**Instructions:**

1. Each question is worth 20 points (including the bonus question meaning the best possible score is 120/100)
2. Please submit it through canvas
3. Assignment won’t be graded if submitted after the deadline
   1. You also will not be able to do Stata peer review 1

**Q1**: As part of a larger examination of how various factors contribute to student achievement, you have been asked to find a couple of pieces of information about a school district. Unfortunately, the relevant data is spread across four different files (student.dta, teacher.dta, school.dta, and subject.dta all in the following subfolder: q1\_data. See the readme file for more details regarding each dataset.

1. What is the mean attendance of students at southern schools?



The mean attendance of students at southern schools is 177.48 days out of the school year (assuming it’s days, there’s no additional information).  
(b) Of all students in high school, what proportion of them have a primary teacher who teaches a tested subject?



54.03 percent of high school students have a primary teacher who teaches a tested subject.  
(c) What is the mean gpa of all students in the district?



The mean gpa for all students in the district is 3.6.  
(d) What is the mean attendance of each middle school?



**Q2:** You are working on a crop insurance project in Kenya. For each household, we have the following information: village name, pixel and payout status.

1. Payout variable should be consistent within a pixel, confirm if that is the case. Create a new dummy variable (pixel\_consistent), this variable =0 if payout variable isn’t consistent within that pixel (i.e. =1 when all the payouts are exactly the same, =0 if there is even a single different payout in the pixel)

 Payouts look uniform within pixels.



1. Usually the households in a particular village are within the same pixel but it is possible that some villages are in multiple pixels (boundary cases). Create a new dummy variable (pixel\_village), =0 for the entire village when all the households from the village are within a particular pixel, =1 if households from a particular village are in more than 1 pixel. Hint: This variable is at village level.

bys village pixel: gen tag = \_n == 1

bys village: gen n\_pixels = sum(tag)

gen pixel\_village = (n\_pixels > 1)

codebook pixel\_village

tab vill if pixel\_village == 1



1. For this experiment, it is only an issue if villages are in different pixels AND have different payout status. For this purpose, divide the households in the following three categories:
   * 1. Villages that are entirely in a particular pixel. (==1)
     2. Villages that are in different pixels AND have same payout status (Create a list of all hhids in such villages) (==2)
     3. Villages that are in different pixels AND have different payout status (==3)

Hint: These 3 categories are mutually exclusive AND exhaustive i.e. every single observation should fall in one of the 3 categories. Note also that the categories may or may not line up with what you created in (a) and (b) so read the instructions closely.

gen hh\_status = .

// hh if village entirely in pixel

replace hh\_status = 1 if pixel\_village == 0

// hh if village is in > 1 pixel & same payout

bys village payout: gen pay\_tag = \_n == 1

bys village: egen n\_payout = sum(pay\_tag)

gen payout\_consistent = (n\_payout == 1) // 1: 1 payout 0: > 1 payout

replace hh\_status = 2 if pixel\_village == 1 & payout\_consistent == 1

// hh if village > 1 pixel & different payout

replace hh\_status = 3 if pixel\_village == 1 & payout\_consistent == 0

codebook hh\_



**Q3:** Faculty members submitted 128 proposals for funding opportunities. Unfortunately, we only have enough funding for 50 grants. Each proposal was assigned randomly to three selected reviewers who each gave a score between 1 (lowest) and 5 (highest). Each person reviewed 24 proposals and assigned a score. **We think it will be better if we normalize the score wrt each reviewer (using unique ids) before calculating the average score**. Add the following columns 1) stand\_r1\_score 2) stand\_r2\_score 3) stand\_r3\_score 4) average\_stand\_score 5) rank (Note: highest score =>1, lowest => 128)

Hint: We can normalize scores using the following formula: (score – mean)/sd, where mean = mean score of that particular reviewer (based on the netid), sd = standard deviation of scores of that particular reviewer (based on that netid). (Hint: we are not standardizing the score wrt reviewer 1, 2 or 3. But by the netID.)

use "$wd/week\_03/04\_assignment/01\_data/q3\_proposal\_review.dta", clear

// renaming vars

rename proposal\_id prop\_id

rename \*, lower // making all variables lower case

rename rewiewer1 reviewer1 // fixing typo

rename averagescore avg\_score

rename standarddeviation stdev

local n = 1

foreach var in review1score reviewer2score reviewer3score {

rename `var' score`n'

local n = `n' + 1

}

// dropping unnecessary variables

drop pi department

// reshaping to long format

reshape long score reviewer, i(prop\_id) j(reviewer\_num)

// adding columns

foreach n in 1 2 3{

bys reviewer: egen mean\_score`n' = mean(score)

gen stand\_r`n'\_score = (mean\_score`n' - avg\_score) / stdev

}

gen average\_stand\_score = (mean\_score1 + mean\_score2 + mean\_score3) / 3

gsort -average\_stand\_score

gen rank = \_n

list rank prop\_id reviewer score if \_n <= 10



**Q4:** We have the information of adults that have computerized national ID card in the following pdf: Pakistan\_district\_table21.pdf. This pdf has 135 tables (one for each district). We extracted data through an OCR software but unfortunately it wasn’t very accurate. We need to extract column 2-13 from the first row (“18 and above”) from each table. Create a dataset where each row contains information for a particular district. The hint do file contains the code to loop through each sheet, you need to find a way to align the columns correctly.

Hint: While the formatting is mostly regular, there are a couple of (pretty minor) anomalies so be sure to look at what your code produces.

I tried my best to remove missing values by trimming, keeping only relevant columns, but I am not sure where else to go from here.

global excel\_t21 "$wd/week\_03/04\_assignment/01\_data/q4\_Pakistan\_district\_table21.xlsx"

import excel "$excel\_t21", clear

clear

\*setting up an empty tempfile

tempfile table21

save `table21', replace emptyok

\*Run a loop through all the excel sheets (135) this will take 1-5 mins because it has to import all 135 sheets, one by one

forvalues i=1/135 {

import excel "$excel\_t21", sheet("Table `i'") firstrow clear allstring //import

display as error `i' //display the loop number

keep if regexm(TABLE21PAKISTANICITIZEN1, "18 AND" )==1 //keep only those rows that have "18 AND"

\*I'm using regex because the following code won't work if there are any trailing/leading blanks

\*keep if TABLE21PAKISTANICITIZEN1== "18 AND"

keep in 1 //there are 3 of them, but we want the first one

rename TABLE21PAKISTANICITIZEN1 table21

keep B C D E F G H I J K L M // keeping relevant columns

foreach var of varlist B C D E F G H I J K L M {

replace `var' = trim(`var')

}

gen table=`i' //to keep track of the sheet we imported the data from

append using `table21'

save `table21', replace //saving the tempfile so that we don't lose any data

}

\*load the tempfile

use `table21', clear

destring B C D E F G H I J K L M, replace ignore("N/A")

list B C D E F G H I J K L M in 1/10



**Q5:** This task involves string cleaning and data wrangling. We scraped data for a school from a [Tanzanian government website](https://onlinesys.necta.go.tz/results/2021/psle/results/shl_ps0101114.htm). Unfortunately, the formatting of the data is a mess. Your task is to extract the following school level variables:

1) number of students that took the test,

2) school average

3) student group (binary, either under 40 or >=40

4) school ranking in council (22 out of 46)

5) school ranking in the region (74 out of 290)

6) school ranking at the national level (545 out of 5664) level dataset with the following variables.

In addition to these variables, also capture the school name and school code in two different columns. Note: This is a school level dataset, and should only contain one row with all the variables. All the school level information is given at the top of this [webpage](https://onlinesys.necta.go.tz/results/2021/psle/results/shl_ps0101114.htm). The page is in Swahili but it should be fairly straightforward to find the relevant information. You can use google translate if you have trouble finding the relevant parts of the webpage.

use "$wd/week\_03/04\_assignment/01\_data/q5\_Tz\_student\_roster\_html.dta", clear

// extracting html\_text

gen html\_text = s

// extracting school name

gen school\_name = regexs(1) if regexm(html\_text, "([A-Z ]+ PRIMARY SCHOOL)")

// extract school code

gen school\_code = regexs(1) if regexm(html\_text, "(PS[0-9]+)")

// extract number of students

gen num\_students = real(regexs(1)) if regexm(html\_text, "WALIOFANYA MTIHANI : ([0-9]+)")

// extracting school average

. gen school\_avg = real(regexs(1)) if regexm(html\_text, "WASTANI WA SHULE : ([0-9]+\.[0-9]+)")

// extract student group

gen student\_group = "Under 40" if regexm(html\_text, "KUNDI LA SHULE : Wanafunzi chini ya 40")

replace student\_group = ">=40" if student\_group == ""

// extracting SR in council

gen rank\_council = regexs(1) + " out of " + regexs(2) if regexm(html\_text, "NAFASI YA SHULE KWENYE KUNDI LAKE KIHALMASHAURI: ([0-9]+) kati ya ([0-9]+)")

// extracting SR in region

gen rank\_region = regexs(1) + " out of " + regexs(2) if regexm(html\_text, "NAFASI YA SHULE KWENYE KUNDI LAKE KIMKOA : ([0-9]+) kati ya ([0-9]+)")

// extract national SR

gen rank\_national = regexs(1) + " out of " + regexs(2) if regexm(html\_text, "NAFASI YA SHULE KWENY E KUNDI LAKE KITAIFA : ([0-9]+) kati ya ([0-9]+)")

keep school\_name school\_code num\_students school\_avg student\_group rank\*



**Bonus Question**: This task involves string cleaning and data wrangling. We scrapped student data for a school from a [Tanzanian government website](https://onlinesys.necta.go.tz/results/2021/psle/results/shl_ps0101114.htm). Unfortunately, the formatting of the data is a mess. Your task is to create a student level dataset with the following variables: schoolcode, cand\_id, gender, prem\_number, name, grade variables for: Kiswahili, English, maarifa, hisabati, science, uraia, average. Note: This is a school level dataset, and should have 16 rows (same as the number of students in that school).

Hint: you can get a better view of the string if you go to the website and view its source (which can be done by right clicking or hitting ctrl/command+U).