# Housing Affordability Data System

Group 3

Kefu Zhu

# 1. Introduction

Housing survey data in 2009, 2011 and 2013 from U.S. Department of Housing and Urban Development website[1] were used in this project. Data from 2009 and 2013 are in raw data file. Only data from 2011 is in SAS data file. These data belong to The Housing Affordability Data System (HADS). HADS 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.

The business interest of this project is to prepare the data to be ready for potential meaningful analysis or comparison. Several questions can be answered after the preparation of the data. Do young people tend to buy expensive houses relative to their income than the old do? Who is living in a house that cost more than his/her income?

# 2. Methods

SAS will be used to perform reading data, data validation, data cleaning and merging datasets in this project. The dataset from 2009 contains 49090 observations. The dataset from 2011 contains 145531 observations. The dataset from 2013 contains 64535 observations. All three datasets contain around 100 variables each. But in this project, I will only use 14 variables for the business interest. Description for each variable used in this project can be seen in Figure 1 below. All three datasets were sorted by **CONTROL**.

| Variable | Description |
|---|---|
| CONTROL | Control number |
| VACANCY | Vacancy status |
| BEDRMS | # of bedrooms in unit |
| NUNITS | # of units in building |
| BUILT | Year unit was built |
| STRUCTURETYPE | Recoded structure type |
| BURDEN | Housing cost as a fraction of income |
| AGE1 | Age of head of household |
| ZINC2 | Household Income |
| ZSMHC | Monthly housing costs |
| VALUE | Current market value of unit |
| UTILITY | Monthly utility cost |
| PER | # of persons in household |
| REGION | Census region |

FIGURE1. Description of Variables Information

[1] U.S. Department of Housing and Urban Development
Retrieved from https://www.huduser.gov/portal/datasets/hads/hads.html

## 2.1 Guidelines

Although all 14 variables have numeric value, **STRUCTURETYPE** and **REGION** are actually categorical variables. Therefore, I will first recode the numeric values in **STRUCTURETYPE** and **REGION** to their corresponding string representations based on the source code[2].

Then I will do some simple exploratory analysis to check if there are any suspicious values. I will also validate and clean the data based on some common sense:

1. The reasonable range for the age of head of household should be between 20 and 90.
2. The current market value of unit should not be non-positive.
3. Monthly utility cost should not be non-negative (zero values are worth investigated).
4. The reasonable range for housing cost as a fraction of income should be between 0 and 1.
5. The number of persons in the household should not be non-positive.
6. The number of units in the building should not be non-positive.

I will perform validation steps mentioned above first on dataset from 2009. After considering all results from validation process of 2009 dataset, I conclude a criterion for cleaning data and use that to clean all three datasets.

After validating and cleaning the datasets, I will merge all three datasets together by **CONTROL** and create a permanent dataset called ***merged***. I will also subset the merged dataset for the business interest.

1. Create a permanent dataset called ***luxury***, which contains subjects who is living in a house that cost more than his/her income.
2. Create three permanents datasets called ***young***, ***middle_age*** and ***old***, which only contain young people, middle aged people and old people respectively.

If any unexpected issues appear during the any process, I will investigate those problems further and adjust the final criterion for cleaning and merging data.

[2] SAS source code used to generate the HADS datasets
Retrieved from https://www.huduser.gov/portal/datasets/hads/hads_source.exe

# 3. Results

## 3.1 Data Validation

### 3.1.1 PER

Based on Figure 2, 4033 observations have value of -6 for **PER** in the housing survey data from 2009. Since **PER** measures the number of persons in household, it is not reasonable to have negative values. However, because the frequency table does not show any missing values, the system might use -6 to indicate missing values or other special meanings.

| PER | -6 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 4033 | 12429 | 15024 | 6896 | 6218 | 2862 | 996 | 382 | 150 | 59 | 22 | 10 | 5 | 4 |

FIGURE2. Frequency Table of PER in 2009 Dataset

### 3.1.2 BEDRMS

Based on Figure 3, 445 observations have value of 0 for **BEDRMS** in the housing survey data from 2009. Because **BEDRMS** measures the number of bedrooms in unit, it is not very likely to have zero value. So I investigated this problem further by printing out the information of other variables for the first 10 observations who do not have bedroom in their units.

| BEDRMS | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 445 | 6020 | 12566 | 19919 | 8082 | 1661 | 299 | 63 | 18 | 10 | 7 |

FIGURE3. Frequency Table of BEDRMS in 2009 Dataset

Among ten printed observations, 9 of them have -6 value in **VALUE** and 8 of them have -6 value in **VACANCY** (Figure 4). Few observations also have suspicious values in other variable like -9 in **AGE1**, **PER**, **ZINC2**, **ZSMHC**, and negative values in **BURDEN**. Since observations that have zero value in **BEDRMS** also have unreasonable values in other variables, I will suggest to investigate more on this problem in future work.

| CONTROL | AGE1 | BEDRMS | PER | REGION | BUILT | VALUE | VACANCY | NUNITS | ZINC2 | ZSMHC | STRUCTURETYPE | UTILITY | BURDEN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| '100129330103' | 36 | 0 | 1 | West | 2006 | 225000 | -6 | 78 | 80000 | 2432 | 50+ units | 64 | 0.36480 |
| '100206390143' | 77 | 0 | 1 | Northwest | 1980 | -6 | -6 | 198 | 12800 | 550 | 50+ units | 0 | 0.51563 |
| '100575110145' | 89 | 0 | 1 | West | 1980 | -6 | -6 | 25 | 24000 | 1460 | 20-49 units | 0 | 0.73000 |
| '100910730103' | -9 | 0 | -6 | West | 2009 | -6 | 1 | 371 | -6 | -6 | 50+ units | 105 | -9.00000 |
| '100914500142' | -9 | 0 | -6 | South | 1985 | -6 | 1 | 8 | -6 | -6 | 5-19 units | 92 | -9.00000 |
| '101379360144' | 25 | 0 | 1 | Northwest | 1985 | -6 | -6 | 90 | 95000 | 1916 | 50+ units | 66 | 0.24202 |
| '101596630103' | 81 | 0 | 1 | Midwest | 2008 | -6 | -6 | 102 | 26568 | 552 | 50+ units | 0 | 0.24932 |
| '102929520146' | 23 | 0 | 2 | West | 1980 | -6 | -6 | 15 | 56000 | 1356 | 5-19 units | 56 | 0.29057 |
| '129641930123' | 89 | 0 | 1 | Midwest | 1980 | -6 | -6 | 37 | 23888 | 399 | 20-49 units | 0 | 0.20044 |
| '141502630133' | 30 | 0 | 1 | Midwest | 1950 | -6 | -6 | 24 | 100000 | 541 | 20-49 units | 25 | 0.06492 |

FIGURE4. First Ten Observations Who Do Not Have Bedroom in 2009 Dataset

### 3.1.3 AGE1

Figure 5 shows the frequency of **AGE1** for observations who are either younger than 20 years old or older than 90 years old. There are 4033 observations who have age of -9 and 429 observations have age of 93. Being -9 years old is definitely impossible in real life. But since there are no missing values in the frequency table and the proportion of observations who have -9 values for **AGE1** is huge (81.67%), it is reasonable to assume that the system use -9 to represent missing values. It was also suspicious that all head of households who are older than 90 years old are actually recorded as 93. So I investigated this problem further by printing out the information of other variables for the first 10 observations who are 93 years old.

| AGE1 | -9 | 14 | 15 | 16 | 17 | 18 | 19 | 93 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 4033 | 2 | 1 | 16 | 291 | 42 | 124 | 429 |

FIGURE5. Frequency Table of AGE1 in 2009 Dataset

Based on Figure 6, all ten printed observations have value of -6 in **VACANCY**.

| CONTROL | AGE1 | BEDRMS | PER | REGION | BUILT | VALUE | VACANCY | NUNITS | ZINC2 | ZSMHC | STRUCTURETYPE | UTILITY | BURDEN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| '100596500146' | 93 | 2 | 1 | Midwest | 1990 | -6 | -6 | 344 | 19223 | 655 | 50+ units | 80.000 | 0.40889 |
| '100653130103' | 93 | 3 | 1 | Midwest | 2004 | 175000 | -6 | 1 | 85880 | 2249 | Single Family | 287.583 | 0.31425 |
| '100871550143' | 93 | 1 | 1 | West | 1985 | -6 | -6 | 4 | 15030 | 2433 | 2-4 units | 33.333 | 1.94251 |
| '100883140141' | 93 | 2 | 2 | Northwest | 1990 | -6 | -6 | 44 | 93602 | 1172 | 20-49 units | 72.000 | 0.15025 |
| '101418970145' | 93 | 3 | 1 | South | 1990 | 250000 | -6 | 1 | 16540 | 288 | Single Family | 125.000 | 0.20895 |
| '101481120646' | 93 | 3 | 1 | Northwest | 1990 | 125000 | -6 | 1 | 1200 | 385 | Single Family | 145.000 | 3.85000 |
| '102238740143' | 93 | 2 | 1 | Northwest | 1985 | 200000 | -6 | 107 | 18060 | 712 | 50+ units | 37.000 | 0.47309 |
| '104589770144' | 93 | 2 | 1 | Midwest | 1985 | -6 | -6 | 217 | 3996 | 2026 | 50+ units | 0.000 | 6.08408 |
| '129006930113' | 93 | 3 | 1 | South | 1960 | 150000 | -6 | 1 | 9775 | 249 | Mobile Home | 207.333 | 0.30568 |
| '129859930113' | 93 | 3 | 2 | West | 1985 | 13000 | -6 | 1 | 50301 | 96 | Mobile Home | 83.000 | 0.02290 |

FIGURE6. First Ten Observations Who Are 93 Years Old in 2009 Dataset

And upon further investigation, all observations who are 93 years old have value of -6 for **VACANCY** in 2009 dataset (Figure 7). Although I did not have more information about the meaning of -6 value for **VACANCY**, this problem is worth to be investigated further in future work.

```
NOTE: No observations were selected from data set WORK.RAW_H2009.
NOTE: There were 0 observations read from the data set WORK.RAW_H2009
      WHERE (AGE1=93) and (VACANCY not = -6);
```

FIGURE7. Partial Log Output Data Validation Process of 2009 Dataset

### 3.1.4 VALUE

17610 observations have negative values in **VALUE** and all those values are -6. Since these observations account for 35.87% of the total number of observations in 2009 dataset. It is not quite possible that those values are simply input errors. So I speculate that -6 value in **VALUE** represents missing values or other special meanings.

### 3.1.5 UTILITY

2510 observations have zero value of **UTILITY** in the housing survey data from 2009. This might indicate that no one is living in those houses. Upon further investigation, 78.29% of those observations have value of -6 in **VACANCY** (Figure 8). Although I did not have information on the meaning of -6 value for **VACANCY**, it is reasonable to guess that -6 represent "Empty". Therefore, I did not remove subjects who have -6 value in **VACANCY** in data cleaning process.

| VACANCY | -6 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 1965 | 349 | 13 | 128 | 28 | 27 |
| Percent | 78.29% | 13.9% | 0.52% | 5.1% | 1.12% | 1.08% |

FIGURE8. Frequency Table of **VACANCY** for subjects who have no monthly utility cost in 2009 dataset

### 3.1.6 BURDEN

4033 observations have value of -9 and 550 observations have value of -1 in **BURDEN**. Negative values might indicate that those people do not have jobs and they receive subsidy from other people or the government to pay for their housing cost. But I still could not explain the situation of having two distinctive negative values in **BURDEN**. This problem should be investigated in future work.

Except negative values in **BURDEN**, there are also 3173 observations which have value greater than 1 for **BURDEN**. One possible explanation is that those people get loans from banks or other financial services for their housing payments. However, based on Figure 9, the median (2.07) of **BURDEN** is way smaller than the mean (23.62). The value of skewness (22.949) is also extremely large compared to the normal threshold of 1. Compared to the minimum, mean and median, the maximum value (8368) is also too large to be reasonable.

| N | Mean | Median | Standard Deviation | Min | Max | Skewness |
|---|---|---|---|---|---|---|
| 3173 | 23.62 | 2.07 | 283.78 | 1 | 8368 | 22.95 |

FIGURE9. Descriptive Statistics of BURDEN in 2009 dataset

All information indicated that the distribution of **BURDEN** is extremely right-skewed and those observations that have huge value in **BURDEN** need to be investigated in future work. It is not quite possible for a person to afford house that is worth thousands times of his/her income.

## 3.2 Data Cleaning

Based on the result from data validation step, I cleaned all three datasets based on following criterions:

1. The number of bedrooms should be greater than zero.
2. The number of units in the building should be greater than zero.
3. The number of persons in the household should be greater than zero.
4. The age of subject should be between 20 and 90.
5. Current market value of the unit should be greater than zero.
6. Monthly utility cost should be greater than zero.

In addition, I decided to adjust my previous criterion about **BURDEN** based on extra research. According to article on Forbes[3], highest value of *Median Home Price Relative To Median Annual Income* in 2012 was 6.8, which correspond to people who live in either Los Angeles or San Francisco. So I decided to only include subjects who have value of **BURDEN** between 0 and 10.

After the cleaning process, the dataset from 2009 contained 29420 observations. The dataset from 2011 contained 80434 observations and the dataset from 2013 contained 34731 observations. All three datasets contained 15 variables.

## 3.3 Merging Data

I first merge all three datasets simply by **CONTROL**. But the log output (Figure 10) shows that the number of observations in the merged dataset is not equal to the sum of observations in 2009, 2011 and 2013 dataset. One possible explanation is these three datasets contain duplicated **CONTROL** values.

```
NOTE: There were 29420 observations read from the data set WORK.H2009.
NOTE: There were 80434 observations read from the data set WORK.H2011.
NOTE: There were 34731 observations read from the data set WORK.H2013.
NOTE: The data set WORK.MERGED has 122887 observations and 15 variables.
```

FIGURE10. Log Output for Merging Three Datasets

[3] High Home Price-to-Income Ratios Hiding Behind Low Mortgage Rates. Forbes.
Retrieved from http://www.forbes.com/sites/zillow/2013/04/16/high-home-price-to-income-ratios-hiding-behind-low-mortgage-rates/#d447e7378df6

Upon further investigation, 21698 observations, which were not included in the merged dataset, have the same control number with at least one other observation. So I decided to only include observations that have the latest information during merging process. In other words, if an observation appears twice in both 2009 and 2011 dataset, only the information from 2011 dataset will be included in the merged dataset. In the ***merged*** dataset, there are 122887 observations and 15 variables.

## 3.4 Subsetting Data

Based on the business interest, I created a permanent dataset called ***luxury***, which contains subjects who is living in a house that cost more than his/her income. The ***luxury*** dataset has 5691 observations and 15 variables.

I also created a permanents dataset called ***young***, which contains subjects who are between 20 and 40 years old; a permanents dataset called ***middle_age***, which contains subjects who are between 41 and 65 years old and a permanents dataset called ***old***, which contains subject who are older than 65 years old. The ***young*** dataset has 24881 observations and 15 variables. The ***middle_age*** dataset has 67729 observations variables. The ***old*** dataset has 30277 observations and 15 variables.

# 4. Future Work

Since -6 appears multiple times in **VACANCY**, **PER**, **ZINC2**, **ZSMHC**, and **VALUE**, it is worth to further investigate the actual meaning of -6. **BURDEN** will also be studied more for the meaning of its two distinctive negative values, -9 and -1, and extreme positive values.

Other problems I will investigate more in the future include the -9 and 93 values in **AGE1** and zero value in **BEDRMS**.