```
* Set the working directory
cd "~/Desktop/Replication_Package"

* Step 1: Load the public school dataset and perform the
aggregation by ZIP code
use "public_school_ny.dta", clear

* Define school categories
gen elementary_school = SCHOOL_LEVEL == "Elementary"
gen middle_school = SCHOOL_LEVEL == "Middle"
gen high_school = SCHOOL_LEVEL == "High"
gen other_school = inlist(SCHOOL_LEVEL, "Other", "Prekindergarten",
 "Secondary", "Ungraded", "Not reported")

* Aggregate by ZIP code
collapse (sum) elementary_schools=elementary_school ///
         (sum) middle_schools=middle_school ///
         (sum) high_schools=high_school ///
         (sum) other_schools=other_school ///
         (mean) avg_student_teacher_ratio=STUTERATIO ///
         (sum) total_free_lunch=TOTFRL, by(LZIP)

* Calculate total schools and rename ZIP code
gen total_schools = elementary_schools + middle_schools +
high_schools + other_schools
rename LZIP zip_code
save "Public_School_Aggregated.dta", replace

* Step 2: Merge with the housing data
use "NY_Housing.dta", clear
merge m:1 zip_code using "Public_School_Aggregated.dta"

* Keep only matched observations
drop if _merge != 3
drop _merge

* Save the merged dataset
save "Housing_Schools_Merged.dta", replace

* Step 3: Load the income dataset
use "Income.dta", clear
duplicates drop zip_code, force
save "Income.dta", replace

* Reload the housing dataset
use "Housing_Schools_Merged.dta", clear
merge m:1 zip_code using "Income.dta"

```

```
* Check merge results
drop if _merge != 3
drop _merge

* Save the final merged dataset
save "Final_Merged_Data.dta", replace

* Step 4: Load the final merged dataset
use "Final_Merged_Data.dta", clear

* Step 5: Clean and Filter the Dataset
keep price bed bath acre_lot house_size zip_code total_free_lunch
total_schools costofliving medianincome avg_student_teacher_ratio
drop if price < 60000 | price > 20000000
drop if bed > 15
drop if bath > 12
drop if house_size > 10000 | house_size < 100
drop if acre_lot > 60
drop if avg_student_teacher_ratio > 20
drop if total_schools > 35
drop if missing(acre_lot)
drop if missing(price) | missing(bed) | missing(bath) | missing(
house_size) | missing(total_free_lunch) | missing(
avg_student_teacher_ratio)

* Step 6: Prepare Variables for Analysis

* Drop unused variables, and log-transform key variables
gen log_price = log(price)
gen log_house_size = log(house_size)
gen log_acre_lot = log(acre_lot)
drop if missing(costofliving) | missing(medianincome) | missing(
log_acre_lot)
gen log_total_free_lunch = log(total_free_lunch)
drop if missing(log_total_free_lunch)

* Step 7: Convert String Variables to Numeric

* Convert costofliving from string to numeric
generate costofliving_num = real(subinstr(subinstr(costofliving,
"$", "", .), ",", "", .))
drop costofliving
rename costofliving_num costofliving
format costofliving %12.2f

* Convert medianincome from string to numeric
generate medianincome_num = real(subinstr(subinstr(medianincome,
"$", "", .), ",", "", .))
```

```
drop medianincome
rename medianincome_num medianincome
format medianincome %12.2f

* Step 8: Standardize Variables

* Standardize house_size
capture drop std_house_size
summarize house_size, detail
local house_size_mean = r(mean)
local house_size_sd = r(sd)
gen std_house_size = (house_size - `house_size_mean') /
`house_size_sd'

* Standardize avg_student_teacher_ratio
capture drop std_avg_student_teacher_ratio
summarize avg_student_teacher_ratio, detail
local avg_student_teacher_ratio_mean = r(mean)
local avg_student_teacher_ratio_sd = r(sd)
gen std_avg_student_teacher_ratio = (avg_student_teacher_ratio -
`avg_student_teacher_ratio_mean') / `avg_student_teacher_ratio_sd'

* Standardize acre_lot
capture drop std_acre_lot
summarize acre_lot, detail
local acre_lot_mean = r(mean)
local acre_lot_sd = r(sd)
gen std_acre_lot = (acre_lot - `acre_lot_mean') / `acre_lot_sd'

* Standardize medianincome
capture drop std_medianincome
summarize medianincome, detail
local medianincome_mean = r(mean)
local medianincome_sd = r(sd)
gen std_medianincome = (medianincome - `medianincome_mean') /
`medianincome_sd'

* Standardize costofliving
capture drop std_costofliving
summarize costofliving, detail
local costofliving_mean = r(mean)
local costofliving_sd = r(sd)
gen std_costofliving = (costofliving - `costofliving_mean') /
`costofliving_sd'

* Step 9: Save the cleaned dataset
save "Final_Merged_Data_Cleaned.dta", replace

```

```
* Step 10: Summary Statistics
summarize price bed bath acre_lot house_size total_free_lunch
total_schools costofliving medianincome avg_student_teacher_ratio

* Step 11: Figures

* Figure 1: Distribution of Log-Transformed Housing Prices
histogram log_price, normal
graph export "figure1.png", as(png) replace

* Figure 2: Residuals for Model 1
regress log_price bed
predict residuals1, resid
histogram residuals1, normal
graph export "figure2.png", as(png) replace

* Figure 3: Residuals for Model 7
regress log_price bed bath c.log_house_size##c.log_acre_lot
total_schools std_avg_student_teacher_ratio log_total_free_lunch
std_medianincome std_costofliving zip_code
predict residuals7, resid
histogram residuals7, normal
graph export "figure3.png", as(png) replace

* Step 12: Run Regression Models

* Model 1: Baseline Model - Bedrooms
regress log_price bed
est store model1

* Model 2: Add Housing Characteristics
regress log_price bed bath std_house_size  std_acre_lot
est store model2

* Model 3: Replace with Log-Transformed House Size and Lot Size
(r-squared improves)
regress log_price bed bath log_house_size  log_acre_lot
est store model3

* Model 4: Add Total Schools
regress log_price bed bath log_house_size  log_acre_lot
total_schools
est store model4

* Model 5: Add Student-Teacher Ratio and Free Lunch
regress log_price bed bath log_house_size  log_acre_lot
total_schools std_avg_student_teacher_ratio log_total_free_lunch
est store model5
```

```
* Model 6: Add Median Income
regress log_price bed bath log_house_size  log_acre_lot
total_schools std_avg_student_teacher_ratio log_total_free_lunch
std_medianincome
est store model6

* Model 7: Add Cost of Living and Interaction
regress log_price bed bath c.log_house_size##c.log_acre_lot
total_schools std_avg_student_teacher_ratio log_total_free_lunch
std_medianincome std_costofliving zip_code
est store model7

esttab model1 model2 model3 model4 model5 model6 model7 using
regression_table.tex, replace ///
label b(3) se stats(r2 N) compress

```