```
************************************************************
****************** Solutions for Topic 5******************
************************************************************
global main_loc "/Users/gesun/Desktop/Bootcamp2023/05_Relational_Data/Class_Exercise"
global data "$main_loc/data"
global logfile "$main_loc/log_files"

clear
clear all
set more off
capture log close


log using "$logfile/topic5.log", replace

* Question A
*********************

*1. keep child_id, child_age, and all variables that store information about the
child's siblings (sib_*). After this step, you should have 112 variables in your
dataset.
*2. This survey asks ten questions for each sibling the child possibly has. (You can
order them first) Please reshape the database.

use "$data/baseline_student_raw.dta",clear

keep child_id child_age sib_*
des, short
order child_id child_age sib_gender* sib_age* sib_high_edu* sib_relation* sib_close*
sib_marital_status* sib_enrolled* sib_class* ///
sib_same_sch* sib_live_*

reshape long sib_gender sib_age sib_high_edu_level sib_relation sib_close
sib_marital_status sib_enrolled sib_class sib_same_sch sib_live_toge, i(child_id
child_age) j(sib_code)

* Question B
*********************

use "$data/lfs_examples_class08.dta",clear

* 1. Please make a graph that shows the relationship between (age-average) wages and
age. (Collapse)
preserve
gen wage_hr =  earnings_week/hours
collapse (mean) wage_hr, by (age)
label var wage_hr "hourly wage"
```

```
twoway scatter wage_hr age || lfit wage_hr age, scheme(s2mono)
graph export "$logfile/wage_age.pdf", as(pdf) name("Graph") replace
restore

* 2. Divide the sample into several 10-year age groups, like 25-35, 35-45, etc. You
need to choose the age dividing start point as the minimum age level in the database.
And then calculate each person's relative hourly wage to the average hourly wage in
their 10-year age group. Suppose one person's hourly wage is 15, and the average wage
in her age group is 20, then the relative wage is 15/20=0.75.
// one way to automatically generate the group number
sum age
scalar min_age = r(min)
scalar max_age = r(max)

gen group = .
replace group = floor((age-min_age)/10) + 1

// to generate the relative wage
gen wage_hr =  earnings_week/hours
bys group: egen wage_ave = mean(wage_hr)
gen relative_wage = wage_hr/wage_ave

* 2. After you get the (age-average) wages, save this after-collapse dataset, and merge
it with your original dataset. Calculate the relative wage with respect to the average
wage of this person's age group. What is the other way that you can directly calculate
the relative wage from the original database without “collapse” and “merge”?

use "$data/lfs_examples_class08.dta",clear

gen wage_hr =  earnings_week/hours
collapse (mean) aveage_wage = wage_hr, by (age)
cd "/Users/gesun/Desktop/Bootcamp/Exercise/Solutions/Day7"
save aveage_wage.dta,replace

cd "/Users/gesun/Desktop/Bootcamp/RCT_Examples"
use lfs_examples_class08.dta,clear
gen wage_hr =  earnings_week/hours
merge m:1 age using
"/Users/gesun/Desktop/Bootcamp/Exercise/Solutions/Day7/aveage_wage.dta"
gen relative_wage = wage_hr/aveage_wage

// Another way:
cd "/Users/gesun/Desktop/Bootcamp/RCT_Examples"
use lfs_examples_class08.dta,clear
gen wage_hr =  earnings_week/hours
bys age: egen aveage_wage = mean(wage_hr)
gen relative_wage = wage_hr/aveage_wage
```

```
* Question C
*********************

log using "$logfile/Day9.log", replace

global parent "$data/baseline_parent_raw.dta"
global student "$data/baseline_student_raw.dta"
global school "$data/baseline_school_cleaned.dta"

*1. import the "baseline_parent_raw.dta" into Stata. How many variables and how many
observations are in this dataset? Hint: try "des,short"
use "$parent",clear
des, short //  7,791 obs and 246 vars

* 2. import the "baseline_student_raw.dta" into Stata. How many variables and how many
observations are in this dataset?
use "$student",clear
des,short //  14,855 obs and 489 vars

* 3. merge the "baseline_parent_raw.dta" and "baseline_student_raw.dta" by the variable
"child_id". How many of them are merged? How many of the un-merged are from the parent
dataset? How many variables are left? Is there something wrong?
merge 1:1 child_id using "$parent" // (not working?)
// One option:
use "$student",clear
des team_id
tostring team_id,replace
merge 1:1 child_id using "$parent" // (another one)
// Another option:
merge 1:1 child_id using "$parent", force

des, short // 14,855 obs,  699 var

/* 4. Go back to the replication folder and find the do-file "0_master_run.do". Before
the merge, the authors have commands like:
    rename * P*
    rename Pchild_id child_id
Can you explain why they include these two lines of commands for the parent data? */
use "$parent",clear

// sib_age*

// rename age Page
// rename gender Pgender

rename * P*
rename Pchild_id child_id
```

```
* 5. Merge the two datasets again, and for this time, please make sure the final
variable number = variables from student + variables from parents - 1
merge 1:1 child_id using "$student"
des, short //  14,855 obs, 735 vars = 489 + 246 - 1 + 1 (_merge)

* 6. Continue to merge this dataset with "baseline_school_cleaned.dta" by the link of
variable
use "$school",clear
des,short  // obs 314, var:313
rename School_ID school_id
tostring school_id,replace

rename * S*
rename Sschool_id school_id

merge 1:m school_id using "$student"
des,short // obs 14855, var: 802 = 313 + 489 - 1 + 1

capture log close
```