HADS custom variable definitions

* OWNRENT (comes from Tenure)
  + 0 – Rental
  + 1 – Owner
* COST06, COST08, COST12, COSTMED
  + Calculated cost to owner, this may be what we want to predict
* UTILITY
  + Is imputed for vacant units using monthly rent, structure type, region and tenure
* BURDEN
  + -1 for houses with no income, otherwise housing cost divided by monthly income
* Assisted housing
  + Overreported, use with caution

Variables that might be useful for prediction

ZINC2 – Household Income

IPOV – Poverty level income (according to the area)

ABLMED, ABL30, ABL50, ABL80 – Median income adjusted for bedrooms

ASSISTED – 0, 1, -9 – Are they receiving some kind of government assistance

FMR – Fair market rent – calculated by HUD as a factor of many things for a given county

LMED, L80, L50, L30 – Income levels of area

APLMED – Income adjusted for persons in household (AHS underestimates sometimes)

age1 – Age of head of household

BEDRMS

BUILT

OWNRENT

METRO3

REGION – Census Region

STATUS

STRUCTURETYPE / TYPE

ZADEQ – Adequacy of unit

ZSMHC – Monthly housing costs – we may have to delete this because it includes mortgage

NUNITS - # units of building

OTHERCOST – ZSMHC may be inconsistent, OTHERCOST is sum of insurance, land rent (not rent), and fees.

PER (# Of persons in household)

ROOMS (# Of Rooms in Unit)

TENURE (Owner/renter status of unit)

TOTSAL – Sum of salary income over all members of household

UTILITY (Monthly utility cost)

VALUE – Current market value (what we are trying to predict)

Coded Variables (taken almost verbatim from IV. A on the pdf):

Adequacy (ZADEQ):

* adequate, moderately inadequate, severely inadequate, vacant–no information
* We will use vacant-no information as our ‘0’variable

Bedrooms (BDRMS) #NOTE – while the pdf says this is coded, it appears it is not, there are 0 to 8 bedrooms in the document. It is possible bedrooms were coded for earlier years:

* PDF
  + Studio, 1, 2, 3, 4 or more
  + Studio will be our 0 variable
* Dataset
  + 0 – 8
  + 0 will be our 0 variable
  + OR we could make this continuous

Location #NOTE: doesn’t appear to be in dataset. There is a region variable, but with 4 categories

* PDF
  + Central city, suburb, nonmetropolitan
  + Nonmetropolitan: 0
* Dataset
  + 1,2,3,4
  + 1 will be 0

METRO3: no ‘categorical’ notes in the pdf

* 1,2,3,4,5,9
* 9 will be 0

Structure Type:

* Single unit, 2-4 units, 5-19 units, 20-49 units, 50+ units, mobile homes, no info
* 1,2,3,4,5,6,-9
* No info (-9) will be 0
* Recoded from Type; we should not use both TYPE and STRUCTURETYPE

Tenure:

* Owner, renter
* Dataset has 1,2,3,-6
* -6 will be 0

Year Built

* After 201035, 2000-2009, 1990-1999, 1980-1989, 1960-1979, 1940-1959, Pre-  1940.
* Pre 1940 will be 0

Discard Variables:

BURDEN – is correlated to ‘Value’

CONTROL – The AHS control number. I believe this is just a private key

COSTX – anything cost related, because we can directly compute the fair market value with that

GL30, GL50, GL80, GLMED – Doesn’t pertain to our dataset

Anything FMR – these are all just recalculated percentages of FMR

ADEQ – Not in model, we use ZADEQ

Anything that relates the housing price to another variable: affordability related to AMI and poverty income