Cleaning Data With Stata

Andy Grogan-Kaylor

27 Feb 2021 19:31:46

# Background

It sometimes seems like 80% of our work as data analysts is cleaning the data, while only 20% is the actual analysis. Here are some Stata commands that are useful in cleaning data.

First we simulate some data to work with, and to clean.

# Simulate Some Data

This section is provided for illustration only, as it may be helpful to see *how* the data was simulated, and the decisions that went into simulating the data. You may also *safely ignore* this section if you like.

Show / Hide Data Simulation Code

. clear all

. set obs 100 // 100 observations  
number of observations (\_N) was 0, now 100

. generate id = \_n // random id

. generate age = rnormal(50,10) // random generated age

. replace age = 200 in 1 // someone is 200 years old!  
(1 real change made)

. generate happy = runiformint(1,5) // randomly generated happiness

. replace happy = 999 in 10 // simulate a missing value  
(1 real change made)

. generate somethingelse = rnormal(0, 1) // something else!

# Look At Some Of The Data

. list in 1/10 // list first 10 observations  
  
 ┌───────────────────────────────────┐  
 │ id age happy somethi~e │  
 ├───────────────────────────────────┤  
 1. │ 1 200 5 -1.458761 │  
 2. │ 2 44.62952 1 -.6263652 │  
 3. │ 3 39.00959 5 .0041983 │  
 4. │ 4 56.09353 5 -.7921791 │  
 5. │ 5 47.48186 5 -.3584003 │  
 ├───────────────────────────────────┤  
 6. │ 6 41.74093 3 .2928382 │  
 7. │ 7 47.3314 4 -.7735971 │  
 8. │ 8 46.24969 4 -.3540395 │  
 9. │ 9 46.3789 1 -.1536392 │  
 10. │ 10 22.72754 999 .698948 │  
 └───────────────────────────────────┘

# Clean The Data!

## Look at Variables (describe, summarize, tabulate)

When we look at variables we are looking for values that don't make sense, or that are outside the plausible range.

. describe // describe the data  
  
Contains data  
 obs: 100   
 vars: 4   
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
 storage display value  
variable name type format label variable label  
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
id float %9.0g   
age float %9.0g   
happy float %9.0g   
somethingelse float %9.0g   
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
Sorted by:   
 Note: Dataset has changed since last saved.

. summarize // descriptive statistics  
  
 Variable │ Obs Mean Std. Dev. Min Max  
─────────────┼─────────────────────────────────────────────────────────  
 id │ 100 50.5 29.01149 1 100  
 age │ 100 49.70431 18.30524 22.72754 200  
 happy │ 100 12.85 99.62185 1 999  
somethinge~e │ 100 -.0207823 .8726572 -2.348869 2.685713

. tabulate happy // tabulation of this particular categorical variable  
  
 happy │ Freq. Percent Cum.  
────────────┼───────────────────────────────────  
 1 │ 24 24.00 24.00  
 2 │ 20 20.00 44.00  
 3 │ 18 18.00 62.00  
 4 │ 17 17.00 79.00  
 5 │ 20 20.00 99.00  
 999 │ 1 1.00 100.00  
────────────┼───────────────────────────────────  
 Total │ 100 100.00

Notice that...

* There are variables in which we may not have interest.
* None of the variables are labelled informatively.
* Variables do not seem to have informative value labels.
* Someone appears to 200 years old.
* There appear to be missing values in the variable happy that need to be recoded.

Remember that the command lookfor is often very helpful in *looking for* a particular variable. e.g. lookfor happy.

## Only keep The Variables Of Interest

We may only be interested in keeping some variables to keep our analytic data set more manageable.

For this particular analysis we may wish to drop the variable called somethingelse.

. keep id age happy // keep only relevant variables

We could also have said drop somethingelse.

## Add *Variable* Labels (label variable "...")

. label variable id "ID" // label variable

. label variable age "Age in Years" // label variable

. label variable happy "Happiness Scale" // label variable

## Create *Value* Labels (label define ...)

. label define happy 1 "Rarely" 2 "Sometimes" 3 "Often" 4 "Always" // create value label

## Attach *Value* Labels To *Variables* (label values ...)

*Variables* and *value labels* can have the same names but are different things. We add the variable label happy to the variable named happy.

. label values happy happy // attach VALUE LABEL happy to VARIABLE happy

## Recode Outliers or Values That Are Errors (recode)

. recode happy (999 = .) // recode values as missing  
(happy: 1 changes made)

. recode age (100/max = 100) // age is topcoded at 100 (may or may not be plausible)  
(age: 1 changes made)

# We describe and summarize The Data And See The Changes That Have Been Made

. describe  
  
Contains data  
 obs: 100   
 vars: 3   
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
 storage display value  
variable name type format label variable label  
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
id float %9.0g ID  
age float %9.0g Age in Years  
happy float %9.0g happy Happiness Scale  
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
Sorted by:   
 Note: Dataset has changed since last saved.

. summarize  
  
 Variable │ Obs Mean Std. Dev. Min Max  
─────────────┼─────────────────────────────────────────────────────────  
 id │ 100 50.5 29.01149 1 100  
 age │ 100 48.70431 11.46535 22.72754 100  
 happy │ 99 2.888889 1.470033 1 5