Purpose: This document provides an overview of the folders needed to produce the datasets that will be used in the prediction exercises.

Date of first writing: 11 February 2025

Author: Javier

- The folder "prediction_YL" has three main subfolders:
    - **"0_raw".** Contains raw data files downloaded directly from Young Lives (YL) repository:
        - "Young Lives 1": contains the first 5 rounds of longitudinal data collection from YL. Questionnaire and dictionaries are also included in mrdoc/pdf.
        - "UKDA-8678-stata": contains the data from the 6th round of longitudinal data collection. Participants were surveyed three times over the phone during the Covid pandemic.
        - Four data files called "COUNTRY_constructed" where COUNTRY can be ethiopia, peru, india and vietnam . These are panel datasets constructed by Young Lives putting the 5 first rounds together and displaying a set of relevant variables. These files are useful because they provide a base on top of which to build our final datasets. The original code to construct those datafiles is located in the folder "Panelconstruction_by_YL", but I have not used it --- I directly took the final dataset provided by Young Lives.
    - **"1_scripts".** Contains Stata do-files generating the final datasets for prediction for each country. To produce the final dataset, simply go to "0_master", include your computer's path in the global root_path, and run that do-file. More specifically, "0_master" cleans the data in the following way:
        - First, "1_cleaning_COUNTRY" builds upon "COUNTRY_constructed" and adds variables of interest for our analysis.
        - Second, "2_construction_gendernorms" and "2_construction_sexeducation" construct, for each country, two outcomes of potential interest: (i) views on gender norms and (ii) sex education knowledge.
        - Third, "3_final" takes the datasets constructed in "1_cleaning_COUNTRY", adds views on gender norms (sexual education information can be added in the future) and sets up the final dataset for the prediction. Among others, getting to the final dataset involves (i) dropping variables from rounds 3 and 4

(because for now we only want to use variables up to round 2 to predict round 5 outcomes) as well as variables in round 5 that are not our outcomes of interest, (ii) dropping predictors that are not asked across the four countries (to maximize comparability), and (iii) dropping variables with no variation. The final datasets for prediction are saved in "2_processed/final_datasets".

  - **"2_processed".** Contains auxiliary datasets used to eventually produce the final datasets as well as the final datasets themselves. The final datasets for prediction can be found in the subfolder "final_datasets". Note that, within "final_datasets" there is a subfolder called "codebooks" that has information on the variables present in the final datasets. There are 2 files for each country. "codebook_country" contains the full list of variables. Most of them are labelled. However, there are some variables that are not labelled (although the names of the variables are hopefully self-explanatory). In any case, "codebook_commonvariables_country_CONSTRUCTED " provides the label for most of those variables. []

Additionally, the following are some key variables that are not labelled in any of the codebooks from above:

- sees_dad_daily: indicator taking the value of 1 if child sees dad on a daily basis (based on variable SEEDAD)
- sees_mom_daily: indicator taking the value of 1 if child sees mom on a daily basis (based on variable SEEMOM)
- mean_symptoms: proportion of symptoms out of the following: STTOOLS BLOOD FEVER COUGH RAPIDB VOMIT APPETITE CONVLSE UNCONS LETHARGY
- health_worse_than_others: indicator taking the value of 1 if child is reported to have worse health conditions than other his age (based on variable HEALTHY)
- desire_more_highsch_parents: indicator taking the value of 1 if the parent wishes that the child will get more education than high school
- mean_disability: proportion of disabilities out of the following: DISAB01 DISAB02 DISAB03 DISAB04 DISAB05 DISAB06 DISAB07 DISAB08 DISAB09
- positive_mean_disability: indicator taking the value of 1 if mean_disability is greater than 0
- number_meals_day: number of meals eaten per day (based on variable FOODTOT)
- three_meals_per_day: indicator taking the value of 1 if child eats at least 3 meals per day
- number_foodgroups_eaten: number of different food groups eaten in the last 24 hours (based on variable FDDIVTOT)

- food_shortage: inidcator taking the value of 1 if household has suffered a food shortage in the last 12 months (based on variable FOODSHRT)
- withdraw_daughter_ifinneed: indicator taking the value of 1 if parent thinks a 12 year old daughter should be removed from school if households needs it financially (based on variable FAMDTR)
- withdraw_son_ifinneed: indicator taking the value of 1 if parent thinks a 12 year old son should be removed from school if households needs it financially (based on variable FAMSON)
- hh_cannot_raise_money: indicator taking the value of 1 if parent thinks it would be completely impossible for the household to raise a given amount of money in a week (based on variable RAISE)
- quant_EXPEARN_parents_r2: quartiles of "at what age should child earn money to suppport household"
- quant_EXPEDU_parents_r2: quartiles of "at what age should child leave full-time education"
- quant_EXPIND_parents_r2: quartiles of "at what age should child be financially independent"
- quant_EXPLEAV_parents_r2: quartiles of "at what age should child leave household"
- quant_EXPMAR_parents_r2: quartiles of "at what age should child get married"
- quant_EXPCHILD_parents_r2: quartiles of "at what age should child have a child"