Data_appendix

Lauren Meyer

3/28/2019

Loading Data

Structure and Names

```
str(small_data)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 3190040 obs. of  18 variables:
##  $ YEAR    : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##   ..- attr(*, "label")= chr "Census year"
##   ..- attr(*, "var_desc")= chr "YEAR reports the four-digit year when the household was enumerated or
##  $ SERIAL  : num  1 2 3 3 3 4 4 4 4 4 ...
##   ..- attr(*, "label")= chr "Household serial number"
##   ..- attr(*, "var_desc")= chr "SERIAL is an identifying number unique to each household record in a
##  $ STATEFIP: 'haven_labelled' int  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "State (FIPS code)"
##   ..- attr(*, "var_desc")= chr "STATEFIP reports the state in which the household was located, using
##   ..- attr(*, "labels")= Named num  1 2 4 5 6 8 9 10 11 12 ...
##   .. ..- attr(*, "names")= chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ METRO   : 'haven_labelled' int  0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Metropolitan status"
##   ..- attr(*, "var_desc")= chr "METRO indicates whether the household resided within a metropolitan a
##   ..- attr(*, "labels")= Named num  0 1 2 3 4
##   .. ..- attr(*, "names")= chr  "Metropolitan status indeterminable (mixed)" "Not in metropolitan are
##  $ OWNERSHP: 'haven_labelled' int  2 2 1 1 1 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Ownership of dwelling (tenure) [general version]"
##   ..- attr(*, "var_desc")= chr "OWNERSHP indicates whether the housing unit was rented or owned by it
##   ..- attr(*, "labels")= Named num  0 1 2
##   .. ..- attr(*, "names")= chr  "N/A" "Owned or being bought (loan)" "Rented"
##  $ HHINCOME: 'haven_labelled' num  10000 38500 90700 90700 90700 27100 27100 27100 27100 27100 ...
##   ..- attr(*, "label")= chr "Total household income"
##   ..- attr(*, "var_desc")= chr "HHINCOME reports the total money income of all household members age
##   ..- attr(*, "labels")= Named num 1e+07
##   .. ..- attr(*, "names")= chr "N/A "
##  $ PHONE   : 'haven_labelled' int  2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Telephone availability"
##   ..- attr(*, "var_desc")= chr "PHONE indicates whether residents of the housing unit had telephone a
##   ..- attr(*, "labels")= Named num  0 1 2 8
##   .. ..- attr(*, "names")= chr  "N/A" "No, no phone available" "Yes, phone available" "Suppressed (20
##  $ CINETHH : 'haven_labelled' int  3 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Access to internet"
##   ..- attr(*, "var_desc")= chr "CINETHH reports whether any member of the household accesses the Inte
##   ..- attr(*, "labels")= Named num  0 1 2 3
##   .. ..- attr(*, "names")= chr  "N/A (GQ)" "Yes, with a subscription to an Internet Service" "Yes, wi
##  $ SEX     : 'haven_labelled' int  1 2 1 2 1 2 2 1 2 2 ...
##   ..- attr(*, "label")= chr "Sex"
##   ..- attr(*, "var_desc")= chr "SEX reports whether the person was male or female."
##   ..- attr(*, "labels")= Named num  1 2
##   .. ..- attr(*, "names")= chr  "Male" "Female"
##  $ AGE     : 'haven_labelled' int  73 31 41 48 16 37 18 17 7 3 ...
##   ..- attr(*, "label")= chr "Age"
##   ..- attr(*, "var_desc")= chr "AGE reports the person's age in years as of the last birthday.\n\nPle
##   ..- attr(*, "labels")= Named num  0 90 100 112 115
##   .. ..- attr(*, "names")= chr  "Less than 1 year old" "90 (90+ in 1980 and 1990)" "100 (100+ in 196
##  $ RACE    : 'haven_labelled' int  2 1 1 1 1 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Race [general version]"
##   ..- attr(*, "var_desc")= chr "With the exception of the 1970-1990 Puerto Rican censuses, RACE was a
##   ..- attr(*, "labels")= Named num  1 2 3 4 5 6 7 8 9
##   .. ..- attr(*, "names")= chr  "White" "Black/African American/Negro" "American Indian or Alaska Nat
##  $ EDUC    : 'haven_labelled' int  2 10 6 6 4 6 5 4 1 0 ...
##   ..- attr(*, "label")= chr "Educational attainment [general version]"
##   ..- attr(*, "var_desc")= chr "EDUC indicates respondents' educational attainment, as measured by th
##   ..- attr(*, "labels")= Named num  0 1 2 3 4 5 6 7 8 9 ...
##   .. ..- attr(*, "names")= chr  "N/A or no schooling" "Nursery school to grade 4" "Grade 5, 6, 7, or
##  $ EMPSTAT : 'haven_labelled' int  3 1 1 3 3 1 3 3 0 0 ...
##   ..- attr(*, "label")= chr "Employment status [general version]"
##   ..- attr(*, "var_desc")= chr "EMPSTAT indicates whether the respondent was a part of the labor forc
##   ..- attr(*, "labels")= Named num  0 1 2 3
##   .. ..- attr(*, "names")= chr  "N/A" "Employed" "Unemployed" "Not in labor force"
##  $ OCC     : 'haven_labelled' num  0 350 6260 0 0 230 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Occupation"
##   ..- attr(*, "var_desc")= chr "Universe Note: \"New Workers\" are persons seeking employment for the
##   ..- attr(*, "labels")= Named num  1880 1920 1930 1940 1950 1960 1970 1980 1990 2000
##   .. ..- attr(*, "names")= chr  "Occupation Codes  [URL omitted from DDI.] (used for 1850-1900 sample
##  $ INCTOT  : 'haven_labelled' num  10000 38500 82000 8700 0 ...
##   ..- attr(*, "label")= chr "Total personal income"
##   ..- attr(*, "var_desc")= chr "INCTOT reports each respondent's total pre-tax personal income or los
##   ..- attr(*, "labels")= Named num  -1e+04 -1e+00 0e+00 1e+00 1e+07
##   .. ..- attr(*, "names")= chr  "$9,900 (1980)" "Net loss (1950)" "None" "$1 or break even (2000, 200
##  $ FTOTINC : 'haven_labelled' num  10000 38500 90700 90700 90700 27100 27100 27100 27100 27100 ...
##   ..- attr(*, "label")= chr "Total family income"
##   ..- attr(*, "var_desc")= chr "FTOTINC reports the total pre-tax money income earned by one's family
##   ..- attr(*, "labels")= Named num  -1e+00 0e+00 1e+07 1e+07
##   .. ..- attr(*, "names")= chr  "Net loss (1950) " "No income (1950-2000, ACS/PRCS) " "Not ascertaine
##  $ DIFFEYE : 'haven_labelled' int  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Vision difficulty"
##   ..- attr(*, "var_desc")= chr "DIFFEYE indicates whether the respondent is blind or has serious dif
##   ..- attr(*, "labels")= Named num  0 1 2
##   .. ..- attr(*, "names")= chr  "N/A" "No" "Yes"
##  $ TRANTIME: 'haven_labelled' num  0 50 45 0 0 25 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Travel time to work"
##   ..- attr(*, "var_desc")= chr "TRANTIME reports the total amount of time, in minutes, that it usual
##   ..- attr(*, "labels")= Named num 0
##   .. ..- attr(*, "names")= chr "N/A "
##  - attr(*, "spec")=
##   .. cols_only(
##   ..   YEAR = col_integer(),
##   ..   DATANUM = col_double(),
##   ..   SERIAL = col_double(),
##   ..   CBSERIAL = col_double(),
##   ..   HHWT = col_double(),
##   ..   REGION = col_integer(),
##   ..   STATEFIP = col_integer(),
##   ..   COUNTYFIP = col_double(),
##   ..   METRO = col_integer(),
##   ..   GQ = col_integer(),
##   ..   OWNERSHP = col_integer(),
##   ..   OWNERSHPD = col_integer(),
##   ..   HHINCOME = col_double(),
##   ..   PHONE = col_integer(),
##   ..   CINETHH = col_integer(),
##   ..   PERNUM = col_double(),
##   ..   PERWT = col_double(),
##   ..   SEX = col_integer(),
##   ..   AGE = col_integer(),
##   ..   RACE = col_integer(),
##   ..   RACED = col_integer(),
##   ..   EDUC = col_integer(),
##   ..   EDUCD = col_integer(),
##   ..   EMPSTAT = col_integer(),
##   ..   EMPSTATD = col_integer(),
##   ..   OCC = col_double(),
##   ..   INCTOT = col_double(),
##   ..   FTOTINC = col_double(),
##   ..   DIFFEYE = col_integer(),
##   ..   TRANTIME = col_double()
##   .. )
```

There are 18 variables in the data, and 3,190,040 observations. Variables: * YEAR: The year of the observation.

```
favstats(~YEAR, data = small_data)
```

```
##   min   Q1 median   Q3  max mean sd       n missing
##  2017 2017   2017 2017 2017 2017  0 3190040       0
```

This dataset contains only the 2017 survey data.

- SERIAL: Unique serial number assigned to each household.

```
favstats(~SERIAL, data = small_data)
```

```
##  min       Q1 median      Q3     max   mean       sd       n missing
##    1 335000.8 692532 1047493 1394399 691840 406411.7 3190040       0
```

Expected distribution of a unique sequential code.

- STATEFIP: Numerical code signifying state.

```
favstats(~STATEFIP, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    1 12     27 42  56 27.68654 16.07883 3190040       0
```

Max of 56 seems odd for a country with 50 states, but I'm guessing those are territory markers such as DC and Guam.

- METRO: Is the household in a metropolitan area or not?

```
favstats(~METRO, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    0  1      3  4   4 2.607806 1.481396 3190040       0
```

This one seems to be a set of numerical codes, not an actual integer value. I'll have to convert this one to a factor variable before doing anything else with it.

- OWNERSHP: Code for whether and how the household owns their dwelling.

```
favstats(~OWNERSHP, data = small_data)
```

```
##  min Q1 median Q3 max    mean        sd       n missing
##    0  1      1  2   2 1.22198 0.5168155 3190040       0
```

This one is also an integer code, with levels 0, 1, and 2. Will need to convert to factor.

- HHINCOME: Total household income.

```
favstats(~HHINCOME, data = small_data)
```

```
##     min    Q1 median     Q3     max     mean      sd       n missing
##  -16200 40300  77000 134000 9999999 565907.2 2101989 3190040       0
```

I'm curious how a household manages to get a negative household income. Maybe it factors in debts and other things?

- PHONE: Availability of a telephone.

```
favstats(~PHONE, data = small_data)
```

```
##  min Q1 median Q3 max     mean        sd       n missing
##    0  2      2  2   2 1.893795 0.4351077 3190040       0
```

Another factor variable.

- CINETHH: Whether any member of the household has internet access.

```
favstats(~CINETHH, data = small_data)
```

```
##  min Q1 median Q3 max     mean        sd       n missing
##    0  1      1  1   3 1.147781 0.6275465 3190040       0
```

A lot of these seem to be factor variables.

- SEX: Binary int, male vs female.

```
favstats(~SEX, data = small_data)
```

```
##  min Q1 median Q3 max     mean        sd       n missing
##    1  1      2  2   2 1.510606 0.4998876 3190040       0
```

As expected, factors numbered 1 and 2.

- AGE: Age of the respondent.

```
favstats(~AGE, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    0 21     42 60  96 41.28723 23.63224 3190040       0
```

Youngest is 0, oldest is 96. Seems accurate.

- RACE: Numerical code for race of respondent.

```
favstats(~RACE, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    1  1      1  1   9 1.820196 1.890103 3190040       0
```

Seems to be another factor variable.

- EDUC: Highest year of schooling/educational attainment.

```
favstats(~EDUC, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    0  4      6  8  11 6.147683 3.252049 3190040       0
```

This survey really likes integer codes.

- EMPSTAT: Employment status.

```
favstats(~EMPSTAT, data = small_data)
```

```
##  min Q1 median Q3 max     mean       sd       n missing
##    0  1      1  3   3 1.499411 1.126316 3190040       0
```

This survey *really* likes integer codes.

- OCC: Occupation of worker.

```
favstats(~OCC, data = small_data)
```

```
##  min Q1 median   Q3  max     mean       sd       n missing
##    0  0   1007 4710 9920 2505.692 2898.312 3190040       0
```

This one also seems coded, but there are a lot more codes than the rest of them.

- INCTOT: Personal income.

```
favstats(~INCTOT, data = small_data)
```

```
##    min    Q1 median    Q3     max    mean      sd       n missing
##  -9100 10600  33500 91000 9999999 1721082 3730118 3190040       0
```

Again, negative income? Unsure how that factors in.

- FTOTINC: Total family income.

```
favstats(~FTOTINC, data = small_data)
```

```
##     min      Q1 median     Q3     max     mean      sd       n missing
##  -16200 35417.5  71000 129000 9999999 573529.2 2130179 3190040       0
```

I'm not sure where this negative income is coming from. At least the numbers otherwise make sense!

- DIFFEYE: Any vision disability, blindness, etc.

```
favstats(~DIFFEYE, data = small_data)
```

```
##  min Q1 median Q3 max    mean        sd       n missing
##    1  1      1  1   2 1.02629 0.1599962 3190040       0
```

This one seems simple, a basic binary code.

- TRANTIME: Length of commute.

```
favstats(~TRANTIME, data = small_data)
```

```
##  min Q1 median Q3 max    mean       sd       n missing
##    0  0      0 20 160 11.6963 20.22848 3190040       0
```

I pity the person who has an 160-minute commute to work, but the rest of it looks good.

Most Pressing, to do: Convert the integer codes to factor levels, possibly rename the factor levels. Decide which variables we want to work with so we don't waste time fixing data we don't end up using.