\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\* End of Course Assignment - "Does depression predict life outcomes?", SIMONE CANTALUPI, 0001092400

\*\* This do file covers Part 0 the final assignment. Therefore, the purpose of this do file is to create a dataset by cleaning the orignal ones, and constructing and labelling new variables

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

clear all

// This path should be replaced with the parent directory with the add health data (the one containing DS001 etc subfolders)

global datapath "/Users/simonecantalupi/Desktop/Workshop/Final project/dataSets"

// This path should be replaced with the parent directory with the add health data (the one containing DS001 etc subfolders)

global projectpath "/Users/simonecantalupi/Desktop/Workshop/Final project/replication0001092400"

// copy this statement the lines below

if "${purpose}" == "grading" {

global datapath "$cd1"

global projectpath "$cd2"

}

use "$datapath/DS0001/21600-0001-Data.dta", clear

gen wave = 1

// IMPORTANT: DO NOT CHANGE THE WORKING DIRECTORY AFTER THIS POINT

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\* SECTION 1 \*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Biological sex \*

gen female = BIO\_SEX == 2 if BIO\_SEX < 6

label var female "Sex of the individual"

label define female 1 "Female" 0 "Male"

\* Race \*

gen white = 1 if H1GI6A == 1 & H1GI8 == 7

replace white = 0 if H1GI6A == 0

gen black = 1 if H1GI6B == 1 & H1GI8 == 7

replace black = 0 if H1GI6B == 0

gen hispanic = 1 if H1GI4 == 1 & H1GI8 == 7

replace hispanic = 0 if H1GI4 == 0

gen raceCategory = 1 if hispanic == 1

replace raceCategory = 2 if white == 1 & raceCategory == .

replace raceCategory = 3 if black == 1 & raceCategory == .

replace raceCategory = 4 if H1GI6A< = 1 & H1GI6B <= 1 & H1GI6C <= 1 & H1GI6D <= 1 & H1GI6E <= 1 & raceCategory == .

label define raceCategory 1 "Hispanic" 2 "White" 3 "Black" 4 "Other Ethnicity"

label values raceCategory raceCategory

label variable raceCategory "Race in 4 categories, derived from HIGI6A-HIGI6E,H1GI4, H1GI8"

label var raceCategory "Race of the individual"

tab raceCategory

\* Self-rated intelligence \*

gen intelligence = H1SE4 if H1SE4 < 96

label var intelligence "Self-rated intelligence (the higher the better)"

\* GPA \*

foreach var of varlist H1ED11 H1ED12 H1ED13 H1ED14 {

cap drop `var'\_new

recode `var' (1 = 4) (2 = 3) (3 = 2) (4 = 1) (5/99 = .), g(`var'\_new)

}

egen gpa = rowmean(H1ED11\_new H1ED12\_new H1ED13\_new H1ED14\_new)

label var gpa "GPA"

\* Depresssed \*

tab1 H1FS6 H1FS1 H1FS2 H1FS3 H1FS4 H1FS5 H1FS6 H1FS7 H1FS8 H1FS9 H1FS10 H1FS11 H1FS12 H1FS13 H1FS14 H1FS15 H1FS16 H1FS19

foreach var of varlist H1FS\* {

replace `var' = . if `var' == 6 | `var' == 8 | `var' == 9

}

// Some variables need reversing because are "positive"

foreach var of varlist H1FS4 H1FS8 H1FS11 H1FS15 {

replace `var' = 3 - `var'

}

// Depression index

egen depressionIndex = rowtotal(H1FS\*), missing

\*\* Replace as missing if more than 1 question is missing

egen missingQus = rowmiss(H1FS\*)

count if missingQus > 1

replace depressionIndex = . if missingQus > 1

replace depressionIndex = depressionIndex / 57

label var depressionIndex "% of possible score over 19 questions related to depression (see Fruewith et al, 2019) for details"

sum depressionIndex

\* Suicide \*

gen suicide = (H1SU4 == 1 | H1SU6 == 1)

replace suicide = . if H1SU4 > 1 | H1SU6 > 1

label var suicide "In the last 12 months have either a family member or a friend tried to commit suicide"

\* Household income \*

gen householdIncome = (PA55\*1000) if PA55 < 9996

label var householdIncome "Derived from PA55, in $'000s"

gen loghouseholdIncome = log(householdIncome+1)

label var loghouseholdIncome "Natural log of household income"

gen missingIncome = householdIncome == .

label var missingIncome "Derived from PA55; income not reported"

\* Age \*

\*\* We have information on birth year, and month.

tab H1GI1Y,nol /\*\* This looks problematic - the numbers are not years. Create a new var \*\*/

gen birthYear = H1GI1Y if H1GI1Y != 96

replace birthYear = birthYear+1900 if birthYear < .

label var birthYear "Birth Year (drved from H1GI1Y)"

gen ageAtSurvey = .

replace ageAtSurvey = 1994 - 1 - birthYear if IYEAR == 94

replace ageAtSurvey = 1995 - 1 - birthYear if (IYEAR == 95 & IMONTH <= H1GI1M & H1GI1M <=12)

replace ageAtSurvey = 1995 - birthYear if (IYEAR == 95 & IMONTH > H1GI1M & H1GI1M <= 12)

label var ageAtSurvey "Age at the time survey was taken. (Drved from H1GI1M H1GI1Y IMONTH IYEAR)"

sum ageAtSurvey

keep wave AID gpa householdIncome loghouseholdIncome female depressionIndex white black hispanic raceCategory ageAtSurvey suicide intelligence

save "$projectpath/datasets/wave1dataset.dta", replace

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\* SECTION 2 \*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

use "$datapath/DS0022/21600-0022-Data.dta", clear

gen wave = 4

\* Depression Index \*

tab1 H4MH18 H4MH19 H4MH20 H4MH21 H4MH22 H4MH23 H4MH24 H4MH25 H4MH26 H4MH27 H4MH28 H4MH2 H4MH3 H4MH4 H4MH5 H4MH6

replace H4MH3 = 3 if H4MH3 == 3 | H4MH3 == 4

replace H4MH4 = 3 if H4MH4 == 3 | H4MH4 == 4

replace H4MH5 = 3 if H4MH5 == 3 | H4MH5 == 4

replace H4MH6 = 3 if H4MH6 == 3 | H4MH6 == 4

foreach var of varlist H4MH20 H4MH24 H4MH25 H4MH4 H4MH5 {

replace `var' = 3 - `var'

}

// Depression index

egen depressionIndex = rowtotal(H4MH18 H4MH19 H4MH20 H4MH21 H4MH22 H4MH23 H4MH24 H4MH25 H4MH26 H4MH27 H4MH28 H4MH2 H4MH3 H4MH4 H4MH5 H4MH6), missing

label var depressionIndex "Depression Index (derived from H1FS1-H1FS19)"

\*\* Replace as missing if more than 1 question is missing

egen missingQus = rowmiss(H4MH18 H4MH19 H4MH20 H4MH21 H4MH22 H4MH23 H4MH24 H4MH25 H4MH26 H4MH27 H4MH28 H4MH2 H4MH3 H4MH4 H4MH5 H4MH6)

count if missingQus > 1

replace depressionIndex = . if missingQus > 1

gen depressed = depressionIndex >= 16 if depressionIndex < .

replace depressionIndex = depressionIndex / 48 /\*\*Make percentage of maximum so comparable to other years \*\*/

label var depressionIndex "% of possible score over 16 questions related to depression (see Fruewith et al, 2019) for details"

\* Age \*

gen birthYear = H4OD1Y

label var birthYear "Birth Year (drved from H1GI1Y)"

gen ageAtSurvey = .

replace ageAtSurvey = IYEAR4 - birthYear -- 1 if (IMONTH4 <= H4OD1M & H4OD1M <=12)

replace ageAtSurvey = IYEAR4 - birthYear

label var ageAtSurvey "Age at the time survey was taken. (Drved from H4OD1M H4OD1M IMONTH4 IYEAR4)"

\* Household Inocme and log Household Income \*

replace H4EC1= . if H4EC1== 96 | H4EC1== 98

gen householdIncome = . if H4EC1== .

replace householdIncome = 5000 if H4EC1 == 1

replace householdIncome = 7500 if H4EC1 == 2

replace householdIncome = 12500 if H4EC1 == 3

replace householdIncome = 17500 if H4EC1 == 4

replace householdIncome = 22500 if H4EC1 == 5

replace householdIncome = 27500 if H4EC1 == 6

replace householdIncome = 35000 if H4EC1 == 7

replace householdIncome = 45000 if H4EC1 == 8

replace householdIncome = 62500 if H4EC1 == 9

replace householdIncome = 87500 if H4EC1 == 10

replace householdIncome = 125000 if H4EC1 == 11

replace householdIncome = 150000 if H4EC1 == 12

label variable householdIncome "Derived from H4EC1"

gen loghouseholdIncome = log(householdIncome+1)

label var loghouseholdIncome "Natural log of household income"

keep wave AID depressionIndex depressed ageAtSurvey householdIncome loghouseholdIncome

save "$projectpath/datasets/wave4dataset.dta", replace

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\* SECTION 3 \*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

use "$datapath/DS0032/21600-0032-Data.dta", clear

gen wave = 5

\* Depression \*

tab1 H5SS0A H5SS0B H5SS0C H5SS0D H5MN5A H5MN1 H5MN2 H5MN3 H5MN4

replace H5MN1 = 4 if H5MN1 == 4 | H5MN1 == 5

replace H5MN2 = 4 if H5MN2 == 4 | H5MN2 == 5

replace H5MN3 = 4 if H5MN3 == 4 | H5MN3 == 5

replace H5MN4 = 4 if H5MN4 == 4 | H5MN4 == 5

foreach var of varlist H5SS0A H5SS0B H5SS0C H5SS0D H5MN5A H5MN1 H5MN2 H5MN3 H5MN4 {

replace `var' = `var' - 1

}

foreach var of varlist H5SS0C H5MN2 H5MN3 {

replace `var' = 3 - `var'

}

// Depression index

egen depressionIndex = rowtotal(H5SS0A H5SS0B H5SS0C H5SS0D H5MN5A H5MN1 H5MN2 H5MN3 H5MN4), missing

label var depressionIndex "Depression Index (derived from H1FS1-H1FS19)"

\*\* Replace as missing if more than 1 question is missing

egen missingQus = rowmiss(H5SS0A H5SS0B H5SS0C H5SS0D H5MN5A H5MN1 H5MN2 H5MN3 H5MN4)

count if missingQus > 1

replace depressionIndex = . if missingQus > 1

gen depressed = depressionIndex >= 16 if depressionIndex < .

replace depressionIndex = depressionIndex / 27 /\*\*Make percentage of maximum so comparable to other years \*\*/

label var depressionIndex "% of possible score over 9 questions related to depression (see Fruewith et al, 2019) for details"

sum depressionIndex

\* College degree \*

tab H5OD11

gen hasCollege = (H5OD11 >= 10)

replace hasCollege = . if missing(H5OD11)

tab hasCollege

label var hasCollege "The individual has a college degree"

label define hasCollege 1 "Has a college degree" 0 "Has not college degree"

\* Depression diagnosed \*

gen depressionDiagn = (H5ID6G == 1)

label var depressionDiagn "The individual has ever been diagnosed with depression"

label define depressionDiagn 1 "Has been diagnosed with depression" 0 "Has not been diagnosed with depression"

\* Medication for depression \*

gen depressionMed = (H5ID6GM == 1)

label var depressionMed "The individual is taking medication for depression (1 = Takes medication)"

label define depressionMed 1 "Takes medication" 0 "Takes not medication"

\* Age \*

gen birthYear = H5OD1Y

label var birthYear "Birth Year (drved from H5OD1Y)"

gen ageAtSurvey = .

replace ageAtSurvey = IYEAR5 - birthYear - 1 if (IMONTH5 <= H5OD1M | H5OD1M <=12)

replace ageAtSurvey = IYEAR5 - birthYear if (IMONTH5 > H5OD1M & H5OD1M <=12)

label var ageAtSurvey "Age at the time survey was taken. (Drved from H5OD1Y H5OD1Y IMONTH5 IYEAR5)"

\* Household Income \*

rename H5EC2 householdIncomeCat

label define H5EC2 1 "<5000" 2 "5000-9999" 3 "10000-14999" 4 "15000-19999" 5 "20000-24999" 6 "25000-29999" 7 "30000-39999" 8 "40000-49999" 9 "50000-74999" 10 "75000-99999" 11 "100000-149999" 12 "150000-199999" 13 ">200000"

label values householdIncomeCat H5EC2

replace householdIncomeCat = 0 if householdIncomeCat == 997 /\*\* These are legitimate skips, which appears to be because they answer 0 in the previous question \*\*/

replace householdIncomeCat = . if householdIncomeCat == 998 | householdIncomeCat == 999

label var householdIncomeCat "Household Income (Categorical), H5EC2"

gen householdIncome = 0 if householdIncomeCat == 0

replace householdIncome = 5000 if householdIncomeCat == 1

replace householdIncome = 7500 if householdIncomeCat == 2

replace householdIncome = 12500 if householdIncomeCat == 3

replace householdIncome = 17500 if householdIncomeCat == 4

replace householdIncome = 22500 if householdIncomeCat == 5

replace householdIncome = 27500 if householdIncomeCat == 6

replace householdIncome = 35000 if householdIncomeCat == 7

replace householdIncome = 45000 if householdIncomeCat == 8

replace householdIncome = 62500 if householdIncomeCat == 9

replace householdIncome = 87500 if householdIncomeCat == 10

replace householdIncome = 125000 if householdIncomeCat == 11

replace householdIncome = 175000 if householdIncomeCat == 12

replace householdIncome = 200000 if householdIncomeCat == 13

label variable householdIncome "Continuous household income, estimated from H5EC2"

sum householdIncomeCat

gen loghouseholdIncome = log(householdIncome+1)

label variable loghouseholdIncome "Log Household Income"

keep wave AID householdIncome householdIncomeCat loghouseholdIncome ageAtSurvey depressionIndex depressionMed depressionDiagn hasCollege

save "$projectpath/datasets/wave5dataset.dta", replace