```
*****************************************
* First Stata command file
*****************************************



log using "k:\debug\race_good.log", replace

/*****************************************************************
* This command file demonstrates how to code the racial background of respondents
* The command file was created on April 12, 2015
*
* The original data  , race2.dta, had 107 observations and 3 variables
* The final data file, race3.dta, had 100 observations and 7 variables




* The racial background was determined by two variables: race and racemore
* We want to create 6 variables, including
        1) Five dummy variables indicating if
           respondents have any of the five culture background
        2) one summary varialbe to indicate the respondent's background.
           The final response category of this summary variable include
        (1) American Indian or Alaskan Native
        (2) Asian/Pacific Islander
        (3) Black/African American
        (4) Hispanic
        (5) White or Caucasian
        (6) Multi-Racial


* The command file achieves the following tasks
* (1) Remove three empty observations
* (2) Remove duplicate records
* (3) Correct the erroneous errors in the race variable
* (4) Use the racemore variable to supplement/recode the race variable
* (5) Use the race variable to generate five dummy variables
* (6) Recode the race variable to include the multi-racial category
* (7) Select variables and save the data file

*******************************************************************/


use "k:\debug\race2.dta", clear

set more 1

*********************************************
* Check the data:
* There were 107 records and 3 string variables
*********************************************

des

*****************************
* Check if everyone had a value on the ID variable
* Dropped three empty records

* The number of records was 104
*****************************

count if id ==""
sum  id race racemore if id ==""
list id race racemore if id ==""

drop if id ==""

count

***************************************
* Check if there were duplicates IDs
***************************************

*************************
* Use the one-way frequency
*************************
tab1 id, mis


******************
```

```
* Use the indicator
*******************
sort id
by id: gen n = _n
by id: gen N = _N
list id N if N >=2

*****************************
* Use the duplicate command
*****************************

duplicates report id

duplicates tag id, generate(N2)
list id N2 if N2 >=1

************************
* Decison:
* I kept only one record for S100, becuase these two records were duplicate
* I kept none of the records for S23, becasue we do not know which record was the accurate one

* The final number of observation should be 100
************************
list if N2 >=1


drop if id =="S100" & n ==2
drop if id == "S23"

count


*******************
* check the data again
*******************
duplicates report id



****************************
* check for typos in the race variable
****************************


tab1 race, mis

replace race = "3,5"     if race == "3,5,,"
replace race = "4,5"     if race == "4, 5"


tab1 race, mis



**************************************************
* Use the information from racemore to code supplement the race variable
**************************************************

gen race_old = race

sort race

list id race_old race racemore if racemore ~=""

***********************************************
* Recode the racial background for three persons
***********************************************

replace race = "3,5,7" if id == "S316"
replace race = "4" if id == "S248"
replace race = "5" if id == "S29"



list id race_old race racemore if inlist(id, "S316","S248","S29")



        ************************************************
        * Create 5 dummy variables for racial background
        ************************************************

*******************************
* Look at the race variable
```

```
*******************************

tab1 race, mis


*******************************************************************************
* Use the parse command to split the race variables into three string variables
*******************************************************************************

split race, parse(,) gen(newrace)
tab1 newrace*, mis


*******************************************************************************
* Change the three string variables to three numeric variables, so I could use the anymatch command later
*******************************************************************************
destring newrace*, replace
tab1 newrace*, mis




        ************************
        * Create the dummay variables
        ************************


egen native   =  anymatch(newrace1 newrace2 newrace3),value(1)
egen asian    =  anymatch(newrace1 newrace2 newrace3),value(2)
egen black    =  anymatch(newrace1 newrace2 newrace3),value(3)
egen hispanic =  anymatch(newrace1 newrace2 newrace3),value(4)
egen white    =  anymatch(newrace1 newrace2 newrace3),value(5)


        *************************
        * check the data
        *************************

list


************************
* Add variable and value labels
************************
label variable native "Racial group: Native American"
label variable asian "Racial group: Asian"
label variable black "Racial group: Black"
label variable hispanic "Racial group: Hispanic"
label variable white "Racial group: White"


label define yesno 1 "yes" 0 "no"
label value native yesno
label value asian yesno
label value black yesno
label value hispanic yesno
label value white yesno



************************
* check the data
************************

        ******************************
        * recode the yrace variables
        ******************************
******************
* recode yrace variable
******************

gen raceold2 = race


tab1 race, mis

replace race = "6" if race =="1,4,5"
replace race = "6" if race =="1,5"
replace race = "6" if race =="1,5,7"
replace race = "6" if race =="3,5"
replace race = "6" if race =="3,5,7"
replace race = "6" if race =="4,5"
replace race = "6" if race =="4,5,7"
replace race = "6" if race =="5,7"
```

```
*******************************
* Add variable label and value labels
*******************************
label variable race "racial background"

label define race                                ///
    1 "American Indian or Alaskan Native"        ///
    2 "Asian/Pacific Islander"                   ///
    3 "Black/African American"                   ///
    4 "Hispanic"                                 ///
    5 "White or Caucasian"                       ///
    6 "Multi-Racial"


destring race, replace
label value race race



tab2 raceold2 race, mis


*******************
* Save the data
*******************

keep id race native asian black hispanic white


save "k:\debug\race3.dta", replace

log close
```