**Decisions I am making.**

* I think we want the final data to be entirely raw values merged together
  + Ie. we don’t create % private in the raw data, just present the data as it comes
  + UPDATE: I’m changing my mind – I think it should be a) raw numbers published and then b) final in % form because the denominators all changes so much that we can’t expect anyone else to do this accurately.
* I think we want it to be panel data with long form years but and long form variables
* Where APIs are not available from Gov – we will publish the raw ODS, XL, CSV files on our github and clean it all directly from there.
  + UPDATE: struggling to read xlsx off github cos of my own bad tempfile skills.
    - But this is enough to convince me it isn’t worthwhile and we should publish all as CSVs and then clean from there.
    - Ultimately then there should be code to download and convert files to CSVs so that we can reproduce this process via code.
      * Even if no one in their right mind would reproduce it – and dockerise versions might struggle
* The final sheet will work by pulling different functions so that the reproducible file looks like eg.
  + Spend\_Data <- clean\_expenditure\_data()
  + Children\_Data <- clean\_SSDA\_data()
  + Housing\_Data <- clean\_homeless\_data()
  + Full\_Data <- merge\_all\_data(Spend\_Data, Children\_Data, Housing\_Data)
    - Magic
* We should include all data available at LA level.
  + Eg. education outcomes by LA
* I think we should keep only variables that are ongoing to this day
  + Ie. if there is a measure of educational attainment at key stage 1 that is used 2010-2013, we get rid of this.
    - There are a whole bunch of these and we will need to make a call about what is worthwhile and what isn’t – for the sake of usability.
      * Eg. one year they report grades by gender… is it worthwhile keeping this one measure? Not in my opinion.
  + Alternative could be only keep them if they run from 2010 to 2023
    - So the vars that are just 2018-23 get removed.
  + This really fucks Key stage 4 where for 6 years it was GCSEs and then changed to English baccalaureate.
    - And the only measure going back is ‘percent passing English and Maths’ but this changes from a\*-c to levels 4/5 in two separate variables from 2017.
    - So I think only include KS4 outcomes from 2016 onwards.
* We should have a ‘terminology’ section to explain terms used in datasets
  + Eg. ‘outcomes’ are defined by DfE not us.
* When confidence intervals of estimates are reported, I’m ignoring them.
  + Eg. average progress 8 score for Key stage 4 comes with upper and lower 95% CI…
    - I guess the argument would be that what were are creating is a dataset for comparison over time, which is different to the precision of the actual number in any given year… not sure that holds though and I’m not just being lazy
* I think one limitation we should be clear of is CLA category we are using (rather than CIN)
* 2010 is year 1 – there are sometimes random variables available before 2010 which I’m ignoring
  + Often the 2010/11 release which has the previous 3 years presented or something.
* Variables are being labelled as per the most recent releases.
  + Even if the DfE’s labels are fucking dumb…
* The denominator should be of children eligible not necessarily a sub-category.
  + Eg. the ‘percent value’ of children receiving intervention for substance use is of all children in care
  + Or eg. the percent of children with SEN statements is the % of all children in care
    - Not of children with substance use identified or SEN identified
    - but will need to check DfE do the sameee
      * UPDATE: correct for sen