

# Package ‘predictrace’

May 9, 2026

**Title** Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data

**Version** 2.0.1

**Description** Predicts the most common race of a surname and based on U.S. Census data, and the most common first named based on U.S. Social Security Administration data.

**Depends** R (>= 2.10)

**URL** <https://github.com/jacobkap/predictrace>

**BugReports** <https://github.com/jacobkap/predictrace/issues>

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**Suggests** knitr, rmarkdown, testthat (>= 2.1.0), covr, spelling

**VignetteBuilder** knitr

**Imports** dplyr

**RoxygenNote** 7.1.1

**Language** en-US

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Jacob Kaplan [aut, cre] (ORCID:  
<<https://orcid.org/0000-0002-0601-0387>>)

**Maintainer** Jacob Kaplan <jkkaplan6@gmail.com>

**Repository** CRAN

**Date/Publication** 2023-07-05 23:20:02 UTC

## Contents

first_names_gender . . . . .	2
first_names_race . . . . .	2
predict_gender . . . . .	3
predict_race . . . . .	4
surnames_race . . . . .	5

**Index****6**


---

first_names_gender	<i>Surnames and number of people of each race with that first name</i>
--------------------	--

---

**Description**

A dataset containing almost 100,000 first names and the proportion of people with that first name that are female and male.

**Usage**

first\_names\_gender

**Format**

A data frame with 99,444 rows and 4 variables:

**name** The person's first name

**probability\_male** Probability that the first is male

**probability\_female** Probability that the first name is female

**likely\_gender** The most likely gender based on the probability of each gender ...

**Source**

<https://www.ssa.gov/oact/babynames/limits.html>

---

first_names_race	<i>Surnames and number of people of each race with that first name</i>
------------------	--

---

**Description**

A dataset containing over 167 thousands surnames and the number of people of each race with that surname. Citation for this data: Tzioumis, Konstantinos (2018) Demographic aspects of first names, Scientific Data, 5:180025 [dx.doi.org/10.1038/sdata.2018.25].

**Usage**

first\_names\_race

**Format**

A data frame with 4,251 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://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TYJKEZ>

---

predict_gender	<i>Find the gender of a first name</i>
----------------	--

---

**Description**

The surname data comes from the United States Social Security Administration (SSA). This data has the number of people with that name that are identified as female or male so the probability female/male is the proportion of all people with that name that are female/male. SSA data is available annually from 1880-2019, this aggregates all years together.

**Usage**

```
predict_gender(name, probability = TRUE)
```

**Arguments**

<b>name</b>	String or vector of strings of the first name that you want to know the gender of.
<b>probability</b>	If TRUE (default) will provide columns for each race with the probability that the first name is of that gender If FALSE, will only return the name, the match-name from the SSA data, and the most likely gender.

**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 gender of the first name (if there are multiple genders with the same probability of a match, it will be a string with each race separated by a comma). If the parameter probability is false, these three columns are all that is returned. Otherwise, columns 4-5 tell the specific probability that the surname is female or male.

**Examples**

```

predict_gender("tyrion")

predict_gender(c("harry", "ron", "hermione", "DEAN", "NEVILLE", "Cho"))
predict_gender("franklin", probability = FALSE)
predict_gender("jacob", probability = FALSE)
predict_gender("jacob", probability = TRUE)

```

---

predict\_race

*Find the race of a surname or first name*

---

**Description**

The surname data comes from the United States Census. The first name data comes from Tzioumis (2018, <dx.doi.org/10.1038/sdata.2018.25>)

**Usage**

```
predict_race(name, probability = TRUE, surname = TRUE)
```

**Arguments**

name	String or vector of strings of surname or first name 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.
surname	If TRUE (default) will return the race based on the inputted name being a surname. If FALSE, will return the race based on the inputted name being a first name.

**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 or first name (if there are multiple races with the same probability of a match, it will be a string with each race separated by a comma). 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 or first name is each race.

**Examples**

```

predict_race("franklin")

predict_race(c("franklin", "Washington", "Jefferson", "Sotomayor", "Liu"))
predict_race("franklin", probability = FALSE)
predict_race("jacob", probability = FALSE, surname = FALSE)
predict_race("jacob", probability = TRUE, surname = FALSE)

```

---

surnames_race	<i>Surnames and number of people of each race with that surname.</i>
---------------	--

---

**Description**

A dataset containing over 167 thousands surnames and the number of people of each race with that surname.

**Usage**

surnames\_race

**Format**

A data frame with 167,408 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](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html) [https://www.census.gov/topics/population/genealogy/data/2000\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2000_surnames.html)

# Index

## \* datasets

first\_names\_gender, 2

first\_names\_race, 2

surnames\_race, 5

first\_names\_gender, 2

first\_names\_race, 2

predict\_gender, 3

predict\_race, 4

surnames\_race, 5