R. R. Day
July 2019
Data Wrangling: Capstone – American Housing Characteristics Trends with
US Census, America Housing Survey (AHS), 2017

The characteristics of the housing stock are assumed to reflect the affordable preferences in the location where and at the time when the houses were built. The 2017 AHS captures a snapshot of the characteristics of the housing stock as were in place at the time of the survey. The geographic locations are composed of nine regions. The records specify whether or not they are in the 15 largest metropolitan areas and identify them if they are. House preferences change or persist over time, captured in decades, and between locations. This data reveals the nature of the changes and the differences between regions and between regions and metropolitan areas.[i]

The US Census Department along with US Housing and Urban Development conducts the AHS every two years. While the AHS contains all housing because the collection is associated with the US Census, this project includes single family detached and single family attached structures. It has been a tenet of the American Dream to live in and preferable own your own house. There is anecdotal evidence that is changing in certain populations, single family houses are still being built with particular characteristics.

The data is from the most recent survey, 2017 and is generally available in the AHS Public Use Files (PUF), which comes in two formats, tabular and flat files, respectively. This project uses the National PUF data downloaded in a flat, comma separated file downloaded directly from the website: https://www.census.gov/programs-surveys/ahs/data/2017/ahs-2017-public-use-file--puf-/ahs-2017-national-public-use-file--puf-.html .

The code book for the AHS is massive because of changes and cross-references to variables and measurements that have changed over decades of surveying and reporting. In 2017, there are 3180 columns reported for each record.

The administrative and geographic variables are complete; none are missing. There is a control variable (CONTROL) assigned to every record. As discussed previously, the dataset was filtered for only single family detached and single family attached houses. In the National PUF, the metropolitan area is captured in the variable named "OMB13CBSA" which indicates US Office of Management and Budget 2013 Core Based Statistical Area. For geographic region, the nine Census Divisions (DIVISION) include all 50 states and the District of Columbia. The map of Divisions is at the website: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf. For privacy issues, State and County level data are not public, similarly to the exact year built, and therefore, are not in the National PUF.

The eight housing characteristic variables are in categorical numeric strings and continuous numeric integer types, respectively. There were 66,752 non-null values for each of my selected variables, filtered for single family, detached and attached, there are 43,910. In preparing the National PUF, some variables have been aggregated into categorical variables. In particular the unit size (UNITSIZE) in square feet is not in uniform increments. UNITSIZE is ordinal and appears as a number in a strings with apostrophes which required changing them to read as integers. The varying increments of square feet will require translation to uniform increments to provide meaningful comparison in relation to some other variables on an as needed basis. At this point the decision to convert UNITSIZE to square footage increments of various sizes, would assign unnecessary precision to the data.

Another housing characteristic variable, how many bathrooms (BATHROOMS), required much the same treatment, ordinal categorical strings turned into integers. Again, to render the analysis meaningful for relation to other variables, values for BATHROOM differentiate between full bath (sink, toilet, and tub or shower) and half baths (only sink and toilet) are ordinal strings in the PUF and were converted to integers for analysis. (For example, differing from the real estate convention listing of "1.5" and '2," a full bath plus a half bath is '2,' and two full baths are '3,' respectively.) However, there were other values which represented combinations of bathroom fixtures that did not constitute the first full bath which required aggregation to indicate that no full bath was present, and given a '0.'

Also, the year built (YRBUILT) is aggregated into decades, as the underlying data are not publicly accessible. The most recent YRBUILT value, '2010,' only includes seven years because the latest survey was in 2017. The houses built earlier than 1920 have the value of '1919' in the PUF. This was replaced with '1910' and will be noted as comprising all the houses built in the 1910's and before.

For non-responses for these characteristics variables, "no report" is encoded as "-9." Of, the ten house characteristics variables, only two, UNITSIZE and whether it included a garage/carport (GARAGE) had "no report" values. Because these specific variables are categorical, and they represent under 20 percent by geographic region (DIVISION), the records will be retained and modified as needed.

Aside from these administrative and characteristics variables, tenure (TENURE) indicates whether the property is owned or rented. The AHS in total is rich with detailed building and related household consumption characteristics, as well as demographics and financial metrics.

---

[i] Modifications of the houses built in other decades are beyond the scope of this project.