```
*************************************************************
******************** Solutions for Topic 2*******************
*************************************************************

***************** Question 2.1 & 2.2 ************************
*************************************************************

// Basic Setting
clear
clear all
set more off
capture log close

cd "/Users/gesun/Desktop/Bootcamp2023/02_Intro_to_Stata/Class_Exercise"

log using "./log_files/topic2_Q1.log", replace


// Coding part

* 0. Import data set "*baseline_census_cleaned.dta*" from the "Analysis Data" subfolder
to Stata.

// Absolute paths
// use "/Users/gesun/Desktop/Bootcamp/Exercise/Solutions/Main Analysis and
Paper/Analysis data/baseline_census_cleaned.dta",clear

// cd "/Users/gesun/Desktop/Bootcamp/RCT_Examples/Main Analysis and Paper/Analysis
Data"
// use baseline_census_cleaned.dta,clear

// Relative paths:
// cd "/Users/gesun/Desktop/Bootcamp/RCT_Examples/Main Analysis and Paper"
// use "./Analysis Data/baseline_census_cleaned.dta", clear
use "$data/baseline_census_cleaned.dta",clear

* 1. Use a command to get the number of observations and the number of variables in
this dataset.
des

* 2. It is a large dataset, and we do not need to use that many variables for today's
practice. Just keep the following the six variables: Ctot_p, Cp_lit, Cp_ill, treatment,
tru, subdistt

* 3. What is each of the six variables' label? According to their labels, can you
understand what these six variables are?
```

```
keep Ctot_p Cp_lit Cp_ill treatment tru district districtname subdistt // Ctot_p is the population; Cp_lit is the number of literates in each village/school/observation; Cp_ill is the nmber of illiterates; treatment indicates whether this observation is in the treatment/control group; tru indicates urban/rural areas; subdistt is the indicator for the subdistrict.

* 4. How many subdistricts are in this dataset?
codebook subdistt
tab subdistt

* 5. How many observations are in the treatment group? Which variable gives you this information?
tab treatment

* 6. How many people are in the **most** populated village?
sum Ctot_p

* 7. Generate a variable "ill_rate" that shows the illiteracy rate in each village, and label it "the illiteracy rate in each village".
gen ill_rate = Cp_ill/Ctot_p
label var ill_rate "the illiteracy rate in each village"

* 8. In "subdistt" 00379, how many villages in this sample are located in urban areas?
tab tru subdistt

* 9. What is the data type of variable "subdistt"? Convert it to the other data type.
des subdistt
destring subdistt, replace

* 10. Can you check this data by adding variable "Cp_lit" and variable "Cp_ill", and then see whether the sum is equal to variable "Ctot_p"? Are they should be equal?
gen check = Cp_ill + Cp_lit
gen diff = check - Ctot_p
sum diff

capture log close



****************** Question 2.3 **************************
*********************************************************



// Basic Setting
clear
clear all
set more off
capture log close
```

```
cd "/Users/gesun/Desktop/Bootcamp2023/02_Intro_to_Stata/Class_Exercise"

log using "./log_files/topic2_Q3", replace

// Coding part

use "./data/baseline_census_cleaned.dta", clear

* 1. Import the data to Stata and find the two variables that relate to the information
of "district".
lookfor district

* 2. Keep the six Ctot_p, Cp_lit, Cp_ill, treatment, tru, subdistt variables, AND the
two new district variables in this dataset.
keep Ctot_p Cp_lit Cp_ill treatment tru subdistt  districtname district

* 3. In this dataset, how many sub-districts are in each district?
tab subdistt district
bys district: egen count_sub_districts = count(subdistt)
tab count_sub_districts district

sort district subdistt
// keep if subdistt[_n] != subdistt[_n+1]

* 4. What are the average village population numbers in treatment and control groups?
bys treatment: egen avg_village_pop = mean(Ctot_p)


* 5. Please generate a series of dummy variables (locate) according to the value of
"subdistt". How many "locate" variables do you generate?
tab subdistt, gen(locate)


* 6. Use ONE line of command to summarize those "locate" variables separately.
sum locate*


* 7. This dataset provides us with the name of each district. Please attach the
district name to the district index through label define/label value (there is a bit of
inconsistency that one district code corresponds to more than one district name. Use
the main one for exercise purposes. But if you encounter the same problem during actual
research, you need to find out why does this situation happen and then decide what to
do)

tab districtname district

destring district, replace
label define distrct_namel 75 "Panipat"  82 "Rohtak"
label values district district_namel
```

```
log close



******************* Question 2.4 ***************************
************************************************************


global main_loc "/Users/gesun/Desktop/Bootcamp2023/02_Intro_to_Stata/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/topic2_Q4.log", replace


* PART A
*********************

use "$data/lfs_examples_class08.dta",clear
keep UqNr age earnings_week black female hours Indus_Sep2003

* 1. Find relevant variables to calculate the hourly wage variable wage_hr
gen wage_hr =  earnings_week/hours

* 2. Generate a histogram of wage_hr
hist wage_hr, scheme(economist)

* 3. Generate a new variable that equals the natural log value of wage_hr: ln_wage_hr
gen ln_wage_hr = log(wage_hr)

* 4. Generate a histogram of ln_wage_hr
hist ln_wage_hr, scheme(tab2)

* 5. Compare the number of variables in these two histograms. Are they the same?
sum wage_hr
return list
scalar wage_num = r(N)
sum ln_wage_hr
return list
scalar lnwage_num = r(N)
```

```
dis wage_num == lnwage_num // Thus the number of observations for these two histograms
are the same. Since all observations in this dataset have positive (non-zero) wages,
taking log will not generate missing values. But it is a good habit to check whether
there are missing values and why when generating new variables.

* 6. Compare those two histograms. Can you guess why we usually take log to variables
like wages?
/* the histogram of wage_hr is right-skewed, just same as other wage distributions.
Very few people earn very large amount of money. This extreme big values (or we can
call it outliers) can have strong influence on regression results. Therefore, a normal
way to take care of skewed wage distribution is to take logs to it. As you can see, the
distribution for ln_wage_hr is quite normal.
Another benefit for taking log on wages is: this gives the regression results good
economic meanings. We will get this back when talking about regression. */

* 7. Can you find the average hourly wage level for black female workers in this
sample?
sum wage_hr if black == 1 & female == 1

* 8. Can you find the average hourly wage level for male workers above-average age?
(Hint: use scalar to store the mean value for the judgment in the next step)
sum age if female == 0
scalar ave_male = r(mean)
sum wage_hr if female == 0 & age > ave_male
sum wage_hr if female == 0 & age > r(mean)



* PART B
*********************

*What does the macro "main_loc" represent? Can you explain this line of code: "global
deid "$main_loc/Analysis data"?
/* “main_loc" represents the main path for this replication, and in my computer, it is
"/Users/gesun/Desktop/Bootcamp2023/Exercise".

"global deid "$main_loc/Main Analysis and Paper/Analysis data" means define a new
global macro "deid", and "deid" represents "$main_loc/Main Analysis and Paper/Analysis
data", or "/Users/gesun/Desktop/Bootcamp2023/Exercise/Main Analysis and Paper/Analysis
data"
*/



* PART C
*********************

// import  "baseline_student_raw.dta" into Stata
use "$data/baseline_student_raw.dta",clear

* 1. Which variable tells you the student's age? How do you find it?
```

```
lookfor age // thus the variable child_age shows student's age

* 2. Find what ".s" ".d" means in the data. What will you do if you want to exclude all missing values?
help missing
/*
according to help document:
.a, .b, .c, ..., .z, which are called the "extended missing values". And Numeric missing values are represented by large positive values.  The ordering is
all nonmissing numbers < . < .a < .b < ... < .z
Therefore, to exclude all the missing values, I need to include a conditional expression: var < .
*/

* 3. There are eleven sib_age* variables asking their siblings' ages. Create a loop using "forvalues" to sum these variables. There are eleven variables for each sibling-related question. Can you tell me why and provide some data evidence?

forvalues i = 1/11{
    sum sib_age`i'
}

forvalues i = 1/11{
    sum sib_gender`i'
}
// the zero observation from this sib_age11 sum results explain why we only have 11 sibling-related variables.

* 4. Can you count how many siblings each student has?
gen sib_count = 0
forvalues i = 1/11{
    replace sib_count = sib_count + 1 if sib_gender`i' <.
}

// Or:

gen sib_count2 = 0
forvalues i = 1/11{
    replace sib_count2 = sib_count2 + 1 if !mi(sib_gender`i')
}

order sib_gender* sib_count
list sib_gender* sib_count in 1/10 // It is very important to double-check your result even the code could run!!!
tab sib_count // It is different from what we get from tab sib_count, why?


* PART D
*********************
```

```
use "$data/baseline_student_raw.dta",clear

* 1. Can you generate a variable that represents the age of each student's youngest married siblings (marital status ==1)? Hint: there are many ways to achieve this goal, and you might need to Google some new commands to realize your ideas.

order sib_age* sib_marital_status*
keep sib_age1 sib_age2 sib_age3 sib_marital_status1 sib_marital_status2 sib_marital_status3
// gen finding_young_married1 = 999
// local i=1
// replace finding_young_married1 = sib_age`i' if sib_age`i' < finding_young_married1 & sib_marital_status`i' ==1

gen finding_young_married1 = 999
forvalues i = 1/3 {
replace finding_young_married1 = sib_age`i' if sib_age`i' < finding_young_married1 & sib_marital_status`i' ==1
}
//Question: do I need to include if !mi(sib_age`i')?
replace finding_young_married1 =. if finding_young_married1 == 999
tab finding_young_married1

// Another possible answer
* generate young_married_`i' to store married sibling's age
forvalues i = 1/3{
    gen young_married`i' = .
}

set trace on
forvalues i = 1/3{
    replace young_married`i' = sib_age`i' if sib_age`i' < . & sib_marital_status`i' ==1
}
* find the youngest married sibling
egen finding_young_married2 = rowmin(young_married*)

// Double checking
dis finding_young_married1  == finding_young_married2

* 2. Get to know the label values for the variables sib_high_edu_level*.
codebook sib_high_edu_level*
label list
```

```
* 3. For families with two children, what's the average highest education level for
these two children(for now, just to take an average of the sib_high_edu_level*)? How
about families with five kids? or six kids? (Let's assume each family has only one
child represented by child_id entering this dataset, in other words, each child_id can
uniquely represent a family. Otherwise, the sibling to child No.1101002 might be the
child No.1101006) Hint: these may require data type conversion. Try "des" if you need
it.
gen sib_count = 0
forvalues i = 1/11{
    replace sib_count = sib_count + 1 if sib_gender`i' <.
}

forvalues i = 1/10 {
    gen sib_edu`i' = 0
    replace sib_edu`i' = sib_high_edu_level`i' if sib_high_edu_level`i' <.
    replace sib_edu`i' = 0 if sib_edu`i' ==12
    replace sib_edu`i' = 0 if sib_edu`i' == 11
}

// gen edu_all = .
gen edu_all = int(0)
forvalues i = 1/10 {
    replace edu_all = sib_edu`i' + edu_all
    tab edu_all
}

gen sib_ave_edu = edu_all/sib_count
bys sib_count: egen sib_edu_by_family = mean(sib_ave_edu)


* 4. What are the problems if I use the result in the last question arguing children in
large families tend to have more/less education?
/*
a. It does not make sense to take mathematical operations to categorical variables.
b. Many children's siblings are just at a young age and have not finished their
schooling years.
*/

* 5.  Pick any other two variables related to their attitudes to gender equality or
their living standard that you are interested in. Please tell me what the data tells
you.

capture log close
```