```
/*******************************************************************************
This command file is for analyzing the impact of race, gender, and aging on depressed mood

The data files were from Add Health public data collected at Waves 3,4, and 5. A total of 3,147
respondents participated in all these three waves of Add Health surveys.

The final data file is w345.dta (3,002 respondents and 9,006 observations) after dropping
respondents with inconsistent racial backgrounds or with missing values on dependent or
independent variables.


File last modified on 5/22/23


Note:

1. The research question is whether respondents of different racial backgrounds differ in the
trajectory of depressive mood between ages 18 and 45.

2. Research hypotheses are if women are more depressed than men between ages 18 and 45 and if
white respondents are more likely than members of minority groups to experience a depressed
mood.

3. Variables needed.
   3.1: The first hypothesis: gender, age, depression, and weight variable
   3.2: the second hypothesis: race, age, depression, and weight variable

4. The data files were from Waves 3, 4, 5 of Add Health survey

5. The depressed mood is measured by three questions asking whether respondents felt depressed,
felt sad, or cannot shake off the blues. These three questions were asked at Waves 3, 4, 5 of
Add Health data. The wordings and response categories are the same for these questions across
waves.

6. Information on Respondent's gender were collected with one question at Waves 3, 4, with same
wordings and response categories.


7. Information on Respondents racial backgrounds were collected with several questions at waves
3 and 5.  Respondents were allowed to report more than one racial background. there is an
additional question measuring the racial backgrounds of respondents if they have more than one
racial backgrounds.

8. Respondent's age at each wave were created by using the interview dates and respondent's
birthdates. Many respondents reported different birthdates at W5 than they did at W3 and W4,
suggesting errors in birthdates reported at W5. I decided to use birthdates at W3 to calculate
respondents' ages at w5.


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

log using "D:\Jason\workshop\Stata workshop\data management_stata editor.log", replace


************************************
* Read in Wave 3 data
* The codebook showed the number of observation is 4,882
* The name of ID variable: aid
************************************

*******************************************************************************************************
*
* Check the accuracy of w3 data
* N = 4,882
* Each value of the ID variable is associated with one record, meaning the data is in wide
format
*******************************************************************************************************
*
use "D:\jason\workshop\Stata workshop\ICPSR_21600\DS0008\21600-0008-Data.dta", clear
```

```
*********************************************************
* change the name of variable from uppercase to lowercase
*********************************************************
rename *, lower

************************
* Check if the number of observation is corrected
************************
count


*****************************************************************************
* check if each value of the ID variable is associated with just one record
*****************************************************************************

duplicates report aid

sort aid
bysort aid: gen n = _n
bysort aid: gen N = _N
tab1 n N, mis
drop n N

*****************************************************************************
* select variables
*****************************************************************************

keep aid iyear3 imonth3 h3od1y h3od1m bio_sex3 h3od4a h3od4b h3od4c h3od4d h3od6 h3od2 h3sp9
h3sp12 h3sp6


*************************************************************************************************
* Create a temporary data file to avoid accidentally overwriting the original data file
*************************************************************************************************

save "D:\jason\workshop\Stata workshop\w3_temp.dta", replace


*****************************************************
* read in the temporary data file
*****************************************************

use "D:\jason\workshop\Stata workshop\w3_temp.dta", clear



*****************************************
* check the attributes of the variables
*****************************************

des aid iyear3 imonth3 h3od1y h3od1m bio_sex3 h3od4a h3od4b h3od4c h3od4d h3od6 h3od2 h3sp9
h3sp12 h3sp6


*****************************************
* Made a copy of these variable
*****************************************
clonevar       Interview_y3   = iyear3
clonevar       Interview_m3   = imonth3
clonevar       birth_y3       = h3od1y
clonevar       birth_m3       = h3od1m
clonevar       female_w3      = bio_sex3
clonevar       white_w3       = h3od4a
clonevar       black_w3       = h3od4b
clonevar       indian_w3      = h3od4c
clonevar       asian_islander_w3 = h3od4d
clonevar       best_category_w3 = h3od6
clonevar       hispanic_w3    = h3od2
clonevar       depressed_w3   = h3sp9
```

```
clonevar       sad_w3         = h3sp12
clonevar       blues_w3       = h3sp6


************************
* Create the Age variable
************************


tab1 Interview_y3 Interview_m3 birth_y3 birth_m3, mis
tab1 Interview_y3 Interview_m3 birth_y3 birth_m3, nol mis


gen cm_interview_w3 = (Interview_y3-1900)*12+Interview_m3
gen cm_birth_w3     = (birth_y3-1900)*12+birth_m3


label variable cm_interview_w3 "interview date at W3, century month"
label variable cm_birth_w3     "birth date from W3, century month"


tab1 cm_interview_w3 cm_birth_w3, mis


gen age_w3_c = (cm_interview_w3-cm_birth_w3)/12
gen age_w3_i = floor((cm_interview_w3-cm_birth_w3)/12)



label variable age_w3_c "respondent's age at w3, continuous"
label variable age_w3_i "respondent's age at w3, integer"

list aid Interview_y3 Interview_m3 birth_y3 birth_m3 age_w3_c age_w3_i in 1/50, nol
sum age_w3_i age_w3_c


***************************
* Respondent's gender at W3
***************************

tab1 female_w3, mis
tab1 female_w3, nol mis

replace female_w3 = female_w3 -1
label define female_w3 1 "(1) female" 0 "(0) male"
label value female_w3 female_w3
label variable female_w3 "respondent's sex at W3"

tab2 bio_sex3 female_w3, mis


*****************************
* Respondent's race at W3
*****************************


tab1 white_w3 black_w3 indian_w3 asian_islander_w3 hispanic_w3, mis

gen str1 s_white_w3 ="."
gen str1 s_black_w3 ="."
gen str1 s_indian_w3 ="."
gen str2 s_asian_islander_w3 =".."
gen str1 s_hispanic_w3 ="."


replace s_white_w3 ="W"              if white_w3 ==1
replace s_black_w3 = "B"             if black_w3 ==1
replace s_indian_w3 ="I"             if indian_w3 ==1
replace s_asian_islander_w3 = "AI"   if asian_islander_w3 ==1
replace s_hispanic_w3 ="H"           if hispanic_w3 ==1
```

```
label variable s_white_w3          "RACE-WHITE-W3, string"
label variable s_black_w3          "RACE-BLACK/AFRICAN AM-W3, string"
label variable s_indian_w3         "RACE-AMER INDIAN/NATIVE Amerian-W3, string"
label variable s_asian_islander_w3 "RACE-ASIAN-ISlander-W3, string"
label variable s_hispanic_w3       "ARE YOU OF HISPANIC ORIGIN-W3, string"



gen race_pattern_w3 = s_white_w3 + "_" + s_black_w3 + "_" + s_indian_w3 + "_" +
s_asian_islander_w3 + "_" + s_hispanic_w3
label variable race_pattern_w3 "racial backgrounds at W3"

des race_pattern_w3
tab1 race_pattern_w3, mis



*********************************************************************
* Incorporate the information from the best race category variable
*********************************************************************

tab2 race_pattern_w3 best_category_w3, mis

gen race_w3 =.
label variable race_w3 "race at w3"
label define race_w3 1 "(1) white"  2 "(2) black"  3 "(3) indians" 4 "(4) asian/islanders" 5
"(5) hispanic"
label value race_w3 race_w3


*code respondents with only one racial backgrounds

replace race_w3 = 1 if race_pattern_w3 =="W_._._.._."
replace race_w3 = 2 if race_pattern_w3 =="._B_._.._."
replace race_w3 = 3 if race_pattern_w3 =="._._I_.._."
replace race_w3 = 4 if race_pattern_w3 =="._._._AI_."
replace race_w3 = 5 if race_pattern_w3 =="._._._.._H"

* code respondents with some hispanic backgrounds
replace race_w3 = 5 if inlist(race_pattern_w3, "._._._AI_H", "._._I_.._H", "._B_._.._H",
"._B_I_.._H")
replace race_w3 = 5 if inlist(race_pattern_w3, "W_._._.._H","W_._._AI_H", "W_._I_.._H",
"W_._I_AI_H")
replace race_w3 = 5 if inlist(race_pattern_w3, "W_B_._.._H", "W_B_._AI_H", "W_B_I_.._H")

replace race_w3 = 1 if best_category_w3 ==1 & inlist(race_pattern_w3,"W_._._AI_.", "W_._I_.._.",
"W_._I_AI_.", "W_B_._.._.", "W_B_I_.._.")

replace race_w3 = 2 if best_category_w3 ==2 & inlist(race_pattern_w3,"W_B_._.._.", "W_B_I_.._.",
"._B_._AI_.", "._B_I_.._.", "W_B_._AI_.", "W_B_I_AI_.")
replace race_w3 = 3 if best_category_w3 ==3 & inlist(race_pattern_w3,"W_._I_.._.", "W_B_I_.._.",
"._B_I_.._.")
replace race_w3 = 4 if best_category_w3 ==4 & inlist(race_pattern_w3,"W_._._AI_.", "W_._I_AI_.")
*replace race_w3 = 5 if best_category_w3 ==5 & inlist(race_pattern_w3,"W_B_._.._.",
"W_B_I_AI_.")

tab1 race_w3, mis



*************************************************
* Depressed mood
*************************************************

* check out the scale items

tab1 depressed_w3 sad_w3 blues_w3, mis
tab1 depressed_w3 sad_w3 blues_w3, mis nol
```

```
* recode the values of 6 through 9 to missing for these variables
recode  depressed_w3 sad_w3 blues_w3 (6/9 =.)

* check the accuracy of variables
tab2 h3sp9     depressed_w3, mis
tab2 h3sp12    sad_w3, mis
tab2 h3sp6     blues_w3, mis


* check the inter-item associations of the depression mood items

factor depressed_w3 sad_w3 blues_w3
alpha depressed_w3 sad_w3 blues_w3


*three methods of aggregating the scale items

* the sum of scale items
gen d_score_w3_sum = depressed_w3+sad_w3+blues_w3

* the rowtotal function of scale items
egen d_score_w3_egen = rowtotal(depressed_w3 sad_w3 blues_w)

* the factor score of the scale items
factor depressed_w3 sad_w3 blues_w3
predict d_score_w3_fs

label variable d_score_w3_sum "depression score at W3, sum score"
label variable d_score_w3_egen "depression score at W3, egen function"
label variable d_score_w3_fs "depression score at W3, factor score"


sum d_score_w3_sum d_score_w3_egen d_score_w3_fs

list aid depressed_w3 sad_w3 blues_w3 d_score_w3_sum d_score_w3_egen d_score_w3_fs if
d_score_w3_sum ==., sepby(aid) nol


*********************************
* generate the indicator variable
*********************************

gen in_w3 =1
label variable in_w3 "respondents of W3"
label define in_w3 0 " (0) not in W3" 1 "(1) in W3"
label value in_w3 in_w3
tab1 in_w3, mis


*****************************
* Save the data
*****************************
sort aid
save "D:\jason\workshop\Stata workshop\w3.dta", replace




************************************
* Read in Wave 4 data
* The codebook showed the number of observation is 4,882
* The name of ID variable: aid
************************************

*****************************************************************************************************
*
* Check the accuracy of w3 data
* N = 5,114
* Each value of the ID variable is associated with one record, meaning the data is in wide
format
```

```
*************************************************************************************************
*
use "D:\jason\workshop\Stata workshop\ICPSR_21600\DS0022\21600-0022-Data.dta" , clear



*********************************************************
* change the name of variable from uppercase to lowercase
*********************************************************
rename *, lower

************************
* Check if the number of observation is correct
************************
count


*****************************************************************************
* check if each value of the ID variable is associated with just one record
*****************************************************************************

duplicates report aid

sort aid
bysort aid: gen n = _n
bysort aid: gen N = _N
tab1 n N, mis
drop n N




*****************************************************************************
* select variables
*****************************************************************************

keep aid iyear4 imonth4 h4od1y h4od1m bio_sex4 h4mh22 h4mh26 h4mh19



*************************************************************************************************
* Create a temporary data file to avoid accidentally overwriting the original data file
*************************************************************************************************

save "D:\jason\workshop\Stata workshop\w4_temp.dta", replace


*****************************************************
* read in the temporary data file
*****************************************************

use "D:\jason\workshop\Stata workshop\w4_temp.dta", clear

*****************************************
* check the attributes of the variables
*****************************************

des aid iyear4 imonth4 h4od1y h4od1m bio_sex4 h4mh22 h4mh26 h4mh19


*****************************************
* Makde a copy of these variables
*****************************************
clonevar       Interview_y4   =      iyear4
clonevar       Interview_m4   =      imonth4
clonevar       birth_y4       =      h4od1y
clonevar       birth_m4       =      h4od1m
clonevar       female_w4      =      bio_sex4
clonevar       depressed_w4   =      h4mh22
clonevar       sad_w4         =      h4mh26
clonevar       blues_w4       =      h4mh19
```

```
************************
* Create the Age variable
************************


tab1 Interview_y4 Interview_m4 birth_y4 birth_m4, mis
tab1 Interview_y4 Interview_m4 birth_y4 birth_m4, nol mis


gen cm_interview_w4 = (Interview_y4-1900)*12+Interview_m4
gen cm_birth_w4     = (birth_y4-1900)*12+birth_m4


label variable cm_interview_w4 "interview date at W4, century month"
label variable cm_birth_w4     "birth date from W4, century month"


tab1 cm_interview_w4 cm_birth_w4, mis


gen age_w4_c = (cm_interview_w4-cm_birth_w4)/12
gen age_w4_i = floor((cm_interview_w4-cm_birth_w4)/12)



label variable age_w4_c "respondent's age at w4, continuous"
label variable age_w4_i "respondent's age at w4, integer"

list aid Interview_y4 Interview_m4 birth_y4 birth_m4 age_w4_c age_w4_i in 1/50, nol
sum age_w4_i age_w4_c



***************************
* Respondent's gender at W4
***************************

tab1 female_w4, mis
tab1 female_w4, nol mis

replace female_w4 = female_w4 -1
label define female_w4 1 "(1) female" 0 "(0) male"
label value female_w4 female_w4
label variable female_w4 "respondent's sex at W4"

tab2 bio_sex4 female_w4, mis


************************************************
* Depressed mood
************************************************

* check out the scale items

tab1 depressed_w4 sad_w4 blues_w4, mis
tab1 depressed_w4 sad_w4 blues_w4, mis nol

* recode the values of 6 through 9 to missing for these variables
recode  depressed_w4 sad_w4 blues_w4 (6/9 =.)

* check the accuracy of variables
tab2 h4mh22    depressed_w4, mis
tab2 h4mh26    sad_w4, mis
tab2 h4mh19    blues_w4, mis



* check the inter-item associations of the depression mood items
```

```
factor depressed_w4 sad_w4 blues_w4
alpha depressed_w4 sad_w4 blues_w4


*three methods of aggregating the scale items

* the sum of scale items
gen d_score_w4_sum = depressed_w4 + sad_w4 + blues_w4

* the rowtotal function of scale items
egen d_score_w4_egen = rowtotal(depressed_w4 sad_w4 blues_w4)

* the factor score of the scale items
factor depressed_w4 sad_w4 blues_w4
predict d_score_w4_fs

label variable d_score_w4_sum "depression score at W4, sum score"
label variable d_score_w4_egen "depression score at W4, egen function"
label variable d_score_w4_fs "depression score at W4, factor score"


sum d_score_w4_sum d_score_w4_egen d_score_w4_fs

list aid depressed_w4 sad_w4 blues_w4 d_score_w4_sum d_score_w4_egen d_score_w4_fs if
d_score_w4_sum ==., sepby(aid) nol



*********************************
* generate the indicator variable
*********************************

gen in_w4 =1
label variable in_w4 "respondents of W4"
label define in_w4 0 " (0) not in W4" 1 "(1) in W4"
label value in_w4 in_w4
tab1 in_w4, mis


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

sort aid
save "D:\jason\workshop\Stata workshop\w4.dta", replace







************************************
* Read in Wave 5 data
* The codebook showed the number of observation is 4,882
* The name of ID variable: aid
************************************

*****************************************************************************************************
*
* Check the accuracy of w3 data
* N = 4,196
* Each value of the ID variable is associated with one record, meaning the data is in wide
format
*****************************************************************************************************
*
use "D:\jason\workshop\Stata workshop\ICPSR_21600\DS0032\21600-0032-Data.dta", clear

*******************************************************
* change the name of variable from uppercase to lowercase
*******************************************************
```

```
rename *, lower

************************
* Check if the number of observation is correctd
************************
count


*******************************************************************************
* check if each value of the ID variable is associated with just one record
*******************************************************************************

duplicates report aid

sort aid
bysort aid: gen n = _n
bysort aid: gen N = _N
tab1 n N, mis
drop n N


*******************************************************************************
* select variables
*******************************************************************************

keep aid iyear5 imonth5 h5od1y h5od1m h5od2a h5od4a h5od4b h5od4f h5od4d h5od4e h5od8 h5od4c
h5od4g h5ss0b h5ss0d h5ss0a


*************************************************************************************************
* Create a temporary data file to avoid accidentally overwriting the original data file
*************************************************************************************************

save "D:\jason\workshop\Stata workshop\w5_temp.dta", replace


*****************************************************
* read in the temporary data file
*****************************************************

use "D:\jason\workshop\Stata workshop\w5_temp.dta", clear

*****************************************
* check the attributes of the variables
*****************************************

des aid iyear5 imonth5 h5od1y h5od1m h5od2a h5od4a h5od4b h5od4f h5od4d h5od4e h5od8 h5od4c
h5od4g h5ss0b h5ss0d h5ss0a



*Add value labels
tab1 iyear5 imonth5 h5od1y h5od1m h5od2a h5od4a h5od4b h5od4f h5od4d h5od4e h5od8 h5od4c h5od4g
h5ss0b h5ss0d h5ss0a, mis


*
label define h5od2a 2 "(2) female" 1 "(1) male"
label value  h5od2a h5od2a

tab1 h5od4a h5od4b h5od4f h5od4d h5od4e h5od8 h5od4c h5od4g, mis
label define yesno 1 "(1) yes" 0 "(1) no"

label value h5od4a yesno
label value h5od4b  yesno
label value h5od4f  yesno
label value h5od4d  yesno
label value h5od4e  yesno
label value h5od4c  yesno
label value h5od4g  yesno
```

```
label define h5od8 1 "(1) White"  2 "(2) Black, African American" 3 "(3) Hispanic" 4 "(4)
Hispanic: Mexican, Mexican American, Chicano" ///
7 "(7) Other Hispanic, Latino, or Spanish origin" 11 "(11) Asian: Filipino" 15 "(15) Other
Asian" 21 "(21) American Indian or Alaska Native" ///
997 "(997) legitimate skip"


label value h5od8 h5od8

tab1 h5od8, mis


label define depressed 1 "(1) never or rarely"  2 "(2) sometimes"  ///
3 "(3) a lot of the time" 4 "(4) most of the time or all of the time"

label value h5ss0a depressed
label value h5ss0b depressed
label value h5ss0d depressed

tab1 h5ss0a h5ss0b h5ss0d, mis



****************************************
* Makde a copy of these variables
****************************************
clonevar       Interview_y5   =      iyear5
clonevar       Interview_m5   =      imonth5
clonevar       birth_y5       =      h5od1y
clonevar       birth_m5       =      h5od1m
clonevar       female_w5      =      h5od2a
clonevar       white_w5       =      h5od4a
clonevar       black_w5       =      h5od4b
clonevar       indian_w5      =      h5od4f
clonevar       asian_w5       =      h5od4d
clonevar       islander_w5    =      h5od4e
clonevar       best_category_w5=     h5od8
clonevar       hispanic_w5    =      h5od4c
clonevar       other_w5       =      h5od4g
clonevar       depressed_w5   =      h5ss0b
clonevar       sad_w5         =      h5ss0d
clonevar       blues_w5       =      h5ss0a


*************************************************************************************************
* Collapse the response categories of best_category_w5 and make them comparable to those at W3
* Therefore, the same syntax can be used for data at both Waves
*************************************************************************************************

recode best_category_w5 (3/7=3) (11/15=4) (21=5) (997=.)
label define best_category_w5 1 "(1) race-white" 2 "(2) race-black" 3 "(3)race-hispanic" 4 "(4)
race-asian" 5 "(5)race-american indians or Alaska Native"
label value best_category_w5 best_category_w5
tab1 best_category_w5, mis
tab2 h5od8 best_category_w5, mis



***********************
* Create the Age variable
***********************


tab1 Interview_y5 Interview_m5 birth_y5 birth_m5, mis
tab1 Interview_y5 Interview_m5 birth_y5 birth_m5, nol mis


gen cm_interview_w5 = (Interview_y5-1900)*12+Interview_m5
gen cm_birth_w5     = (birth_y5-1900)*12+birth_m5
```

```
label variable cm_interview_w5 "interview date at W5, century month"
label variable cm_birth_w5     "birth date from W5, century month"


tab1 cm_interview_w5 cm_birth_w5, mis


gen age_w5_c = (cm_interview_w5-cm_birth_w5)/12
gen age_w5_i = floor((cm_interview_w5-cm_birth_w5)/12)



label variable age_w5_c "respondent's age at w5, continuous"
label variable age_w5_i "respondent's age at w5, integer"

list aid Interview_y5 Interview_m5 birth_y5 birth_m5 age_w5_c age_w5_i in 1/50, nol
sum age_w5_i age_w5_c




**************************
* Respondent's gender at W5
**************************

tab1 female_w5, mis
tab1 female_w5, nol mis

replace female_w5 = female_w5 -1
label define female_w5 1 "(1) female" 0 "(0) male"
label value female_w5 female_w5
label variable female_w5 "respondent's sex at W5"

tab2 h5od2a female_w5, mis


******************************
* Respondent's race at W5
******************************


tab1 white_w5 black_w5 indian_w5 asian_w5 islander_w5 hispanic_w5, mis


tab2 asian_w5 islander_w5, mis

gen asian_islander_w5 = .
replace asian_islander_w5 =1 if asian_w5 ==1 | islander_w5 ==1
label variable asian_islander_w5 "RACE-Asian or Islanders-W5"
label value asian_islander_w5 yesno
tab1 asian_islander_w5, mis



gen str1 s_white_w5 ="."
gen str1 s_black_w5 ="."
gen str1 s_indian_w5 ="."
gen str2 s_asian_islander_w5 =".."
gen str1 s_hispanic_w5 ="."
gen str1 s_other_w5 ="."


replace s_white_w5 ="W"              if white_w5 ==1
replace s_black_w5 = "B"             if black_w5 ==1
replace s_indian_w5 ="I"             if indian_w5 ==1
replace s_asian_islander_w5 = "AI"   if asian_islander_w5 ==1
replace s_hispanic_w5 ="H"           if hispanic_w5 ==1
replace s_other_w5 ="O"              if other_w5 ==1
```

```
label variable s_white_w5          "RACE-WHITE-W5, string"
label variable s_black_w5          "RACE-BLACK/AFRICAN AM-W5, string"
label variable s_indian_w5         "RACE-AMER INDIAN/NATIVE Amerian-W5, string"
label variable s_asian_islander_w5 "RACE-ASIAN-ISlander-W5, string"
label variable s_hispanic_w5       "ARE YOU OF HISPANIC ORIGIN-W5, string"
label variable s_other_w5          "ARE YOU OF OTHER RACIAL BACKGROUNDS-W5, string"


gen race_pattern_w5 = s_white_w5 + "_" + s_black_w5 + "_" + s_indian_w5 + "_" +
s_asian_islander_w5 + "_" + s_hispanic_w5 + "_" + s_other_w5
label variable race_pattern_w5 "racial backgrounds at W5"

des race_pattern_w5
tab1 race_pattern_w5, mis



*********************************************************************
* Incorporate the information from the best race category variable
*********************************************************************

tab2 race_pattern_w5 best_category_w5, mis



gen race_w5 =.
label variable race_w5 "race at w5"
label define race_w5 1 "(1) white"  2 "(2) black"  3 "(3) indians" 4 "(4) asian/islanders" 5
"(5) hispanic"
label value race_w5 race_w5


*code respondents with only one racial backgrounds

replace race_w5 = 1 if race_pattern_w5 =="W_._._.._._."
replace race_w5 = 2 if race_pattern_w5 =="._B_._.._._."
replace race_w5 = 3 if race_pattern_w5 =="._._I_.._._."
replace race_w5 = 4 if race_pattern_w5 =="._._._AI_._."
replace race_w5 = 5 if race_pattern_w5 =="._._._.._H_."


*code hispanic background
replace race_w5 = 5 if race_pattern_w5 == "W_._._.._H_."
replace race_w5 = 5 if race_pattern_w5 == "W_._._AI_H_."
replace race_w5 = 5 if race_pattern_w5 == "W_._I_.._H_."
replace race_w5 = 5 if race_pattern_w5 == "W_B_._.._H_."
replace race_w5 = 5 if race_pattern_w5 == "W_B_I_.._H_."
replace race_w5 = 5 if race_pattern_w5 == "._B_._.._H_."
replace race_w5 = 5 if race_pattern_w5 == "._B_I_.._H_."
replace race_w5 = 5 if race_pattern_w5 == "._B_I_AI_H_."
replace race_w5 = 5 if race_pattern_w5 == "._._I_.._H_."
replace race_w5 = 5 if race_pattern_w5 == "._._I_.._H_O"
replace race_w5 = 5 if race_pattern_w5 == "._._._AI_H_."
replace race_w5 = 5 if race_pattern_w5 == "._._._.._H_O"

*code white background
replace race_w5 = 1 if race_pattern_w5 == "W_._I_.._._O" & best_category_w5 ==1
replace race_w5 = 1 if race_pattern_w5 == "W_B_I_.._._." & best_category_w5 ==1
replace race_w5 = 1 if race_pattern_w5 == "W_._._.._._O" & best_category_w5 ==1
replace race_w5 = 1 if race_pattern_w5 == "W_B_._.._._." & best_category_w5 ==1
replace race_w5 = 1 if race_pattern_w5 == "W_._._AI_._." & best_category_w5 ==1
replace race_w5 = 1 if race_pattern_w5 == "W_._I_.._._." & best_category_w5 ==1

*code black background

replace race_w5 = 2 if race_pattern_w5 == "W_B_I_AI_._." & best_category_w5 ==2
replace race_w5 = 2 if race_pattern_w5 == "W_B_._AI_._." & best_category_w5 ==2
replace race_w5 = 2 if race_pattern_w5 == "._B_._.._._O" & best_category_w5 ==2
replace race_w5 = 2 if race_pattern_w5 == "._B_._AI_._." & best_category_w5 ==2
```

```
replace race_w5 = 2 if race_pattern_w5 == "._B_I_._._._." & best_category_w5 ==2
replace race_w5 = 2 if race_pattern_w5 == "W_B_I_._._._." & best_category_w5 ==2
replace race_w5 = 2 if race_pattern_w5 == "W_B_._._._._." & best_category_w5 ==2

*code asian/islanders background

replace race_w5 = 4 if race_pattern_w5 == "._B_I_AI_._O" & best_category_w5 ==4
replace race_w5 = 4 if race_pattern_w5 == "W_B_._AI_._." & best_category_w5 ==4
replace race_w5 = 4 if race_pattern_w5 == "W_._._AI_._." & best_category_w5 ==4

*check the variable
tab2 race_pattern_w5 race_w5, mis



************************************************
* Depressed mood
************************************************

* check out the scale items

tab1 depressed_w5 sad_w5 blues_w5, mis
tab1 depressed_w5 sad_w5 blues_w5, mis nol


* check the inter-item associations of the depression mood items

factor depressed_w5 sad_w5 blues_w5
alpha depressed_w5 sad_w5 blues_w5


*three methods of aggregating the scale items

* the sum of scale items
gen d_score_w5_sum = depressed_w5+sad_w5+blues_w5

* the rowtotal function of scale items
egen d_score_w5_egen = rowtotal(depressed_w5 sad_w5 blues_w)

* the factor score of the scale items
factor depressed_w5 sad_w5 blues_w5
predict d_score_w5_fs

label variable d_score_w5_sum "depression score at W3, sum score"
label variable d_score_w5_egen "depression score at W3, egen function"
label variable d_score_w5_fs "depression score at W3, factor score"


sum d_score_w5_sum d_score_w5_egen d_score_w5_fs

list aid depressed_w5 sad_w5 blues_w5 d_score_w5_sum d_score_w5_egen d_score_w5_fs if
d_score_w5_sum ==., sepby(aid) nol


*********************************
* generate the indicator variable
*********************************

gen in_w5 =1
label variable in_w5 "respondents of W3"
label define in_w5 0 " (0) not in W3" 1 "(1) in W5"
label value in_w5 in_w5
tab1 in_w5, mis


*****************************
* Save the data
*****************************
sort aid
save "D:\jason\workshop\Stata workshop\w5.dta", replace
```

```
************************
* Merge data
************************

use "D:\jason\workshop\Stata workshop\w3.dta"
merge 1:1 aid using "D:\jason\workshop\Stata workshop\w4.dta"
tab1 in_w3 in_w4, mis
drop _merge

sort aid
count

merge 1:1 aid using "D:\jason\workshop\Stata workshop\w5.dta"

drop _merge

tab1 in_w3 in_w4 in_w5, mis

recode in_w3 (.= 0)
recode in_w4 (.= 0)
recode in_w5 (.= 0)

gen in_all = in_w3*100 + in_w4*10 +in_w5
label variable in_all "the waves that respondents participated"


*******************************************************
* Look at the number of survey participants across waves
*******************************************************
tab1 in_all, mis


****************************************
* Select the 3,147 eligible respondents
****************************************

keep if in_all ==111
count


***********************************************
* check the size of analytic sample
***********************************************

sum age_w*_i female_w* race_w* d_score_w*_fs


***********************************************
* Generate an indicator to select respondents
***********************************************


*************************************************************
* check the consistency in the gender variable across waves
*************************************************************

tab2 female_w3 female_w4, mis
tab2 female_w3 female_w5, mis

*******************************************************
* check the consistency in the race variable across waves
*106 respondents with inconsistent information on race
*******************************************************

tab2 race_w3 race_w5, mis

count if race_w3 ~= race_w5
```

```
*************************************************************************************************
***********
* check the consistencies of birthdate
* Many Birthdates at W5 differs from those at Waves 3 and 4
* Use Birthdates at Wave 3 to calculate ages at each wave
* If birthdates differ between Waves 3 and 4, use all available information to decide valid
birth dates
*************************************************************************************************
***********

list aid birth_y3 birth_m3 birth_y4 birth_m4 birth_y5 birth_m5 if cm_birth_w3 ~= cm_birth_w4,
nol
list aid birth_y3 birth_m3 birth_y4 birth_m4 birth_y5 birth_m5 if cm_birth_w4 ~= cm_birth_w5,
nol



gen birth_y =.
gen birth_m =.
label variable  birth_y "birth year, corrected"
label variable birth_m "birth month, corrected"

count if cm_birth_w3 == cm_birth_w4

replace birth_y = birth_y3 if cm_birth_w3 == cm_birth_w4
replace birth_m = birth_m3 if cm_birth_w3 == cm_birth_w4


list aid birth_y3 birth_m3 birth_y4 birth_m4 birth_y5 birth_m5 if cm_birth_w3 ~= cm_birth_w4,
nol


replace birth_y =. if aid =="91573099"
replace birth_m =. if aid =="91573099"

replace birth_y = birth_y3 if aid == "94574536"
replace birth_m = birth_m3 if aid == "94574536"

replace birth_y = birth_y4 if aid == "99575329"
replace birth_m = birth_m4 if aid == "99575329"


sum birth_y birth_m


gen ncm_birth = (birth_y -1900)*12 + birth_m
label variable ncm_birth "birth date, century month, corrected"


gen nage_w3_c = (cm_interview_w3-ncm_birth)/12
gen nage_w3_i = floor((cm_interview_w3-ncm_birth)/12)

gen nage_w4_c = (cm_interview_w4-ncm_birth)/12
gen nage_w4_i = floor((cm_interview_w4-ncm_birth)/12)

gen nage_w5_c = (cm_interview_w5-ncm_birth)/12
gen nage_w5_i = floor((cm_interview_w5-ncm_birth)/12)


label variable nage_w3_c "new age variable at W3, continuous)
label variable nage_w3_i "new age variable at W3, integer)
label variable nage_w4_c "new age variable at W4, continuous)
label variable nage_w4_i "new age variable at W4, integer)
label variable nage_w5_c "new age variable at W5, continuous)
label variable nage_w5_i "new age variable at W5, integer)

list aid Interview_y3 Interview_m3 cm_interview_w3 birth_y birth_m ncm_birth nage_w3_c nage_w3_i
if cm_birth_w3 ~= cm_birth_w4, nol
list aid Interview_y4 Interview_m4 cm_interview_w4 birth_y birth_m ncm_birth nage_w4_c nage_w4_i
if cm_birth_w3 ~= cm_birth_w4, nol
```

```
list aid Interview_y5 Interview_m5 cm_interview_w5 birth_y birth_m ncm_birth nage_w5_c nage_w5_i
if cm_birth_w3 ~= cm_birth_w5 in 1/100, nol


***************
*check the depressed mood variables
******************

sum d_score_w*_sum d_score_w*_egen d_score_w*_fs

*****************************
* Select variables
*****************************

keep aid d_score_w*_fs nage_w*_i race_w3 race_w5 female_w3
count


tab1 female_w3 race_w3 d_score_w3_fs d_score_w4_fs race_w5 d_score_w5_fs nage_w3_i nage_w4_i
nage_w5_i, mis
egen removed = rowmiss(female_w3 race_w3 d_score_w3_fs d_score_w4_fs race_w5 d_score_w5_fs
nage_w3_i nage_w4_i nage_w5_i)
label variable removed "respondents with missing values on one of the variables"
tab1 removed, mis


count if race_w3 ~= race_w5

keep if (race_w3 == race_w5)
keep if removed ==0


*************************
* final sample size is 3,002
*************************
count




********************
* select variables and Reshape the data
********************

keep aid female_w3 race_w3 d_score_w3_fs d_score_w4_fs d_score_w5_fs nage_w3_i nage_w4_i
nage_w5_i


tab1 d_score_w3, mis


*****************************************************
* Rename the variable names before reshaping the data
*****************************************************
rename female_w3              female
rename  race_w3               race
rename  d_score_w3_fs  d_score_w3
rename  d_score_w4_fs  d_score_w4
rename  d_score_w5_fs  d_score_w5
rename  nage_w3_i             nage_w3
rename  nage_w4_i             nage_w4
rename  nage_w5_i             nage_w5


reshape long d_score_w nage_w, i(aid) j(wave)
label variable wave "which wave the data were from"

rename d_score_w d_score_fs
rename nage_w    nage

label variable d_score_fs "depressed mood, factor score"
```

```
label variable nage       "age, use age st W3 to correct for errors at W5"
count

save "D:\jason\workshop\Stata workshop\w345.dta", replace
log close
```