Estimate of relative league quality where 2015 MLB = 1.00

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

adj_pyth Adjusted pythagorean record; =(SOS*unadj_Pyth)/((2*unadj_Pyth*SOS)-SOS-unadj_Pyth+1)

no_1_player Best player according to combo of actual PA/IP and talent

no_2_player 2nd-best player according to combo of actual PA/IP and talent

Source

http://baseball-reference.com , http://chadwick-bureau.com, Fangraphs

---

mlb_elo *MLB Elo*

Description

The raw data behind the stories: "The Complete History Of MLB" https://projects.fivethirtyeight.com/complete-history-of-mlb/ and "MLB Predictions" https://projects.fivethirtyeight.com/2017-mlb-predictions/.

Usage

```
mlb_elo
```

Format

Because of R package size restrictions, only a preview of the first 10 rows of this dataset is included; to obtain the entire dataset (1871 to 2017 games) see Examples below. A data frame with 10 rows representing Elo ratings and 26 variables:

date The date of the game.

season The season within which the game was played.

neutral No description provided.

playoff No description provided.

team1 One team that participated in the game.

team2 The other team that participated in the match.

elo1_pre The Elo rating for team1 prior to the game.

elo2_pre The Elo rating for team2 prior to the game.

elo_prob1 No description provided.

elo_prob2 No description provided.

elo1_post The Elo rating for team1 after the game.

elo2_post The Elo rating for team2 after the game.

rating1_pre No description provided.

rating2_pre No description provided.

pitcher1 An identifier of the pitcher

pitcher2 No description provided.

pitcher1_rating No description provided.

pitcher2_rating No description provided.

pitcher1_adj No description provided.

pitcher2_adj No description provided.

rating_prob1 No description provided.

rating_prob2 No description provided.

rating1_post No description provided.

rating2_post No description provided.

score1 The number of runs scored by team1.

score2 The number of runs scored by team2.

Source

See https://github.com/fivethirtyeight/data/blob/master/mlb-elo/README.md

Examples

```
# To obtain the entire dataset, run the code inside the following if statement:
if(FALSE){
  library(tidyverse)
  mlb_elo <- read_csv("https://projects.fivethirtyeight.com/mlb-api/mlb_elo.csv") %>%
    mutate(
      playoff = as.factor(playoff),
      playoff = ifelse(playoff == "<NA>", NA, playoff),
      neutral = as.logical(neutral)
    )
}
```

murder_2015_final *A Handful Of Cities Are Driving 2016's Rise In Murder*

Description

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

Usage

```
murder_2015_final
```

Format

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

city Name of city

state Name of state

murders_2014 Total murders in 2014

murders_2015 Total murders in 2015

change 2015 - 2014

Source

Unknown

murder_2016_prelim *A Handful Of Cities Are Driving 2016's Rise In Murder*

Description

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

Usage

```
murder_2016_prelim
```

Format

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

city Name of city

state Name of state

murders_2015 Number of murders in 2015

murders_2016 Number of murder in 2016 (as of `as_of` date)

change 2016 - 2015

source Source of data

as_of 2016 murders up to this date

Source

Listed as source variable in dataset

---

nba_carmelo *The Complete History Of The NBA 2017-18 NBA Predictions*

Description

The raw data behind the story "The Complete History Of The NBA" https://projects.fivethirtyeight.com/complete-history-of-the-nba/ and our "2017-18 NBA Predictions" https://projects.fivethirtyeight.com/2018-nba-predictions/

Usage

```
nba_carmelo
```

Format

Because of R package size restrictions, only a preview of the first 10 rows of this dataset is included; to obtain the entire dataset (1871 to 2017 games) see Examples below. A data frame with 10 rows representing games and 20 variables:

date Date

season Season year, 1947-2018

neutral TRUE if the game was played on neutral territory, FALSE if not

playoff TRUE if the game was a playoff game, FALSE if not

team1 The name of one participating team

team2 The name of the other participating team

elo1_pre Team 1's Elo rating before the game

elo2_pre Team 2's Elo rating before the 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's Elo rating after the game

elo2_post Team 2's Elo rating after the game

carmelo1_pre Team 1's CARMELO rating before the game

carmelo2_pre Team 2's CARMELO rating before the game

carmelo1_post Team 1's CARMELO rating after the game

carmelo2_post Team 2's CARMELO rating after the game

carmelo_prob1 Team 1's probability winning based on CARMELO rating

carmelo_prob2 Team 2's probability of winning based on CARMELO rating

score1 Points scored by Team 1

score2 Points scored by Team 2

Source

See https://projects.fivethirtyeight.com/nba-model/nba_elo.csv

Examples

```
# To obtain the entire dataset, run the following code:
library(tidyverse)
library(janitor)
nba_carmelo <- read_csv("https://projects.fivethirtyeight.com/nba-model/nba_elo.csv") %>%
  clean_names() %>%
  mutate(
    team1 = as.factor(team1),
    team2 = as.factor(team2),
    playoff = ifelse(playoff == "t", TRUE, FALSE),
    playoff = ifelse(is.na(playoff), FALSE, TRUE),
    neutral = ifelse(neutral == 1, TRUE, FALSE)
  )
```

`nba_draft_2015` *Projecting The Top 50 Players In The 2015 NBA Draft Class*

Description

The raw data behind the story "Projecting The Top 50 Players In The 2015 NBA Draft Class" https://fivethirtyeight.com/features/projecting-the-top-50-players-in-the-2015-nba-draft-class/ An analysis using this data was contributed by G. Elliott Morris as a package vignette at http://fivethirtyeight-r.netlify.com/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)

Source

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

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

Description

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

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 http://nbatattoos.tumblr.com/

nfltix_div_avgprice *Who Goes To Meaningless NFL Games And Why?*

Description

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

Usage

```
nfltix_div_avgprice
```

Format

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

event NFL divisional game info

division NFL division

avg_tix_price Average ticket price

Source

StubHub

nfltix_usa_avg *Who Goes To Meaningless NFL Games And Why?*

Description

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

Usage

```
nfltix_usa_avg
```

Format

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

team Name of NFL team

avg_tix_price Average ticket price

Source

StubHub

nflwr_aging_curve *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

```
nflwr_aging_curve
```

Format

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

age_from Beginning age

age_to Ending age

trypg_change Change in TRY per game from one age-year to next

Source

Unknown

nflwr_hist *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

```
nflwr_hist
```

Format

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

pfr_player_id Player identification code at Pro-Football-Reference.com

player_name The player's name

career_try Career True Receiving Yards

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

career_wowy The amount by which career_ranypa exceeds what would be expected from his QBs' (age-adjusted) performance without the receiver

bcs_rating The number of yards per game by which a player would outgain an average receiver on the same team, after adjusting for teammate quality and age (update of http://www.sabernomics.com/sabernomics/index.php/2005/02/ranking-the-all-time-great-wide-receivers/)

Source

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

nfl_elo *The Complete History Of The NFL 2017 NFL Predictions*

Description

The raw data behind the story "The Complete History of the NFL" https://projects.fivethirtyeight.com/complete-history-of-the-nfl/ And our "2017 NFL Predictions" https://projects.fivethirtyeight.com/2017-nfl-predictions/

Usage

```
nfl_elo
```

Format

Because of R package size restrictions, only a preview of the first 10 rows of this dataset is included; to obtain the entire dataset (1920 to 2018 games) see Examples below. A data frame with 10 rows representing games and 14 variables:

date Date

season Season year, 1920-2018

neutral TRUE if the game was played on neutral territory, FALSE if not

playoff No description provided

team1 The name of one participating team

team2 The name of the other participating team

elo1_pre Team 1's Elo rating before the game

elo2_pre Team 2's Elo rating before the 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's Elo rating after the game

elo2_post Team 2's Elo rating after the game

score1 Points scored by Team 1

score2 Points scored by Team 2

Source

See https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv # To obtain the entire dataset, run the following code: library(tidyverse) library(janitor) nfl_elo <- read_csv("https://projects.fivethirtyeight.com api/nfl_elo.csv") clean_names() mutate(team1 = as.factor(team1), team2 = as.factor(team2), neutral = ifelse(neutral == 1, TRUE, FALSE))

---

nfl_fandom_google *How Every NFL Team's Fans Lean Politically*

Description

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

Usage

```
nfl_fandom_google
```

Format

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

dma Designated Market Area

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

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

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

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

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

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

trump_2016_vote The percentage of voters in the region who voted for Trump in the 2016 Presidential Election

Source

Google Trends https://g.co/trends/5P8aa

See Also

nfl_fandom_surveymonkey

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
nfl_fandom_google_tidy <- nfl_fandom_google %>%
  gather(sport, search_traffic, -c("dma", "trump_2016_vote")) %>%
  arrange(dma)
```

---

nfl_fandom_surveymonkey

How Every NFL Team's Fans Lean Politically

Description

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

Usage

```
nfl_fandom_surveymonkey
```

Format

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

team NFL team

total_respondents Total number of poll respondents who ranked the given team in their top 3 favorites

asian_dem Number of Asian, democrat poll respondents who ranked the given team in their top 3 favorites

black_dem Number of Black, democrat poll respondents who ranked the given team in their top 3 favorites

hispanic_dem Number of Hispanic, democrat poll respondents who ranked the given team in their top 3 favorites

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

white_dem Number of White, democrat poll respondents who ranked the given team in their top 3 favorites

total_dem Total number of democrat poll respondents who ranked the given team in their top 3 favorites

asian_ind Number of Asian, independent poll respondents who ranked the given team in their top 3 favorites

black_ind Number of Black, independent poll respondents who ranked the given team in their top 3 favorites

hispanic_ind Number of Hispanic, independent poll respondents who ranked the given team in their top 3 favorites

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

white_ind Number of White, independent poll respondents who ranked the given team in their top 3 favorites

total_ind Total number of independent poll respondents who ranked the given team in their top 3 favorites

asian_gop Number of Asian, republican poll respondents who ranked the given team in their top 3 favorites

black_gop Number of Black, republican poll respondents who ranked the given team in their top 3 favorites

hispanic_gop Number of Hispanic, republican poll respondents who ranked the given team in their top 3 favorites

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

white_gop Number of White, republican poll respondents who ranked the given team in their top 3 favorites

total_gop Total number of republican poll respondents who ranked the given team in their top 3 favorites

gop_percent Percent of fans (who ranked the team in their top 3 favorite NFL teams) who are republicans

dem_percent Percent of fans who are democrats

ind_percent Percent of fans who are independent

white_percent Percent of fans who are White

nonwhite_percent Percent of fans who are not White

Source

See https://github.com/fivethirtyeight/data/tree/master/nfl-fandom/NFL_fandom_data-surveymonkey.csv

See Also

nfl_fandom_google

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
nfl_fandom_surveymonkey_tidy <- nfl_fandom_surveymonkey %>%
  gather(key = race_party, value = percent,
         -c("team", "total_respondents", "gop_percent", "dem_percent",
            "ind_percent", "white_percent", "nonwhite_percent")) %>%
  arrange(team)
```

`nfl_fav_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

Usage

```
nfl_fav_team
```

Format

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

team Name of NFL team

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

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

players Players - Effort on the field, likability off it

future_wins Future wins - Projected wins over next 5 seasons

bandwagon Bandwagon Factor - Are the team's next 5 years likely to be better than their previous 5?

tradition Tradition - Championships/division titles/wins in team's entire history

bang_buck Bang for the buck - Wins per fan dollars spent

behavior Behavior - Suspensions by players on team since 2007, with extra weight to transgressions vs. women

nyc_prox Proximity to New York City

stlouis_prox Proximity to St. Louis

afford Affordability - Price of tickets, parking and concessions

small_market Small Market - Size of market in terms of population, where smaller is better

stadium_exp Stadium experience - Quality of venue; fan-friendliness of environment; frequency of game-day promotions

coaching Coaching - Strength of on-field leadership

uniform Uniform - Stylishness of uniform design, according to Uni Watch's Paul Lukas

big_market Big Market - Size of market in terms of population, where bigger is better

Source

http://www.espn.com/sportsnation/teamrankings, http://www.allourideas.org/nflteampickingsample

---

`nfl_suspensions` *The NFL's Uneven History Of Punishing Domestic Violence*

Description

The raw data behind the story "The NFL's Uneven History Of Punishing Domestic Violence" https://fivethirtyeight.com/features/nfl-domestic-violence-policy-suspensions/.

Usage

```
nfl_suspensions
```

Format

A data frame with 269 rows representing National Football League players and 7 variables:

name first initial.last name

team team at time of suspension

games number of games suspended (one regular season = 16 games)

category personal conduct, substance abuse, performance enhancing drugs or in-game violence

description description of suspension

year year of suspension

source news source

Source

http://en.wikipedia.org/wiki/List_of_players_and_coaches_suspended_by_the_NFL, http://www.spotrac.com/fines-tracker/nfl/suspensions/

`nutrition_pvalues` *You Can't Trust What You Read About Nutrition*

Description

The raw data behind the story "You Can't Trust What You Read About Nutrition" https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition.

Usage

```
nutrition_pvalues
```

Format

A data frame with 27716 rows representing Regression fits for p-hacking and 3 variables:

food Name of food (response/dependent variable)

characteristic Name of characteristic (predictor/independent variable)

p_values P-value from regression fit

Source

See https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition

---

police_deaths *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-pol

Usage

```
police_deaths
```

Format

A data frame with 22800 rows representing Police officers/dogs who lost their lives and 7 variables:

person Name of person/canine who died

cause_of_death Cause of death

date Date of event

year Year of event

canine TRUE if canine, FALSE if human

dept_name Name of police department

state State of police department

Source

Officer Down Memorial Page https://www.odmp.org/

police_killings *Where Police Have Killed Americans In 2015*

Description

The raw data behind the story "Where Police Have Killed Americans In 2015" https://fivethirtyeight.com/features/where-police-have-killed-americans-in-2015.

Usage

```
police_killings
```

Format

A data frame with 467 rows representing People who died from interactions with police and 34 variables:

name Name of deceased

age Age of deceased

gender Gender of deceased

raceethnicity Race/ethnicity of deceased

month Month of killing

day Day of incident

year Year of incident

streetaddress Address/intersection where incident occurred

city City where incident occurred

state State where incident occurred

latitude Latitude, geocoded from address

longitude Longitude, geocoded from address

state_fp State FIPS code

county_fp County FIPS code

tract_ce Tract ID code

geo_id Combined tract ID code

county_id Combined county ID code

namelsad Tract description

lawenforcementagency Agency involved in incident

cause Cause of death

armed How/whether deceased was armed

pop Tract population

share_white Share of pop that is non-Hispanic white

share_black Share of pop that is black (alone, not in combination)

share_hispanic Share of pop that is Hispanic/Latino (any race)

p_income Tract-level median personal income

h_income Tract-level median household income

county_income County-level median household income

comp_income ‘h_income‘ / ‘county_income‘

county_bucket Household income, quintile within county

nat_bucket Household income, quintile nationally

pov Tract-level poverty rate (official)

urate Tract-level unemployment rate

college Share of 25+ pop with BA or higher

Source

See https://github.com/fivethirtyeight/data/tree/master/police-killings

---

`police_locals` *Most Police Don’t Live In The Cities They Serve*

Description

The raw data behind the story "Most Police Don’t Live In The Cities They Serve" https://fivethirtyeight.com/features/most-police-dont-live-in-the-cities-they-serve/.

Usage

```
police_locals
```

Format

A data frame with 75 rows representing cities and 8 variables:

city U.S. city

force_size Number of police officers serving that city

all Percentage of the total police force that lives in the city

white Percentage of white (non-Hispanic) police officers who live in the city

non_white Percentage of non-white police officers who live in the city

black Percentage of black police officers who live in the city

hispanic Percentage of Hispanic police officers who live in the city

asian Percentage of Asian police officers who live in the city

Details

The dataset includes the cities with the 75 largest police forces, with the exception of Honolulu for which data is not available. All calculations are based on data from the U.S. Census.

The Census Bureau numbers are potentially going to differ from other counts for three reasons:

1. The census category for police officers also includes sheriffs, transit police and others who might not be under the same jurisdiction as a city’s police department proper. The census category won’t include private security officers.

2. The census data is estimated from 2006 to 2010; police forces may have changed in size since then.
3. There is always a margin of error in census numbers; they are estimates, not complete counts.

Note: Missing values means that there are fewer than 100 police officers of that race serving that city.

Source

See https://github.com/fivethirtyeight/data/tree/master/police-locals

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
police_locals_tidy <- police_locals %>%
   gather(key = "race", value = "perc_in", all:asian)
```

---

pres_2016_trail *The Last 10 Weeks Of 2016 Campaign Stops In One Handy Gif*

Description

The raw data behind the story "The Last 10 Weeks Of 2016 Campaign Stops In One Handy Gif" https://fivethirtyeight.com/features/the-last-10-weeks-of-2016-campaign-stops-in-one-handy-gif/

Usage

```
pres_2016_trail
```

Format

A data frame with 177 rows representing 2016 Republican and Democratic candidate campaign trail stops and 5 variables:

candidate Clinton or Trump

date The date of the event

location The location of the event

lat Latitude of the event location

lng Longitude of the event location

Source

https://hillaryspeeches.com/, http://www.conservativedailynews.com/

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-

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 http://www.presidency.ucsb.edu

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

See https://fivethirtyeight.com/features/do-pulitzers-help-newspapers-keep-readers/

---

ratings *An Inconvenient Sequel*

Description

The raw data behind the story "Al Gore's New Movie Exposes The Big Flaw In Online Movie Ratings" https://fivethirtyeight.com/features/al-gores-new-movie-exposes-the-big-flaw-in-online-mov

Usage

```
ratings
```

Format

A data frame with 80053 rows representing movie ratings and 27 variables:

timestamp The date at which the rating was recorded.

respondents The number of respondents in a category associated with a given timestamp.

category The subgroups of respondents differentiated by demographics like gender, age, and nationality.

link The website associated with a given category's responses.

average The average rating reported by a given category.

mean The mean rating reported by a given category.

median The median rating reported by a given category.

votes_1 The count of votes denoting a rating of one that respondents gave.

votes_2 The count of votes denoting a rating of two that respondents gave.

votes_3 The count of votes denoting a rating of three that respondents gave.

votes_4 The count of votes denoting a rating of four that respondents gave.

votes_5 The count of votes denoting a rating of five that respondents gave.

votes_6 The count of votes denoting a rating of six that respondents gave.

votes_7 The count of votes denoting a rating of seven that respondents gave.

votes_8 The count of votes denoting a rating of eight that respondents gave.

votes_9 The count of votes denoting a rating of nine that respondents gave.

votes_10 The count of votes denoting a rating of ten that respondents gave.

pct_1 The percentage of votes denoting a rating of one that respondents gave.

pct_2 The percentage of votes denoting a rating of two that respondents gave.

pct_3 The percentage of votes denoting a rating of three that respondents gave.

pct_4 The percentage of votes denoting a rating of four that respondents gave.

pct_5 The percentage of votes denoting a rating of five that respondents gave.

pct_6 The percentage of votes denoting a rating of six that respondents gave.

pct_7 The percentage of votes denoting a rating of seven that respondents gave.

pct_8 The percentage of votes denoting a rating of eight that respondents gave.

pct_9 The percentage of votes denoting a rating of nine that respondents gave.

pct_10 The percentage of votes denoting a rating of ten that respondents gave.

Source

IMBD http://www.imdb.com/title/tt6322922/ratings and see https://github.com/fivethirtyeight/data/tree/master/inconvenient-sequel

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
library(stringr)
ratings_tidy <- ratings %>%
  gather(votes, count, -c(timestamp, respondents, category, link, average, mean, median)) %>%
  arrange(timestamp)
```

---

riddler_castles *Can You Rule Riddler Nation?*

Description

The raw data behind the story "Can You Rule Riddler Nation?" https://fivethirtyeight.com/features/can-you-rule-riddler-nation/. Analysis of the submitted solutions can be found at: https://fivethirtyeight.com/features/can-you-save-the-drowning-swimmer/

Usage

```
riddler_castles
```

Format

A data frame with 1387 rows representing submissions and 11 variables:

castle1 Number of troops out of 100 send to castle 1

castle2 Number of troops out of 100 send to castle 2

castle3 Number of troops out of 100 send to castle 3

castle4 Number of troops out of 100 send to castle 4

castle5 Number of troops out of 100 send to castle 5

castle6 Number of troops out of 100 send to castle 6

castle7 Number of troops out of 100 send to castle 7

castle8 Number of troops out of 100 send to castle 8

castle9 Number of troops out of 100 send to castle 9

castle10 Number of troops out of 100 send to castle 10

reason Why did you choose your troop deployment?

Source

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

See Also

riddler_castles2

Examples

```
# To convert data frame to tidy data (long) format, run
library(tidyverse)
library(stringr)
riddler_castles_tidy<-riddler_castles %>%
   gather(key = castle , value = soldiers, castle1:castle10) %>%
   mutate(castle = as.numeric(str_replace(castle, "castle","")))
```

---

riddler_castles2 *The Battle For Riddler Nation, Round 2*

Description

The raw data behind the story "The Battle For Riddler Nation, Round 2" https://fivethirtyeight.com/features/the-battle-for-riddler-nation-round-2/. Analysis of the submitted solutions can be found at: https://fivethirtyeight.com/features/how-much-should-you-bid-for-that-painting

Usage

```
riddler_castles2
```

Format

A data frame with 932 rows representing submissions and 11 variables:

castle1 Number of troops out of 100 send to castle 1

castle2 Number of troops out of 100 send to castle 2

castle3 Number of troops out of 100 send to castle 3

castle4 Number of troops out of 100 send to castle 4

castle5 Number of troops out of 100 send to castle 5

castle6 Number of troops out of 100 send to castle 6

castle7 Number of troops out of 100 send to castle 7

castle8 Number of troops out of 100 send to castle 8

castle9 Number of troops out of 100 send to castle 9

castle10 Number of troops out of 100 send to castle 10

reason Why did you choose your troop deployment?

Source

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

See Also

riddler_castles

Examples

```
# To convert data frame to tidy data (long) format, run
library(tidyverse)
library(stringr)
riddler_castles_tidy<-riddler_castles2 %>%
   gather(key = castle , value = soldiers, castle1:castle10) %>%
   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

sandy_311 *The (Very) Long Tail Of Hurricane Recovery*

Description

The raw data behind the story "The (Very) Long Tail Of Hurricane Recovery" https://projects.fivethirtyeight.com/sandy-311/

Usage

```
sandy_311
```

Format

A data frame with 1783 rows representing dates and 25 variables:

date Date

nyc_311 No description provided.

acs The number of emergency hotline (311) calls made to the Administration for Children's Services related to Hurricane Sandy on the given date

bpsi The number of emergency hotline (311) calls made to Building Protection Systems, Inc related to Hurricane Sandy

cau The number of emergency hotline (311) calls made to the Community Affairs Unit related to Hurricane Sandy

chall The number of emergency hotline (311) calls made to the City Hall related to Hurricane Sandy

dep The number of emergency hotline (311) calls made to the Department of Environmental Protection related to Hurricane Sandy

dob The number of emergency hotline (311) calls made to the Department of Buildings related to Hurricane Sandy

doe The number of emergency hotline (311) calls made to the Department of Education related to Hurricane Sandy

dof The number of emergency hotline (311) calls made to the Department of Finance related to Hurricane Sandy

dohmh The number of emergency hotline (311) calls made to the Department of Health and Mental Hygiene related to Hurricane Sandy

dpr The number of emergency hotline (311) calls made to the Department of Parks and Recreation related to Hurricane Sandy

fema The number of emergency hotline (311) calls made to the Federal Emergency Management Agency related to Hurricane Sandy

hpd The number of emergency hotline (311) calls made to the Department of Housing Preservation and Development related to Hurricane Sandy

hra The number of emergency hotline (311) calls made to the Human Resources Administration related to Hurricane Sandy

mfanyc The number of emergency hotline (311) calls made to the Mayor's Fund to Advance NYC related to Hurricane Sandy

mose The number of emergency hotline (311) calls made to the Mayor's Office of Special Enforcement related to Hurricane Sandy

nycem The number of emergency hotline (311) calls made to Emergency Management related to Hurricane Sandy

nycha The number of emergency hotline (311) calls made to the New York City Housing Authority related to Hurricane Sandy

nyc_service The number of emergency hotline (311) calls made to NYC Service related to Hurricane Sandy

nypd The number of emergency hotline (311) calls made to the New York Police Department related to Hurricane Sandy

nysdol The number of emergency hotline (311) calls made to the NYC Department of Labor related to Hurricane Sandy

sbs The number of emergency hotline (311) calls made to Small Business Services related to Hurricane Sandy

nys_emergency_mg The number of emergency hotline (311) calls made to NYS Emergency Management related to Hurricane Sandy

total The total number of emergency hotline (311) calls made related to Hurricane Sandy

Source

Data from NYC Open Data https://data.cityofnewyork.us/City-Government/311-Call-Center-Inquiry/tdd6-3ysr, Agency acronyms from the Data Dictionary. See also https://github.com/fivethirtyeight/data/tree/master/sandy-311-calls

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
sandy_311_tidy <- sandy_311 %>%
  gather(agency, num_calls, -c("date", "total")) %>%
  arrange(date) %>%
  select(date, agency, num_calls, total) %>%
  rename(total_calls = total) %>%
  mutate(agency = as.factor(agency))
```

---

san_andreas *The Rock Isn't Alone: Lots Of People Are Worried About 'The Big One'*

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-

Usage

```
san_andreas
```

Format

A data frame with 1013 rows representing respondents and 11 variables:

worry_general In general, how worried are you about earthquakes?

worry_bigone How worried are you about the "Big One," a massive, catastrophic earthquake?

will_occur Do you think the "Big One" will occur in your lifetime?

experience Have you ever experienced an earthquake?

prepared Have you or anyone in your household taken any precautions for an earthquake (packed an earthquake survival kit, prepared an evacuation plan, etc.)?

fam_san_andreas How familiar are you with the San Andreas Fault line?

fam_yellowstone How familiar are you with the Yellowstone Supervolcano?

age Age

female Gender

hhold_income How much total combined money did all members of your HOUSEHOLD earn last year?

region US Region

Source

See https://github.com/fivethirtyeight/data/tree/master/san-andreas

---

senate_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

`senators` *Senator Dataset*

Description

Senator Dataset

Usage

```
senators
```

Format

Because of R package size restrictions, only a preview of the first 10 rows of this dataset is included; to obtain the entire dataset see Examples below. A data frame with 10 rows representing tweets and 10 variables:

created_at The date and time the tweet was posted

user The user posting the tweet

text The text of the tweet

url The link to the tweet

replies The number of replies to the tweet

retweets The number of retweets

favorites The number of favorites

bioguide_id The poster's member ID from the "Biographical Directory of the United States Congress"

party The poster's political party affiliation

state The state the poster represents in Congress

Details

Data collected on Oct 19 and 20

Source

Twitter

See Also

`twitter_presidents`

Examples

```
# To obtain the entire dataset, run the code inside the following if statement:
if(FALSE){
  library(tidyverse)
  url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/twitter-ratio/senators.csv"
  senators <- read_csv(url) %>%
    mutate(
      party = as.factor(party),
      state = as.factor(state),
```

```
    created_at = as.POSIXct(created_at, tz = "GMT", format = "%m/%d/%Y %H:%M"),
    text =  gsub("[^\x01-\x7F]", "", text)
  ) %>%
  select(created_at, user, everything())
}
```

`spi_global_rankings` *Current SPI ratings and rankings for men's club teams*

Description

The raw data behind the stories "Club Soccer Predictions" https://projects.fivethirtyeight.com/soccer-predictions/ and "Global Club Soccer Rankings" https://projects.fivethirtyeight.com/global-club-soccer-rankings/.

Usage

```
spi_global_rankings
```

Format

A data frame with 453 rows representing soccer rankings and 7 variables:

name The name of the soccer club.

league The name of the league to which the club belongs.

rank A club's current global ranking.

prev_rank A club's previous global ranking

off Offensive rating for a given team (the higher the value the stronger the team's offense).

def Defensive rating for a given team (the lower the value the stronger the team's defense).

spi A club's SPI score.

Source

See https://github.com/fivethirtyeight/data/blob/master/soccer-spi/README.md

See Also

spi_matches

spi_matches *Match-by-match SPI ratings and forecasts back to 2016*

Description

The raw data behind the stories "Club Soccer Predictions" https://projects.fivethirtyeight.com/soccer-predictions/ and "Global Club Soccer Rankings" https://projects.fivethirtyeight.com/global-club-soccer-rankings/.

Usage

```
spi_matches
```

Format

A data frame with 10182 rows representing soccer matches and 13 variables:

date The date that the match took place.

league_id A numerical identifier of the league within which the match was played.

team1 One team that participated in the match.

team2 The other team that participated in the match.

spi1 The SPI score of team1.

spi2 The SPI score of team2.

prob1 The probability that team1 would have won the match.

prob2 The probability that team2 would have won the match.

probtie The probability that the match would have resulted in a tie.

proj_score1 The predicted number of goals that team1 would have scored.

proj_score2 The predicted number of goals that team2 would have scored.

score1 The number of goals that team1 scored.

score2 The number of goals that team2 scored.

xg1

xg2

nsxg1

nsxg2

adj_score1

adj_score2

Source

See https://github.com/fivethirtyeight/data/blob/master/soccer-spi/README.md

See Also

spi_global_rankings

steak_survey *How Americans Like Their Steak*

Description

The raw data behind the story "How Americans Like Their Steak" https://fivethirtyeight.com/features/how-americans-like-their-steak/.

Usage

```
steak_survey
```

Format

A data frame with 550 rows representing respondents and 15 variables:

respondent_id Respondent ID

lottery_a not sure

smoke Is respondent a smoker?

alcohol Is respondent a drinker?

gamble Is respondent a gambler?

skydiving Is respondent a skydiver?

speed not sure

cheated not sure

steak not sure

steak_prep Preferred steak preparation

female Is respondent female?

age Age

hhold_income Household income

educ Education level

region Region of US

Source

See https://fivethirtyeight.com/features/how-americans-like-their-steak/

tarantino *A Complete Catalog Of Every Time Someone Cursed Or Bled Out In A Quentin Tarantino Movie*

Description

The raw data behind the story "A Complete Catalog Of Every Time Someone Cursed Or Bled Out In A Quentin Tarantino Movie" https://fivethirtyeight.com/features/complete-catalog-curses-deaths-quen
An analysis using this data was contributed by Olivia Barrows, Jojo Miller, and Jayla Nakayama as a package vignette at http://fivethirtyeight-r.netlify.com/articles/tarantino_swears.html.

Usage

```
tarantino
```

Format

A data frame with 1894 rows representing curse/death instances and 4 variables:

movie Film title

profane Whether the event was a profane word (TRUE) or a death (FALSE)

word The specific profane word, if the event was a word

minutes_in The number of minutes into the film the event occurred

Source

See https://github.com/fivethirtyeight/data/tree/master/tarantino

tennis_events_time *Why Some Tennis Matches Take Forever*

Description

The raw data behind the story "Why Some Tennis Matches Take Forever" https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/.

Usage

```
tennis_events_time
```

Format

A data frame with 205 rows representing tournaments and 5 variables:

tournament Name of event

surface Court surface used at the event

sec_added Seconds added per point for this event on this surface in years shown, from regression model controlling for players, year and other factors

year_start Start year for data used from this tournament in regression

year_end End year for data used from this tournament in regression

Source

See https://github.com/fivethirtyeight/data/tree/master/tennis-time

See Also

tennis_players_time and tennis_serve_time

---

tennis_players_time *Why Some Tennis Matches Take Forever*

Description

The raw data behind the story "Why Some Tennis Matches Take Forever" https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/.

Usage

```
tennis_players_time
```

Format

A data frame with 218 rows representing players and 2 variables:

player Player Name

sec_added Weighted average of seconds added per point as loser and winner of matches, 1991-2015, from regression model controlling for tournament, surface, year and other factors

Source

See https://github.com/fivethirtyeight/data/tree/master/tennis-time

See Also

tennis_events_time and tennis_serve_time

---

tennis_serve_time *Why Some Tennis Matches Take Forever*

Description

The raw data behind the story "Why Some Tennis Matches Take Forever" https://fivethirtyeight.com/features/why-some-tennis-matches-take-forever/.

Usage

```
tennis_serve_time
```

Format

A data frame with 120 rows representing serves and 7 variables:

server Name of player serving at 2015 French Open

sec_between Time in seconds between end of marked point and next serve, timed by stopwatch app

opponent Opponent, receiving serve

game_score Score in the current game during the timed interval between points

set Set number, out of five

game Score in games within the set

date Date

Source

See https://github.com/fivethirtyeight/data/tree/master/tennis-time

See Also

tennis_events_time and tennis_players_time

---

tenth_circuit *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

Usage

```
tenth_circuit
```

Format

A data frame with 954 rows representing cases and 13 variables:

title Name of the case

date Date of decision

federalreporter_cit Case citation, as listed in the Federal Reporter Series

westlaw_cit Case citation, Westlaw format

issue Issue number, in cases divided into multiple issues

weight Weight per issue (total weight per case equals one)

judge1 Name of first judge

judge2 Name of second judge

judge3 Name of third judge

vote1_liberal Vote of first judge. 1 = liberal, 0 = conservative.

vote2_liberal Vote of second judge. 1 = liberal, 0 = conservative.

vote3_liberal Vote of third judge. 1 = liberal, 0 = conservative.

category Category of case, immigration or discrimination

Note

In immigration cases, partial relief to immigration petitioner is coded as liberal because the petitioner typically seeks just one core remedy (e.g., withholding of removal, adjustment of status, or asylum); in discrimination cases, partial relief is coded as multiple issues because the plaintiff often seeks separate remedies under multiple claims (e.g., disparate treatment, retaliation, etc.) and different sources of law.

Source

See https://github.com/fivethirtyeight/data/tree/master/tenth-circuit

---

`trumpworld_issues` *What the World Thinks of Trump*

Description

The raw data behind the story "What the World Thinks of Trump" https://fivethirtyeight.com/features/what-the-world-thinks-of-trump/: Trump World Issues Dataset

Usage

```
trumpworld_issues
```

Format

A data frame with 185 rows representing countries and 6 variables:

country The country whose population is being polled

net_approval The difference in the number of respondents from the given country who approve and who disapprove of the issue (Trump proposal) in question (approve-disapprove)

approve The number of respondents from the given country who approve of the issue (Trump proposal)

disapprove The number of respondents who disapprove of the issue

dk_refused undefined

issue The specific trump policy proposal being posed. Specifically: 1: Withdraw support for international climate change agreements 2: Build a wall on the border between the U. S. and Mexico 3: Withdraw U.S. support from the Iran nuclear weapons agreement 4: Withdraw U.S. support for major trade agreements 5: Introduce tighter restrictions on those entering the U.S. from some majority-Muslim countries

Source

Pew Research Center http://www.pewresearch.org/fact-tank/2017/07/17/9-charts-on-how-the-world-sees

See Also

trumpworld_polls

`trumpworld_polls` *What the World Thinks of Trump*

Description

The raw data behind the story "What the World Thinks of Trump" https://fivethirtyeight.com/features/what-the-world-thinks-of-trump/: Trump World Polls Dataset.

Usage

```
trumpworld_polls
```

Format

A data frame with 32 rows representing years and 40 variables:

year Year the poll was conducted

avg The average percentage people who answered the poll question positively (support the president or have a favorable view of the U.S.)

canada The percentage of people from Canada who answered the poll question positively

france The percentage of people from France who answered the poll question positively

germany The percentage of people from Germany who answered the poll question positively

greece The percentage of people from Greece who answered the poll question positively

hungary The percentage of people from Hungary who answered the poll question positively

italy The percentage of people from Italy who answered the poll question positively

netherlands The percentage of people from Netherlands who answered the poll question positively

poland The percentage of people from Poland who answered the poll question positively

spain The percentage of people from Spain who answered the poll question positively

sweden The percentage of people from Sweden who answered the poll question positively

uk The percentage of people from the U.K. who answered the poll question positively

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

Pew Research Center http://www.pewresearch.org/fact-tank/2017/07/17/9-charts-on-how-the-world-sees

See Also

trumpworld_issues

Examples

```
# To convert data frame to tidy data (long) format, run:
library(tidyverse)
trumpworld_polls_tidy <- trumpworld_polls %>%
  gather(country, percent_positive, -c("year", "avg", "question"))
```

---

trump_approval_poll *How Popular is Donald Trump*

Description

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

Usage

```
trump_approval_poll
```

Format

A data frame with 3051 rows representing individual polls and 20 variables:

subgroup The subgroup the poll falls into as defined by the type of people being polled (all polls, voters, adults)

start_date The date the polling began

end_date The date the polling concluded

pollster The polling group which produced the poll

grade The grade for President Trump that the respondents' approval ratings correspond to

sample_size The sample size of the poll

population The type of people being polled (a for adults, lv for likely voters, rv for registered voters)

weight The weight fivethirtyeight gives the poll when determining approval ratings based on historical accuracy of the pollster

approve The percentage of respondents who approve of the president

disapprove The percentage of respondents who disapprove of the president

adjusted_approve The percentage of respondents who approve of the president adjusted for systematic tendencies of the polling firm

adjusted_disapprove The percentage of respondents who 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

https://projects.fivethirtyeight.com/trump-approval-data/approval_polllist.csv and https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv

See Also

trump_approval_trend

`trump_approval_trend` *How Popular is Donald Trump*

Description

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

Usage

```
trump_approval_trend
```

Format

A data frame with 1044 rows representing poll trends and 11 variables:

subgroup The subgroup the poll falls into as defined by the type of people being polled (all polls, voters, adults)

modeldate The date the model was created

approve_estimate Estimated approval ratings

approve_high Higher bound of the estimated approval percentage

approve_low Lower bound of the estimated approval percentage

disapprove_estimate Estimated disapproval percentage

disapprove_high Higher bound of the estimated disapproval percentage

disapprove_low Lower bound of the estimated disapproval percentage

timestamp Date and time the model was compiled

Details

The Variable "president" was removed because all values were "Donald Trump"

Source

https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv

See Also

`trump_approval_poll`

trump_news *How Trump Hacked The Media*

Description

The raw data behind the story "How Trump Hacked The Media" https://fivethirtyeight.com/features/how-donald-trump-hacked-the-media/.

Usage

```
trump_news
```

Format

A data frame with 286 rows representing lead stories and 3 variables:

date Date of lead story about Donald Trump.

major_cat Story classification

detail

Source

Memeorandum http://www.memeorandum.com/.

trump_twitter *The World's Favorite Donald Trump Tweets*

Description

The raw data behind the story "The World's Favorite Donald Trump Tweets" https://fivethirtyeight.com/features/the-worlds-favorite-donald-trump-tweets/. Tweets posted on twitter by Donald Trump (@realDonaldTrump). An analysis using this data was contributed by Adam Spannbauer as a package vignette at http://fivethirtyeight-r.netlify.com/articles/trump_twitter.html.

Usage

```
trump_twitter
```

Format

A data frame with 448 rows representing tweets and 3 variables:

id

created_at

text

Source

Twitter https://twitter.com/realdonaldtrump

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

Description

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

Usage

```
tv_hurricanes
```

Format

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

date Date

harvey The percent of sentences in TV news that mention Hurricane Harvey on the given date

irma The percent of sentences in TV news that mention Hurricane Irma

maria The percent of sentences in TV news that mention Hurricane Maria

jose The percent of sentences in TV news that mention Hurricane Irma

Source

Internet TV News Archive https://archive.org/details/tv and Television Explorer https://television.gdeltproject.org/cgi-bin/iatv_ftxtsearch/iatv_ftxtsearch

See Also

mediacloud_hurricanes, mediacloud_states, mediacloud_online_news, mediacloud_trump, tv_hurricanes_by_network, tv_states, google_trends

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

Description

The raw data behind the story "The Media Really Started Paying Attention to Puerto Rico When Trump Did" https://fivethirtyeight.com/features/the-media-really-started-paying-attention-to-pue
TV Hurricanes by Network Data.

Usage

```
tv_hurricanes_by_network
```

Format

A data frame with 84 rows representing dates and 6 variables:

date Date

query The hurricane in question

bbc_news The percent of sentences on the BBC News TV channel on the given date that mention the hurricane in question

cnn The percent of sentences on CNN News that mention the hurricane in question

fox_news The percent of sentences on Fox News that mention the hurricane in question

msnbc The percent of sentences on MSNBC News that mention the hurricane in question

Source

Internet TV News Archive https://archive.org/details/tv and Television Explorer https://television.gdeltproject.org/cgi-bin/iatv_ftxtsearch/iatv_ftxtsearch

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-pue
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 https://television.gdeltproject.org/cgi-bin/iatv_ftxtsearch/iatv_ftxtsearch

See Also

mediacloud_hurricanes, mediacloud_states, mediacloud_online_news, mediacloud_trump, tv_hurricanes, tv_hurricanes_by_network, google_trends

---

twitter_presidents *The Worst Tweeter in Politics Isn't Trump*

Description

The raw data behind: "The Worst Tweeter in Politics Isn't Trump" https://fivethirtyeight.com/features/the-worst-tweeter-in-politics-isnt-trump/

Usage

```
twitter_presidents
```

Format

A data frame with 6439 rows describing individual tweets and 8 variables:

created_at The date and time the tweet was posted

user The user posting the tweet

text The text of the tweet

url The link to the tweet

replies The number of replies to the tweet

retweets The number of retweets

favorites The number of favorites

Details

Presidents Dataset:

Data on President Obama's tweets collected on Oct 20, President Trump's tweets collected on Oct 23.

Source

Twitter https://twitter.com/BarackObama and https://twitter.com/realDonaldTrump

See Also

senators

undefeated *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/

Usage

```
undefeated
```

Format

A data frame with 2125 rows representing boxing matches and 4 variables:

name Name of boxer

url URL with the boxer's record

date Date of the match

wins Number of cumulative wins for the boxer including the match at the specified date

Source

Box Rec http://boxrec.com/

unisex_names *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-

Usage

```
unisex_names
```

Format

A data frame with 919 rows representing names and 5 variables:

name First names from the Social Security Administration

total Total number of living Americans with the name

male_share Percentage of people with the name who are male

female_share Percentage of people with the name who are female

gap Gap between male_share and female_share

Source

Social Security Administration https://www.ssa.gov/oact/babynames/limits.html. See https://github.com/fivethirtyeight/data/tree/master/unisex-names.

US_births_1994_2003 *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-f

Usage

```
US_births_1994_2003
```

Format

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

year Year

month Month

date_of_month Day

date POSIX date

day_of_week Abbreviation of day of week

births Number of births

Source

Centers for Disease Control and Prevention's National Center for Health Statistics

See Also

US_births_2000_2014

US_births_2000_2014 *Some People Are Too Superstitious To Have A Baby On Friday The 13th*

Description

The raw data behind the story "Some People Are Too Superstitious To Have A Baby On Friday The 13th" https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-f

Usage

```
US_births_2000_2014
```

Format

A data frame with 5479 rows representing dates and 6 variables:

year Year

month Month

date_of_month Day

date POSIX date

day_of_week Abbreviation of day of week

births Number of births

Source

Social Security Administration

See Also

US_births_1994_2003.

---

weather_check *Where People Go To Check The Weather*

Description

The raw data behind the story "Where People Go To Check The Weather" https://fivethirtyeight.com/features/weather-forecast-news-app-habits/.

Usage

```
weather_check
```

Format

A data frame with 928 rows representing respondents and 9 variables:

respondent_id Respondent ID

ck_weather Do you typically check a daily weather report?

weather_source How do you typically check the weather?

weather_source_site If they responded "A specific website or app" when asked how they typically check the weather, they were asked to write-in the app or website they used.

ck_weather_watch If you had a smartwatch (like the soon to be released Apple Watch), how likely or unlikely would you be to check the weather on that device?

age Age

female Gender

hhold_income How much total combined money did all members of your HOUSEHOLD earn last year?

region US Region

Source

The source of the data is a Survey Monkey Audience poll commissioned by FiveThirtyEight and conducted from April 6 to April 10, 2015. See https://github.com/fivethirtyeight/data/tree/master/weather-check

Index