Package ‘predictrace’

May 22, 2019

Title   Predict the Race of a Given Surname Using Census Data
Version 1.0.0
Description Predicts the most common race of a surname based on U.S. Census data.
Depends R (>= 2.10)
License MIT + file LICENSE
Encoding UTF-8
LazyData true
Suggests knitr, rmarkdown, testthat (>= 2.1.0), covr, spelling
VignetteBuilder knitr
Imports dplyr, tidyr, here
RoxygenNote 6.1.1
Language en-US
NeedsCompilation no
Author Jacob Kaplan [aut, cre] (<https://orcid.org/0000-0002-0601-0387>)
Maintainer Jacob Kaplan <jkaplan6@gmail.com>
Repository CRAN
Date/Publication 2019-05-22 07:50:03 UTC

R topics documented:

predict_race ................................................................. 2
surnames_race ............................................................ 2

Index 4
predict_race  Find the race of a surname

Description
Find the race of a surname

Usage
predict_race(name, probability = TRUE)

Arguments
- name: String or vector of strings of surname that you want to know the race of.
- probability: If TRUE (default) will provide columns for each race with the probability that the surname is of that race. If FALSE, will only return the name, the match-name from the Census data, and the most likely race.

Value
A data.frame with three or nine columns: The first column has the name as inputted, the second column has the cleaned up name (no spaces or punctuation, all lowercase), the third column tells the likely race of the surname. If the parameter probability is false, these three columns are all that is returned. Otherwise, columns 4-9 tell the specific probability that the surname is each race.

Examples
predict_race("franklin")
predict_race(c("franklin", "Washington", "Jefferson", "Sotomayor", "Liu"))
predict_race("franklin", probability = FALSE)

surnames_race  Surnames and number of people of each race with that surname.

Description
A dataset containing over 162 thousands surnames and the number of people of each race with that surname.

Usage
surnames_race
surnames_race

Format

A data frame with 162,254 rows and 8 variables:

- **name**  Surname
- **likely_race**  The most likely race based on the probability of each race
- **probability_american_indian**  Probability that the surname is American Indian
- **probability_asian**  Probability that the surname is Asian
- **probability_black**  Probability that the surname is Black
- **probability_hispanic**  Probability that the surname is Hispanic
- **probability_white**  Probability that the surname is White
- **probability_2races**  Probability that the surname is two or more races ...

Source

https://www.census.gov/topics/population/genealogy/data/2010_surnames.html
Index

*Topic datasets
   surnames_race, 2

predict_race, 2

surnames_race, 2