1. Downloaded all full year consolidated data files from 2008 to 2019 (didn’t want to include covid). Each file contains data on two panels – the one for which this is the first year and the one for which it is the second – I separate them into two separate dataframes in R since we are interested in since we are interested in using the first-year data for each panel to predict second year outcomes.
2. Renaming variables:
   1. Technical: many variables refer to rounds 1/3, 2/4 or 3/5 according to which panel – e.g. the variable codes the data of round 1 for the panel that started this year and the data of round 3 for the panel that started the previous year. I rename the variables to have the suffix \_rx where x is the round matching the panel (since we have already separated the panels). Any variable that refers to yearly data (would have for example 08 in the variable name) is changed to refer to the panel year (1 or 2) by removing the calendar year and replacing it with the suffix \_y1/\_y2. The same suffix is added for variables that code end of year information (1231 in the end of the original variable name). Some variables refer to other years (e.g. in the 2008 data there are variables that refer to 2007, to all years before 2007 and to 2009) replace number of year prior and year after with \_y0/\_y1/\_y2/\_y3 so we can avoid for example using 2009 data in the prediction for the panel that started in 2008.
   2. Rename variables with the following prefixes to define variable categories (according to variables categories in the documentation + a little bit of manual renaming):
      1. survey administrative and id variables: prefix - surv\_ad\_
      2. demographic variables: prefix – dem\_
      3. Income and Tax Filing Variables: prefix – inc\_
      4. PERSON-LEVEL CONDITION VARIABLES: prefix – cond\_
      5. Health Status Variables: prefix – hstat\_
      6. DISABILITY DAYS VARIABLES: prefix – dd\_
      7. ACCESS TO CARE VARIABLES: prefix – acc\_
      8. EMPLOYMENT VARIABLES: prefix – emp\_
      9. HEALTH INSURANCE VARIABLES: prefix – ins\_
      10. Expenditure and charging variables: prefix exp\_
      11. Utilization variables: prefix – util\_
      12. Weights variables: prefix – w\_
   3. For now, did this for the first year of panels 13-15 (years 2008-2010) since more naming work was needed for next panels, should be easy to add at least some of them later.
   4. In the second year of each panel rename the expenditure variables with prefix y\_exp\_ and utilization variables with y\_util\_
   5. Merge both years of each panel – inner\_join – keeps only people with data in both years.
   6. Merge panels – variables that are not available in a specific panel get NA.