**AMERICAN HOUSING SURVEY: HOUSING AFFORDABILITY DATA SYSTEM**

*Download*: [Excel](http://archive.ics.uci.edu/ml/machine-learning-databases/00240/) Workbook, ReadMe

**Abstract**:  is a set of files derived from the 1985 and later national American Housing Survey (AHS) and the 2002 and later Metro AHS. This system categorizes housing units by affordability and households by income, with respect to the Adjusted Median Income, Fair Market Rent (FMR), and poverty income. It also includes housing cost burden for owner and renter households. These files have been the basis for the worst case needs tables since 2001. [[1]](#footnote-1)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 1881 | **Area:** | Housing | |
| **Attribute Characteristics:** | N/A | **Number of Attributes:** | 18 | **Missing** **Values?** | | N/A  -6 |

**Source:**

https://www.huduser.gov/portal/datasets/hads/hads.html

**Updated Codebook:**

Original Codebook credited to https://www.huduser.gov/portal/datasets/hads/hads.html

Updates made by Kim Kirk “Business Statistics and Analysis” for Rice University.

**Data Set Information:**

The *Housing Affordability Data System (HADS)* is a set of files derived from the 1985 and later national American Housing Survey (AHS) and the 2002 and later Metro AHS. This system categorizes housing units by affordability and households by income, with respect to the Adjusted Median Income, Fair Market Rent (FMR), and poverty income. It also includes housing cost burden for owner and renter households. These files have been the basis for the worst case needs tables since 2001. The data files are available for public use, since they were derived from AHS public use files and the published income limits and FMRs. We are providing these files give the community of housing analysts the opportunity to use a consistent set of affordability measures.[[2]](#footnote-2)

Data variables include:

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Explanation |
| CONTROL | Character | Primary key |
| AGE 2011 | Numeric | Age of head of household |
| ALL\_OTHER\_METROS | Numeric | Dummy variable for the Metropolitan status:  Not Central City |
| REGION\_WEST | Numeric | Dummy variable for the census region West |
| REGION\_MIDWEST | Numeric | Dummy variable for the census region Midwest |
| REGION\_NORTHEAST | Numeric | Dummy variable for the census region Northeast |
| LN\_LMED\_2011 | Numeric | Log transformed Area Median Income |
| LN\_FMR\_2011 | Numeric | Log transformed Fair Market Monthly Rent |
| VALUE 2013 | Numeric | Current market value of unit in 2013 |
| BEDRMS 2011 | Numeric | Number of Bedrooms in the unit |
| LN\_BUILT\_2011 | Numeric | Log transformed Year the unit was built |
| LN\_VALUE\_2011 | Numeric | Log transformed Current Market Value of the unit |
| ROOMS 2011 | Numeric | Number of rooms in the unit |
| LN\_PER 2011 | Numeric | Log transformed number of persons in household |
| LN\_ZINC2\_2011 | Numeric | Log transformed Annual Household income |
| ZADEQ\_ADEQUACY | Numeric | Dummy variable for Adequacy of unit:  Moderately Inadequate  Severely Inadequate |
| LN\_UTILITY\_2011 | Numeric | Log transformed Monthly utilities cost (gas, oil, electricity, other fuel, trash collection, and water) |
| LN\_OTHERCOST\_2011 | Numeric | Log transformed Sum of ‘other monthly costs’ such as Home owners’ or renters’ insurance, Land rent (where distinct from unit rent), Condominium fees (where applicable), Other mobile home fees (where applicable). |

**Data Summaries Calculated:**

* Mean
* Standard Error
* Median
* Mode
* Standard Deviation
* Sample Variance
* Kurtosis
* Skewness
* Range
* Minimum
* Maximum
* Sum
* Count

**Units for Data Summaries and Variables:**

|  |  |
| --- | --- |
| **Name** | **Unit Measurement** |
| CONTROL | N/A |
| AGE 2011 | Age in years |
| LN\_LMED\_2011 | Dollars |
| LN\_FMR\_2011 | Dollars |
| VALUE 2013 | Dollars |
| BEDRMS 2011 | One bedroom structure |
| LN\_BUILT\_2011 | Years |
| LN\_VALUE\_2011 | Dollars |
| ROOMS 2011 | One room structure |
| LN\_PER 2011 | One person |
| LN\_ZINC2\_2011 | Dollars |
| LN\_UTILITY\_2011 | Dollars |
| LN\_OTHERCOST\_2011 | Dollars |

**Data Transformations:**

1. Dataset was downloaded from host website: <https://www.huduser.gov/portal/datasets/hads/hads.html>
2. Data was cleaned to include only single-family houses, flats, apartments with Fair Market Value of more than $1000.00 owned in 2013 as this was the subset that stakeholders wanted to focus on.
   * Missing values made up 4.1% of the data and were discarded per stakeholder input.
3. VALUE 2013 was determined to be the outcome variable.
4. AGE 2011, METRO3, REGION, LMED\_2011, FMR\_2011, VALUE\_2013, BEDRMS 2011, BUILT\_2011, VALUE\_2011, ROOMS 2011, PER 2011, ZINC2\_2011, ZADEQ\_ADEQUACY, UTILITY\_2011, OTHERCOST\_2011 were determined to be the predictor variables.
5. LMED\_2011, OTHERCOST\_2011, FRM\_2011, BUILT\_2011, VALUE\_2013, PER\_2011, ZINC2\_2011, UTILITY\_2011 had exponential distributions and a natural logarithmic transformation was applied for a normal distribution.
6. New variables created include:
   * LN\_LMED\_2011
   * LN\_FMR\_2011
   * LN\_VALUE 2013
   * LN\_BUILT\_2011
   * LN\_VALUE\_2011
   * LN\_PER 2011
   * LN\_ZINC2\_2011
   * LN\_UTILITY\_2011
   * LN\_OTHERCOST\_2011
7. 1000 observations were randomly selected as Test data; the rest of the data set was used as Training data for the predictive model.
8. Worksheets were created to hold four categories for the analysis:
   * Summary Report identifies and answers the business question/what was measured
   * Descriptive Statistics identifies and calculates descriptive statistics
   * Graphs & Charts displays histograms, scatterplots, bar graphs
   * Statistical Tests holds results for statistical tests

1. https://www.huduser.gov/portal/datasets/hads/hads.html [↑](#footnote-ref-1)
2. https://www.huduser.gov/portal/datasets/hads/hads.html [↑](#footnote-ref-2)