The raw data is in “Data\weekdayend_all.xlsx”. The survey is still ongoing but will be finalized over next weekend (around June 6). All variables can be put into three main categories (all the shorthand notations can be found in the questionnaire):

(1) Background information for workers

Demographics: DM (morning + evening) JA (morning) TR (morning) WF (evening) JH (evening)

Type: AT (morning + evening) WM (morning) A11 – A36 (grading) B1 – B12 (grading) dexterity (grading)

(2) Information of jobs

Outside options: OA – OC (morning)

Jobs in HIP: IA – ID (morning), IT (morning)

Jobs in HIP, follow-up: IB – ID (weekend), IA (weekend)

(3) Additional question, less of the current interest: WA (evening) PO (evening)

We can skip (3) for the moment. Ideally, Ezana would be in charge of (1) because there might be some open-ended questions with Amharic typed there, and Max would be in charge of (2). Both groups of variables are of equal importance: we want to see if good “type” of workers and workers with higher misperceptions are more likely to be affected by the information.

Here’s my usual way to conduct variable cleaning:

a. I would look at the distribution of the values first. (In Stata I would simply do “tabulate X”.) If it’s a continuous variable, you can see if there’s some weirdly large or small values. If it’s a 0-1 or order variable, you can first check the codes for values in the surveyCTO questionnaire (sheet “Choice”).
b. Dealing with outliers: For weird values in continuous variables, if it’s an obvious mistake (for example, some enumerator might type “2000001” instead of “2000”), you can simply write a command to replace the value. If it’s not an obvious mistake but the values are too large in a sense, you can trim the outliers or winsorize the variable.
c. Dealing with order variables: Some questions might allow values like “-7”, “-8”, or “-9” to indicate “I don’t know”, “I’m not sure”, or “not applicable”. Usually we treat “not applicable” as missing, but for “I don’t know” or “I’m not sure” it depends. Please write down your reasoning whenever you’re dealing with situations like this (whether and why you treat the negative values as missing or 0).
d. Some variables should be cleaned in a group. For example, there’s a minimal salary question “Would you accept a job if paid 600/700/…/2000 birr a month?” Eventually we want one continuous variable that indicates worker’s minimal salary.
e. String variables: when dealing with places, names, or other open-ended questions, I would first see if a thorough cleaning of this variable is necessary. For example, if in the future we need to use information of a specific woreda, we might need to clean the woreda names (and there’ll be a lot of typos). If not, then we should think what useful information we can get out of this variable without thorough cleaning.

f. Labeling: If you want to name the variables in the most convenient way for coding, please feel free to do so, but you should make sure we can easily find the link between old labels and new labels (we can’t make changes to the labels in the surveyCTO anymore).

g. For the person who is going to clean the variables from grading test, I uploaded the original grading test sheets in “Data\sheetpic”. If you find something weird in the grading results, you can check the original picture (see variable “sheetpic” in the xlsx file).

For reproducibility, please make sure to follow these rules:

a. Please don’t make changes directly on the raw data “weekdayend_all.xlsx”. For temporary files in the process of data cleaning, please put them in “Analysis\temp”.

b. Please put your relevant codes under “Analysis\codes”.

c. When writing codes, please put necessary notes in the key steps. I will also go through the codes at some point.

d. If a code file gets too longer, you can try break it down into a few shorter code files. I was trained in old-schooled computer science before college (I learned C++!), so feel free to use any for-loops or whatever commands to simplify the codes if need be.

e. If you want to generate any tables or graphs to visualize the patterns, please put them in “Analysis\figures” or “Analysis\tables”.

f. Since the data collection is still ongoing, I will update the excel file probably two more times in the next few weeks. All the variable cleaning should not be finalized for these two weeks, but you can write down the structure of the codes so it’s much easier for you to update the codes once we have a larger sample size.