33. Checking IPUMS download

Daniel Zoleikhaeian

2023-05-24

```
## 2010 IPUMS data
df <- read.csv('C:\\Users\\danie\\Documents\\Joshi Lab Materials\\usa_00002.csv')
head(df)
```

```
##   YEAR SAMPLE SERIAL CBSERIAL HHWT  CLUSTER CITY CITYPOP STRATA GQ PERNUM PERWT
## 1 2010 201001      1       69   96 2.01e+12    0       0  40001  1      1    96
## 2 2010 201001      2       80   97 2.01e+12 4170    1952 220001  1      1    97
## 3 2010 201001      2       80   97 2.01e+12 4170    1952 220001  1      2   128
## 4 2010 201001      2       80   97 2.01e+12 4170    1952 220001  1      3   182
## 5 2010 201001      3      140   90 2.01e+12    0       0  10001  1      1    90
## 6 2010 201001      4      224   82 2.01e+12    0       0 130001  1      1    82
##   RACE RACED HISPAN HISPAND RACHSING
## 1    1   100      0       0        1
## 2    1   100      0       0        1
## 3    1   100      0       0        1
## 4    1   100      0       0        1
## 5    1   100      0       0        1
## 6    1   100      0       0        1
```

```
df_additional<- read.csv('C:\\Users\\danie\\Documents\\Joshi Lab Materials\\3 Studies Dataset\\Dataset I
# head(df_additional)

df_additional$cEnrollmentYear <- substr(df_additional$cEnrollmentDateYear, 1,4)
unique(df_additional$cEnrollmentYear) # get from 2008 to 2020 from ACS
```

```
##  [1] "2008" "2010" "2011" "2009" "2012" "2005" "2015" "2016" "2017" "2018"
## [11] "2019" "2020" ""     "2013" "2014"
```

```
# showing that city populations match up
LA <- df$CITYPOP[df$CITY == 3730 & df$YEAR == 2010][1]
LA * 100
```

```
## [1] 3797100
```

```
1917672 +   1879472 # this is the data from the excel sheet i've been using
```

```
## [1] 3797144
```

```
colnames(df)
```

```
##  [1] "YEAR"     "SAMPLE"   "SERIAL"   "CBSERIAL" "HHWT"     "CLUSTER"
##  [7] "CITY"     "CITYPOP"  "STRATA"   "GQ"       "PERNUM"   "PERWT"
## [13] "RACE"     "RACED"    "HISPAN"   "HISPAND"  "RACHSING"
```

```
unique(df$RACED)
```

```
##   [1] 100 200 669 817 802 304 500 398 310 666 379 620 361 400 662 856 640 399
##  [19] 700 826 801 901 610 670 845 812 362 308 814 905 813 600 323 315 689 663
##  [37] 815 307 830 837 915 660 672 907 838 685 884 902 811 824 854 676 630 314
##  [55] 834 816 324 835 674 819 679 833 321 303 671 821 822 312 673 823 904 664
##  [73] 306 892 667 932 923 372 370 371 941 886 661 302 970 867 682 855 903 680
##  [91] 684 353 841 675 698 320 325 950 980 914 864 317 311 906 883 354 318 952
## [109] 355 357 305 358 677 309 885 319 836 832 933 911 699 913 960 862 934 881
## [127] 825 974 981 990 678 350 925 940 692 930 861 954 882 931 912 943 863 961
## [145] 953 962 356 352 942 951 973 963 935 982 984 972 971 955
```

Things to Note:

- Could not find mutually exclusive race/ethnicity categories for total data
  - only found them for samples from each year
- IPUMS USA has easily accessible 1-year sample data
  - They track mutually exclusive categories via HISPAN and RACE identifiers for each "individual"
  - Each individual has a "weight", which states how many individuals this person represents in the total population
    - Will treat as multiplier

Caveats from IPUMS database

- City populations are rounded to the nearest hundred
  - Uses information from the Decennial census where possible
  - Otherwise, uses the ACS survey data
- RACHSING variable already puts people into mutually exclusive categories
- but combines Asian and Pacific Islander into one category
- Assigns Hispanic of any race (including multiracial) to Hispanic
- Assigns Non-hispanic multiracial people to a single category
  - Does this by inference. Predicted the race of an indidivual based on their age, sex, region, and the urbanization level and racial diversity of their data district. See documentation text in html file.
  - Suggestion: don't use RACHSING variable

Suggested variables to use from IPUMS

- YEAR
- CITY
- CITYPOP
- RACE
- HISPAN

RACE Information

```
race_df <- data.frame(Value = 1:9,
                      Label = c('White', 'Black/African American', 'American Indian or Alaska Native',
knitr::kable(race_df)
```

Value	Label
1	White
2	Black/African American
3	American Indian or Alaska Native
4	Chinese
5	Japanese
6	Other Asian or Pacific Islander
7	Other race
8	Two major races
9	Three or more major races

Suggestions:

- Combine Asian and Pacific Islander into one category
  - Alternative: use RACED variable to separate Asian and Pacific Islander
  - May have to go in one by one to assign each ethnicity to Asian or Pacific Islander (See documentation)
- Combine Mixed race/Other race into one category (National Equity Atlas does this)

HISPAN Information

```
hispan_df <- data.frame(Value = c(0:4, 9),
                        Label = c('Not Hispanic', 'Mexcian',
                                  'Puerto Rican', 'Cuban', 'Other',
                                  'Not Reported'))
knitr::kable(hispan_df)
```

Value	Label
0	Not Hispanic
1	Mexcian
2	Puerto Rican
3	Cuban
4	Other
9	Not Reported

Suggestions:

- Combine “Not Reported” with “Not Hispanic”
- Combine all else into “Hispanic”

Suggested Race/Ethnicity Categories moving forward

- Non-Hispanic (NH) White
- NH Asian/Pacific Islander
  - Note: National Equity Atlas has separate category for Pacific Islander
- NH African American
- NH Native American/Alaska Native
- NH Mixed/Other
- Hispanic