Housing Affordability Data System
STAT 440
Group 3 Project Proposal

Kefu Zhu

1. **Background information**

Data is downloaded from U.S. Department of Housing and Urban Development website (https://www.huduser.gov/portal/datasets/hads/hads.html). The data belongs to The Housing Affordability Data System (HADS). 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.

2. **Description of Data File**

Three data files are used in this project. Those are housing data from HADS for 2009, 2011 and 2013. Each data file contains thousands of observations and less than a hundred variables. For example, in the data file of 2013, there are 64535 observations and 99 variables. The data is also sorted by control number.

3. **Research Interests**

Cleaning the data to be ready for potential meaningful analysis or comparison, such as do young people tend to buy expensive houses relative to their income than the old do.

4. **Data Preparation**

(1) Convert numeric value in METRO3 into corresponding string and make the data more readable

(2) Extract only needed variables from the original dataset, such as BURDEN (Housing cost as a fraction of income), AGE1 (Age of head of household) and etc.

(3) Read two raw data file into SAS (Data from 2009 and 2013)

(4) Merge or concatenate data from 2011 and 2013 together

(5) Validate variables based on common sense

  (a) Age of head of household should not be smaller than 20 or greater than 100 (or even 90)

  (b) The value of BURDEN should not be negative

(6) Subset data based on meaningful condition to find some insights

  (a) People who is living in a house that cost more than his/her income