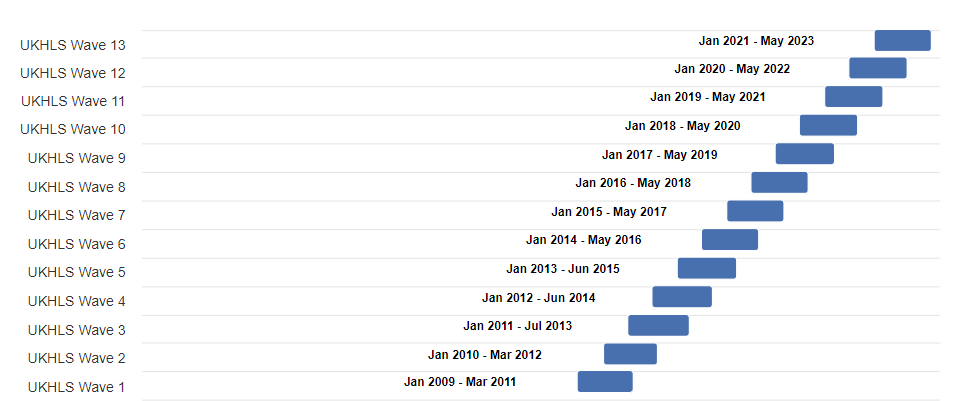
# Key documents

* [Master variable table](https://docs.google.com/spreadsheets/d/1BcuyN424izOygPfoOgIWMKJoF5xn9tUI9XkPxa5BPKo/edit?gid=1796593779#gid=1796593779)

# Introduction to Usoc (UKHLS) & the datasets I’ve created

Understanding Society or the UK Household Longitudinal Survey is a longitudinal survey of 40,000 households that has been running since 2009. There are currently 13 waves. The timeline for data collection for each wave can be found in the diagram below



Source: <https://www.understandingsociety.ac.uk/documentation/mainstage/survey-timeline/>

Usoc has a very complex structure but the key things to be aware of are:

* Respondents may join the survey for the first time in wave 1, wave 13, or anywhere in between!
* Respondents may appear in one wave, disappear, and then reappear in a later wave
* Variables are split between one of several data files corresponding both to the level at which the information is recorded (individual or household) and the age of the respondent (adult or child).

The key datafiles for our purposes are:

* xwavedat - a single cross wave file containing stable characteristics of individuals, such as date of birth, country of birth, ethnicity, which is typically collected only once in the lifetime of the study. These variables are picked from different data files and put into this file.
* xwaveid - a single cross wave file containing some basic sampling information from each wave such as interview outcomes.
* w\_indall - a wave specific file containing some key information from ALL PEOPLE in the household (e.g. age, relationship to other members of the household)
* w\_indresp - a wave specific file containing individual-level data collected from responding ADULTS (e.g. job and health information)
* w\_hhresp - a wave specific file containing household level information (e.g. household income, housing tenure etc)
* w\_child - a wave specific file containing information pertaining to children as reported by their parents and guardians
* peach - a synthetic wide cross-wave data file derived by the Usoc team which contains some key variables on children. All variables present in this dataset have been pulled from other datasets originally.

Essential variables to know

* pidp - the cross-wave unique identifier for each individual. This never changes.
* hidp - the wave specific household identifier for each household. All individuals with the same hidp live in the same household. This can change across waves.

# About the datasets I’ve created

## Overview

I’ve create two main datasets

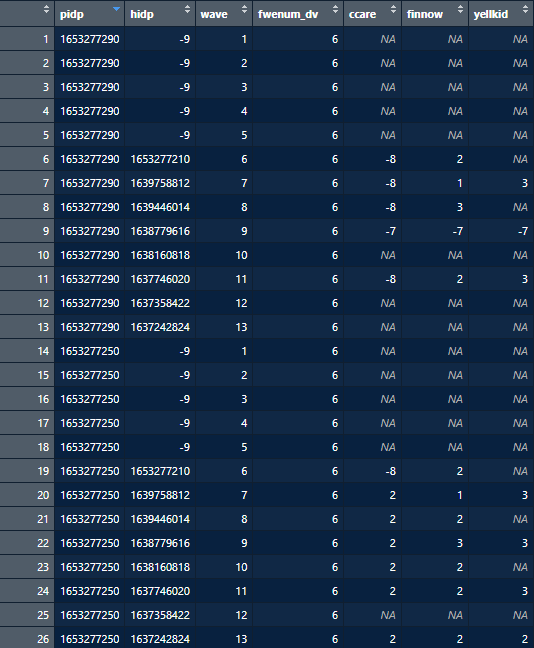
1. ukhls\_master\_enriched\_spine
2. family

|  | ukhls\_master\_enriched\_spine | family |
| --- | --- | --- |
| Who? | Everyone who has ever been part of any wave of Usoc | Any child who we ever observe in Usoc once when under the age of 5. The file contains all observations of that child (pre and post 5) |
| Variables | 515 | 749 |
| Each row contains info about | The individual only and their household | A given child, their household, and their parents |
| Rows | 1,676,207 | 228,254 |
| Observations uniquely identified by | pidp + wave | pidp + wave |
| Unique individuals | 128,939 adults and children | 17,558 children |
| Observations per individual | 13 | 13 |

## Ukhls\_master\_enriched\_spine

The first dataset - ukhls\_master\_enriched\_spine - contains 13 rows (one for each wave), for every person who has ever been enumerated as part of Usoc

### Example data



The snippet above shows the records of two individuals with a random selection of variables

* Rows 1-13 show individual with pidp 1653277290, with one record for each Usoc wave 1-13
  + Most of their variables in waves 1-5 are ‘NA’
  + This is because this individual was first enumerated in the survey in wave 6.
    - We can confirm this by looking at the ‘fwenum\_dv’ variable which records the first wave a person was enumerated.
  + They also failed to participate in wave 10, took part in wave 11, and then disappeared again for waves 12-13
* Any variables that come from the xwavedat or xwaveid data files will not be ‘NA’ missing even in waves where the individual didn’t participate.
  + This is because they contain useful (often but not always time-invariant) information that is useful to have regardless of the wave.
  + For example, the first wave in which a person was interviewed, a person’s ethnicity, whether or not they were born in the UK.
* Variables that come from any other data file will be ‘NA’ missing in waves where an individual didn’t participate
* Rows 14-26 contain records for individual 1653277250.
  + This person also first joined in wave 6
  + In wave 6 the variable ‘ccare’ which asks ‘do you use any childcare for your children’ has the value -8.
    - **Minuses in Usoc always indicate missing information, but their value is informative about why the info is missing**
    - -8 means ‘inapplicable’ - this person likely didn’t have a child
    - By the next wave they do, but they don’t use childcare for them
  + In wave 6 the variable ‘yellkid’ is NA missing
    - This is because this variable is only asked in odd waves
    - By wave 7 we see that they yell at their kid ‘sometimes’

## 

## Family

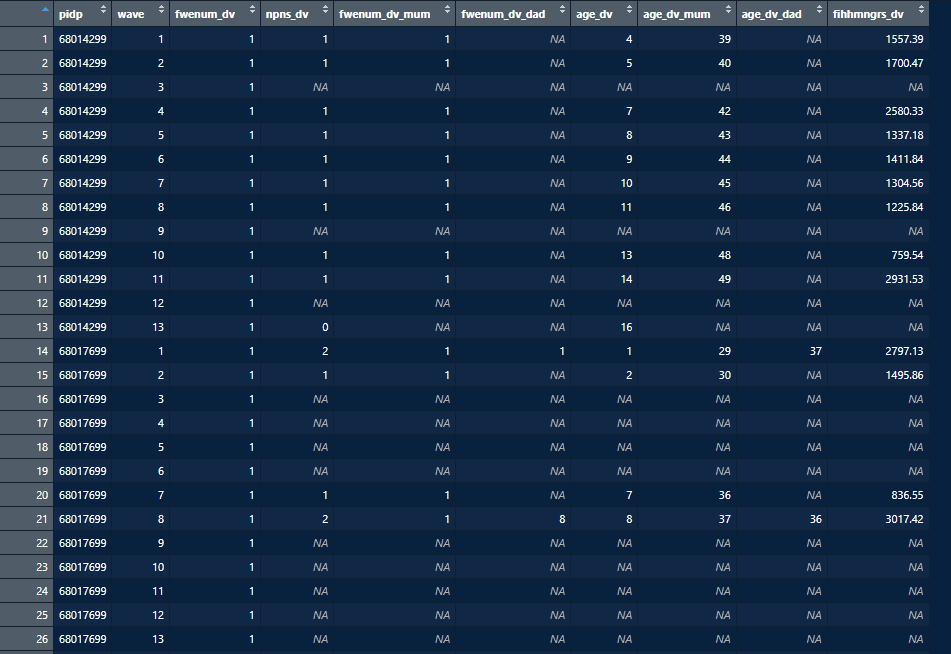
This dataset is formed from ukhls\_master by doing two things

1. Identifying the pidps of any child who we ever observe once under the age of five
2. Filtering the dataset for only individuals with those pidps
3. Identifying the pidps of their mothers and fathers and then joining information from their mothers and fathers onto the child’s row

All variables that come from mothers and fathers are labelled with the suffixes \_mum and \_dad

* **NB: the code use for joining makes use of the variables mnspid and fnspid**
  + These are the pidps of the **natural/step or adopted** mothers or fathers
  + So e.g. a child’s father could change over the course of the study from their natural to their step father
  + We wouldn't distinguish between this in the variables
  + If you wanted to discriminate you would need to change the joining variable to mnpid/ fnpid (natural mother or natural fathers pidp)

### Example data



This shows the records of two children with a short selection of variables

* Rows 1-13 show child with pidp 68014299 in waves 1 to 13
  + We first see them in wave 1 when they are aged 4.
  + We can see from variable npns\_dv (number of parents in the household) that this is a single parent family
  + This is confirmed by the fact that only the ‘\_mum’ variables are not NA missing. They live with their mum only. She is 39 in wave a
  + They come and go between waves, sometimes participating, sometimes not.
  + Their household gross income (fihhmngrs\_dv) fluctuates between waves
  + By wave 13, when they are 16, they appear to be living without any parents
    - In this wave their household gross income is NA missing. This is because I take the household variables from the mother’s record. I do this because I want household variables (such an household income) to be available in the waved prior to the child’s birth
* Rows 14-26 show child with pidp 68017699
  + We first see them in wave 1 aged 1.
  + In wave 1 they live with their mum and dad. Their dad is 8 years older than their mum.
  + By wave 2 they are living with just their mum, and their household income has significantly dropped.
  + They drop out of the survey between waves 3 and 6 then reappear in wave 7.
  + By wave 8 they are 8 years old and living with a new dad who is a year younger than their mum. Their household income has risen significantly.

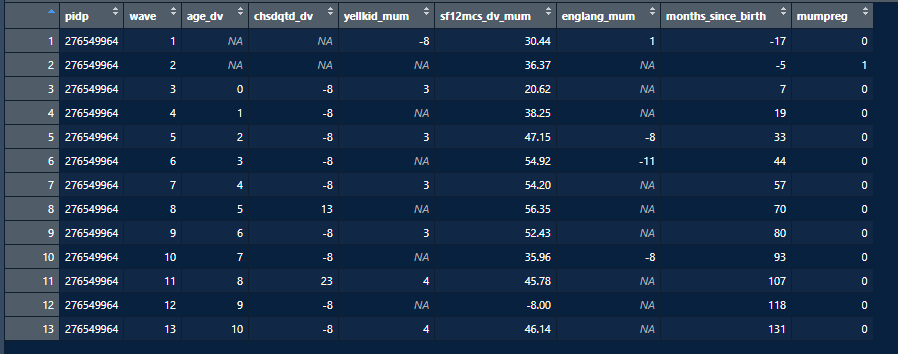
## Why have we got 13 records for everyone even if they didn’t participate in 13 waves?

Hopefully you can see by now that the complexity of the Usoc dataset makes it preferable to keep every record until the specific analytical purpose is determined and all relevant variables created.

Consider this scenario - you want to analyse SDQ scores for 5 year olds and include information about

1. their mother’s mental health in pregnancy and
2. whether or not their mother yells at them
3. whether their mother’s first language is english

First of all you would want to create a sample of children who have observations at age 5 and information from their parents in pregnancy. Here are the records from one child who fits that description (pidp = 276549964)



* The SDQ scores at age 5 come from wave 8 (chsdqtd\_dv)
* The child was born in wave 3, but their mum was pregnant during wave 2. So we would want the information on maternal mental health (sf12mcs\_dv\_mum) from wave 2.
* Their mum is first interviewed in wave 1 when she confirms that she speaks English as a first language. Thereafter she is not asked again. So we need information on englang\_mum from wave 1.
* Yellkid is only asked in odd waves so if we were to take the value in wave 8 (at the same time as the SDQ score) then it would be missing and we would have to drop this child from our analysis. It would probably be better to take it from the most recent non-missing wave (i.e. wave 7, where we learn that their mum yells at them ‘sometimes’)

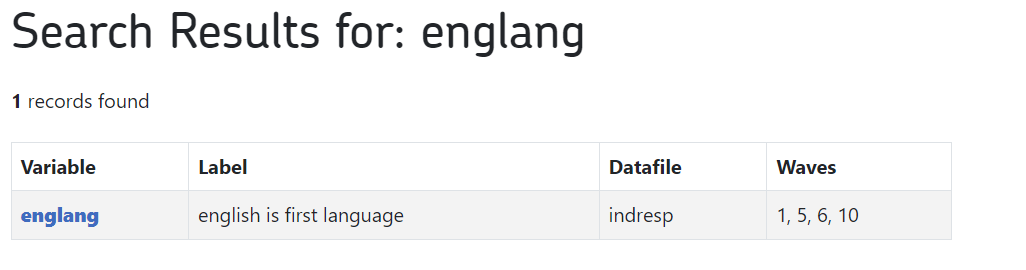
It’s impossible to plan for all of these complexities in advance so the best thing is to keep every record until the final analytical sample is created.

## Finding out more about any variable

All the variables in the final dataset that I have created are logged here in the [master variable table.](https://docs.google.com/spreadsheets/d/1BcuyN424izOygPfoOgIWMKJoF5xn9tUI9XkPxa5BPKo/edit?gid=1796593779#gid=1796593779)  They have been categorised for easier reference.

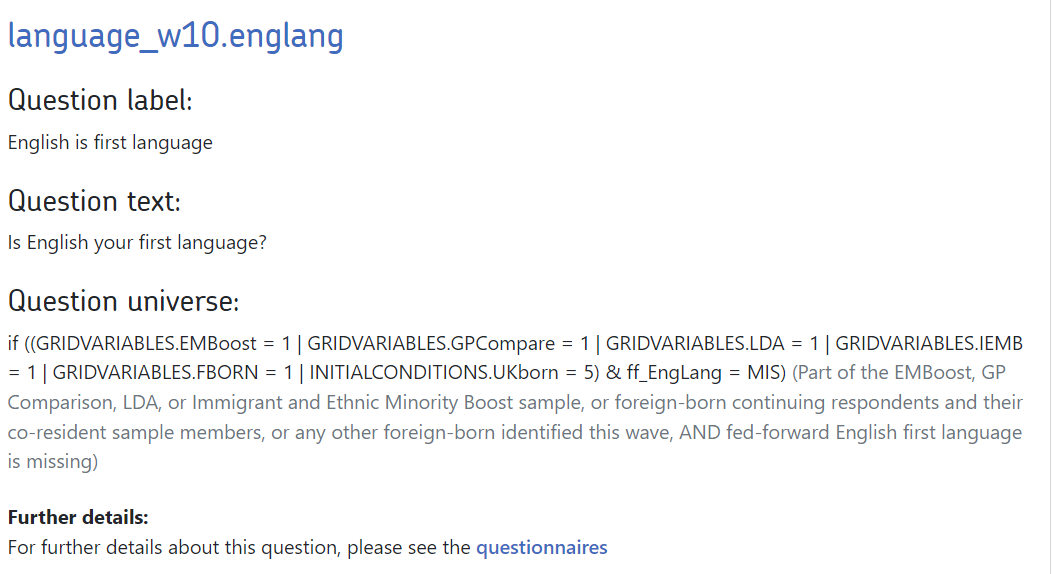
If you want to understand anything further about the variables the next place to go would be Usoc’s [variable search](https://www.understandingsociety.ac.uk/documentation/mainstage/variables/?wpsolr_fq%5B0%5D=index_term_str%3A&wpsolr_fq%5B1%5D=variable_wave_mainstage_str%3A)

e.g.



The variable englang records whether english is someone’s first language and is asked in waves 1,5,6 and 10.

Clicking on the link for the variable we get more information about the question and the universe of people that it is asked of…



What this tells us is that….

* The question is only asked of people who are part of a particular sample or not born in the UK.
* The question is only asked **once** of each sample member (fed\_forward English first language is missing). So if they have been asked it in wave 1 they won’t then be asked it again.

## Further resources

I would highly recommend skimming through this [Understanding Society training playlist](https://www.youtube.com/watch?v=si0rCQvruz4&list=PLN2x0DavrncJyrWhNzv2hdVD9k7NHieMg) on youtube to familiarise yourself with the structure of the dataset.

The Usoc team have also just created a [new user pathway](https://www.understandingsociety.ac.uk/help/new-user-pathway/) with lots of helpful links

Weighting is a huge topic with lots of resources: [start here](https://www.understandingsociety.ac.uk/documentation/mainstage/user-guides/main-survey-user-guide/weighting-guidance/)

You can also ask more complicated questions related to variables/ weighting etc on the [help forums](https://iserredex.essex.ac.uk/support/projects/support). You usually receive a response from the study experts in 1 to 2 days.

# Guide to the code

The code I’ve written so far is intended to be flexible and create a dataset that can easily be adapted for any analytical purpose.

## At a glance

| file | purpose | input | output |
| --- | --- | --- | --- |
| 01-create-data-spine | In this code file I first create a cross-wave long ‘data spine’ which is then customised by merging in all the additional variables likely to be needed for our study. | individual raw usoc files: xwave, indall, indresp, hhresp, lsoa, peach, child and some external data on inflation. | ukhls\_master\_enriched\_spine - a long data spine enriched with variables from the raw usoc files containing entries for everyone every observed  ukhls\_master\_vars - a csv listing all the variables in ukhls\_master\_enriched\_spine and a record of which usoc datafile they originated from |
| 02-derive-variables | This code file derives some key information related to household income, benefit receipt and childcare usage | ukhls\_master\_enriched\_spine - a long data spine enriched with variables from the raw usoc files containing entries for everyone every observed | ukhls\_master\_enriched\_spine\_dv - a long data spine enriched with variables from the raw usoc files and the new derived variables |
| 03-create-family-dataset | This code file creates the **family** dataset from **ukhls\_master** with only records for any child who we ever observe at age five or under.  It then adds to that child data information from their parents and households  The end output is still in the format of a dataspine - a long dataset with records for the child at any wave, regardless of whether or not they were interviewed.This format enables you to examine patterns of non-response, and to create lagged variables based on e.g. parent's income prior to birth, tailored to your analyses | ukhls\_master\_enriched\_spine - a long data spine enriched with variables from the raw usoc files containing entries for everyone every observed | Family - a long data set containing entries for every child ever observed in Usoc under the age of 5 and corresponding information from their parents. |
| 04-create-analytical sample | Just in case of use - this code file creates an example analytical sample for a hypothetical scenario in which you want to analyse SDQ scores for 5 year olds and include information about   1. their mother’s mental health in pregnancy and 2. whether or not their mother yells at them 3. whether their mother’s first language is english | Family - a long data set containing entries for every child ever observed in Usoc under the age of 5 and corresponding information from their parents. | Sample - a dataset containing **\*one row per child\*** with variables relevant to the hypothetical scenario |

## 01-create-data-spine

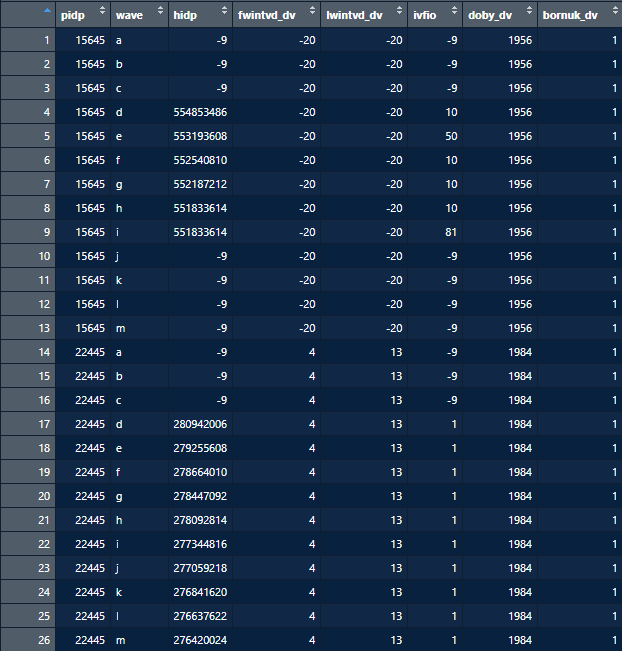
### What is a data spine?

The data spine records key information about each sample member, regardless of age, for all waves of the Understanding Society study, regardless of whether or not they were actually interviewed (remember that not everyone in Understanding Society completes a full interview at every wave).

The data spine is built using xwavedat and xwaveid and has a long structure in which each row contains information for a specific person in a given wave. The combination of pidp and wave uniquely identify each row. The dataspine contains 13 records (one for each wave) for **every adult and child** ever enumerated regardless of the number of interviews they have given. This makes it very large but also very flexible for any intended future analyses which might involve creating lagged variables or investigating patterns of non-response.

Building a data spine enables you to better understand each individual’s record of engagement with the study, and thus take into account the effects of attrition and non-response in your analyses.

A snippet with a selection of variables and rows is shown below.



You can use both Understanding Society’s [variable search](https://www.understandingsociety.ac.uk/documentation/mainstage/variables/) and my master [variable table](https://docs.google.com/spreadsheets/d/1BcuyN424izOygPfoOgIWMKJoF5xn9tUI9XkPxa5BPKo/edit?gid=1796593779#gid=1796593779) to read this dataset

* Rows 1-13 contain records for an individual with the unique identifier (pidp) 15645 born in 1956 in the UK
  + We can see that in wave a-c they have no household identified (hidp)
  + Looking at the variable [ivfio](https://www.understandingsociety.ac.uk/documentation/mainstage/variables/ivfio/), which stands for individual interview outcome, we can see that this person has in fact never given a full interview. They were first approached in wave d and then consistently refused until wave i after which their household was no longer interviewed.
* Rows14-26 contain records for an individual with the unique identifier 22445 born in 1984 in the UK
  + They were first interviewed in wave 4 (fwintvd\_dv) and last interviewed in wave 13 (lwintvd\_dv), giving a full interview each time (ivfio)

### Enriching the data spine

Once we have the data spine we can now join on variables from other datasets to create the ‘ukhls\_master\_enriched\_spine’ output. At each stage I also record where the variables came from and add this information to a dataframe - ‘master\_vars’. This will be helpful later on when I come to join information from parents and children together.

### The code file in detail

| Chunk | Explanation |
| --- | --- |
| LIBRARIES & GETTERS | Loads relevant libraries and getters  **Getters**  Getters are either for single cross wave datasets or multiple wave specific datasets   * Single cross-wave datasets   + Get\_usoc\_slxwaveid\_xwav(filename, fixed\_cols, start\_cols, end\_cols)   + get\_usoc\_peach (wave\_variable, fixed\_cols, start\_cols, end\_cols) * Multiple wave specific datasets   + get\_usoc\_sldata\_long(filename, fixed\_cols, start\_cols, end\_cols)   + get\_usoc\_sllsoa\_long()   Single cross-wave getters simply get and return the dataset with the variables specified. Row are uniquely identified by pidp.  Multiple cross wave getters return a long dataset with responses for each wave bound together. Rows are uniquely identified by a combination of pidp (or hidp) and wave.  Most getters allow you to specify which variables from the raw dataset to return. You can input these in three ways   * fixed\_cols - simply input the full name of the variable * start\_cols - input a prefix and all variables with that prefix will be selected * end\_cols - input a suffix and all variables with that prefix will be selected   All three must be input into the function, even if you input a blank value (“ “) for one of the three.  Inputting a value of “” (no space) will return all variables  The peach getter contains an additional argument ‘wave\_variable’. This is specific to the structure of the peach datafile.  **You also need to set your AWS Credentials here**  I’m sure there is a way to do it permanently but I haven’t worked out how  **NB: Redirecting the getters from my local drive to AWS radically increased the time it takes for all of the getters to run** |
| CREATE AN X WAVE LONG DATA SPINE | Uses xwaveid and xwavedat to create the long data spine  Creates the master\_vars dataframe which records each variable in the final dataset and where it came from |
| AUGMENT THE SPINE WITH TIME-VARYING INDIVIDUAL ADULT INFORMATION | Selects variables from datafiles indall and indresp and merges them onto the master dataset - ukhls\_master  Records where these variables came from using master\_vars  NB: loading the indresp file can take about 5 minutes |
| AUGMENT THE SPINE WITH TIME-VARYING HOUSEHOLD INFORMATION | Selects variables from datafiles hhresp and lsoa and merges them onto the master dataset - ukhls\_master  Records where these variables came from using master\_vars |
| AUGMENT THE SPINE WITH TIME-VARYING INDIVIDUAL CHILD INFORMATION | Selects variables from datafiles child and peach and merges them onto the master dataset - ukhls\_master  When selecting the variables from peach I differentiate between variables which come from specific sets of questions asked in different waves (pregnancy, newborn and parstyle). This is because each has a different associated wave\_variable which records in which wave the question was asked. e.g. the pregnancy questions are asked in the wave prior to the child’s birth. We need the specific wave\_variable to merge these variables onto the master file. |
| AUGMENT THE SPINE WITH EXTERNAL INFLATION INDEX INFORMATION | Adds information on the inflation index each year from a dataset downloaded from the [bank of england’s inflation calculator](https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7bt/mm23?referrer=search&searchTerm=d7bt) .  This is used later on to create inflation adjusted income variables |
| SAVE THE MASTER DATA SET & MASTER VARS | Saves out the results |

## 02-derive-variables

All derived variables are listed in the [master variable table](https://docs.google.com/spreadsheets/d/1BcuyN424izOygPfoOgIWMKJoF5xn9tUI9XkPxa5BPKo/edit?gid=1796593779#gid=1796593779) as ‘LJ DERIVED’ in the ‘original datasets’ column.

### The code file in detail

| Chunk | Explanation |
| --- | --- |
| DERIVE THE INCOME VARIABLES | This section creates a series of ‘real’ household income variables that simply adjust the existing household income variables for inflation using the bank of england’s inflation index  All variables are inflated to 2023 prices |
| DERIVE THE BENEFITS VARIABLES | In this section I derive a series of variables indicating whether or not respondents receive certain types of benefit (e.g. universal credit, housing benefit etc).   * I do this two ways (once using the variables from the indresp file and once using the income file) and then check the answers against each other. This is because there are issues with using either method. * This is necessary because the way that benefits have been captured in the indresp file has changed considerably across waves making it hard to be confident that everything relevant is captured. * The way that benefts are captured in the income file is consistent across waves but it is not possible to distinguish between those who don't receive a given benefit and those with missing information. * Ultimately, I use the variable captured using the indresp file as the master (as we can distinguish nos from missings) but sub in some information from the variable captured using the income file where relevant. * The two are merged in the final stage of this process so there is no need to go into the details of this unless necessary |
| DERIVE THE CHILDCARE VARIABLES | In this section I derive variables relating to use of childcare including   * Use of formal or informal childcare (yes/no) * Hours spent per week in formal childcare * Hours spent per week in informal childcare   **Health warning: The way that childcare usage is recorded in Usoc is a complete mess! Different systems and questions are used at each wave.**  I’ve done my best to work through it and clean it up but I would approach the resulting variables with caution and cross-reference against other sources. E.g. Using my variable use\_formal which records whether or not children make any use of formal childcare I find really low percentages of three year olds who use childcare (36% in wave a rising to 53% in wave m). It should be around 90%. These values aren’t weighted but I would still be very cautious using these variables. |
| RECORD THE VARIABLES IN MASTER VARS | In this section I record where the variables that each of the derived variables are based on came from. |
| SAVE THE MASTER DATA SET | In this section I save out the results |

## 03-create-family-dataset

This code file creates a dataset with records for any child who we ever observe at age five or under. It then joins the child's information (e.g. regarding developmental questions) with information from their parents (e.g. benefit claims) and household (e.g. imd status). The end output is still in the format of a dataspine - a long dataset with records for the child at any wave, regardless of whether or not they were interviewed.

### The code file in detail

| Chunk | Explanation |
| --- | --- |
| DECIDE THE LEVEL AT WHICH EACH VARIABLE APPLIES | In this part of the code I create lists of the names of variables that   1. Belong only to children (e.g. developmental questions) 2. Belong only to individual parents (e.g. mental health) 3. Belong to both parents and children (e.g. ethnicity) 4. Belong to the household (e.g. whether area lived in is rural or urban)   I do this different depending on the original data source file of the variables   1. For xwave and indall variables - I do it by hand using judgement 2. For all other datasets I am able to use their source dataset to easily determine which category they should belong to. E.g. all variable from ‘indresp’ belong to individual parents, all variables from hhresp belong to households.   I use the ‘master\_vars’ dataframe to identify the origin of each variable.  The end results are four list of variable names corresponding to each category   1. **child** - Belong only to children (e.g. developmental questions) 2. **parents -** Belong only to individual parents (e.g. mental health) 3. **both** - Belong to both parents and children (e.g. ethnicity) 4. **Hh -** Belong to the household (e.g. whether area lived in is rural or urban) |
| CREATE A CHILD DATASET | In this section I create a separate dataframe ‘children’   * I filter the master df for all the records for any child who we ever observe once under the age of 5. * I select only variables relevant to children, households, or both adults and children.   I then do some data cleaning   * make some useful time-invariant variables appear in every wave * Some variables from the peach dataset are provided by both ‘parent 1’ and ‘parent 2’ e.g. parenting styles. The variables are named xxx\_par1 and xx\_par2. I identify who ‘par1’ and ‘par2’ refers to and rename the variables xx\_mum and xx\_dad. I rename them xx\_other1 and xx\_other2 if a different adult (e.g. grandmother) is referred to. |
| CREATE PARENT DATASET | In this section I create separate dataframes for the mothers and fathers of the children in the ‘children’ dataframe - ‘mothers’ & ‘fathers’   * I filter the master df for records whose pidps match the mnspid or fnspid variables in ‘children’. This identified records for the natural, step or adopted parents of those children. * I select variables relevant to parents and both adults and children * Additionally when creating the ‘mothers’ dataframe I also select the household variables   + I do this so that we can have information on household characteristics in the waves prior to birth. |
| JOIN PARENTS AND CHILDREN | I then join the child and mum datasets together to create a ‘family’ dataset  First I need to create bespoke joining variable: mnspid\_join and fnspid\_join   * To ensure that we get information from parents available in the waves prior to those in which the children were born I need to ensure that the mnspid and fnspid variables are not missing in the waves prior to birth. * To do I will pull the first observed values of mnspid and fnspid into the waves prior to birth (because mnspid and fnspid refer to either natural/step or adopted parents their value could change across waves e.g. if natural father was present at birth, then moved out, and was then replaced by a step father)   Then I join children and mothers using ‘mnspid\_join’   * Because I’ve pulled household information from children and mothers this creates a lot of excess variables. * I clean this up, using information from the mothers only when information from the children is NA missing (i.e. before their birth)   Then I join children and fathers using ‘fnspid\_join’ |
| CREATE THE MUMPREG AND MONTHS\_SINCE\_BIRTH VARIABLES | Finally I create two useful variables that might be helpful for later analysis   * Months\_since\_birth - a variable which counts backwards and forwards the months since the child’s birth. E.g.   + if the interview is conducted 4 months before child’s birth, months\_since\_birth = -4   + If the interview is conducted when the child is age 2yrs and 3 months, months\_since\_birth = 27 * Mumpreg - indicates whether or not the child’s mother is pregnant with the given child.   + =1 if months\_since\_birth is between -9 & -1 |
| SAVE THE FILE | I save the new dataset as ‘family’ |

## 04-create-analytical-sample

This is probably totally redundant for you Jess (as I’m sure you know this better than me!) but just in case it is of use this code files creates an example analytical sample for a hypothetical scenario in which you want to analyse SDQ scores for 5 year olds and include information about

1. their mother’s mental health in pregnancy and
2. whether or not their mother yells at them
3. whether their mother’s first language is english

### The code file in detail

| Chunk | Explanation |
| --- | --- |
| CREATE TAILORED VARIABLES | In this section I create variables tailored to the research question |
| PULL TAILORED VARIABLES INTO EVERY WAVE | In this section I ensure that the tailored variables are available in every wave. |
| IDENTIFY THE CORRECT WEIGHTS | (I haven’t done this but but an essential step) You would then need to select the correct weights for your research question - this can be straightforward or fiendishly difficult. Our dataset contains every potential weight. To pick the correct one make sure you consult the [Usoc weighting guidance](https://www.understandingsociety.ac.uk/documentation/mainstage/user-guides/main-survey-user-guide/weighting-guidance/) and ask a question in the Usoc [help forums](https://iserredex.essex.ac.uk/support/projects/support) if unsure. |
| SELECT ONLY RELEVANT WAVES | Create the final sample by ensuring that only one wave per child is included |

# What if I want to…

## Add a new variable

1. Go to the [Understanding Society Variable Search](https://www.understandingsociety.ac.uk/documentation/mainstage/variables/?wpsolr_fq%5B0%5D=index_term_str%3AIncomes%3A%20Benefits%20and%20Allowances%20and%20Pensions&wpsolr_fq%5B1%5D=variable_wave_mainstage_str%3A01) and search for the variable you want. Identify the original datafile that the variable is contained in.

The procedure for adding a new variable to the dataset will be slightly different depending on the original datafile

2a. If the variable is contained in hhresp, indresp, child, peach or parstyle

* Simply add the variable to the ‘select’ statement under the relevant section in 01-create-enriched-spine.R
* Rerun 01-create-enriched-spine, then 02-derive-variables then 03-create-family-dataset

2b. If the variable is contained in xwaveid, xwavedat or indall

* Add the variable to the ‘select’ statement under the relevant section in 01-create-enriched-spine.R
* Rerun 01-create-enriched-spine and 02-derive-variables
* Decide the level at which the variable applies
  + In 03-create-family-dataset chunk *‘*DECIDE THE LEVEL AT WHICH EACH VARIABLE APPLIES *’*: add the new variable to either the ‘parents’ ‘child’, ‘both’ or ‘hh’ lists
* Rerun 03-create-family-dataset

2c. If the variable is contained in a new dataset

* In 01-create-enriched-spine.R :
  + use the ‘get\_usoc\_sldata\_long function with the name of the new dataset to load all waves of the new dataset
    - E.g. get\_usoc\_sldata\_long(“youth”)
  + Select the relevant variables
  + Left\_join the new variables onto the master dataset
  + Add the names of the variables and their origin dataset to ‘master\_vars’
* Rerun 01-create-enriched-spine and 02-derive-variables
* Decide the level at which the new variable applies
  + You are likely to be able to do this based on the dataset from which it is drawn from
    - i.e. all variables from the youth file would apply only to children so could be added to the ‘child\_rest’ list based on their origin dataset.
* Rerun 03-create-family-dataset

## Change the sample

As a reminder the ukhls\_master dataset is made up of everyone who has ever been part of any wave of Usoc.

The family dataset is currently made up of any child who we ever observe in Usoc once when under the age of 5. The file contains all observations of that child (pre and post 5).

The family dataset is made from the ukhls\_master dataset by filtering observations and then joining parents and child variables together.

You can easily change the the filtering conditions in 03-create-family-dataset code chunk CREATE A CHILD DATASET/ SAMPLE SELECTION

## 