We broke down the housing data for analysis based on location criteria. The first level by which the data was broken down was census region. There are four census regions in the USA. The map below shows the regions which are numbered as follows: 1 = Northeast, 2 = Midwest, 3 = South, and 4 = West. The regions were further broken down into two categories: Central City and Suburban. These categories are not necessarily the best breakdown that could have been achieved, as to call everything outside of a city suburban is a gross overstatement. However, this is what was provided in the HADS dataset. (The Census Bureau officially breaks down each of the four regions as shown in the map, however these breakdowns were not included in the dataset we analyzed.)

To analyze each of the eight sub-regions (census region and city or suburb), we did the following:

1. We loaded the data into Excel and checked each of the columns we planned to use and identified what value was used as ‘null’ by the creators of the datasets.
2. In our Map function, we checked for these values. Rows with null values were given a count of 0. Furthermore, households with null incomes were given an income of 0.
3. If none of the values under consideration were null, we calculated a metric we called rating. Rows with null values were also given a rating of 0.
4. The rating was calculated as follows: the age of the head of household mod 10 was added to the number of persons in the household. Then 1 was added if the household had the “owned” while 2 was added for a status of “rented.”
5. A key of the form “Region number, City or Suburb” was emitted with a value of the form “count, rating, income”.
6. In the reduce function, we created variables for total count, income count, rating sum and income sum.
7. We obtained a count of the total households rated by summing the counts of the households. Since households with null values (which were not rated) were given a count of 0, they could be added to the total count without affecting the value, alleviating the need for if statements. The ratings could be summed in the same way since households with null values were also given a rating of zero.
8. The income value given in the dataset was described as a sum of wage and salary income. Some households reported a value of 0, which was separate from the null value for this column. (A possible explanation for this is that null values were used for vacant homes while 0 was used for income that was unreported. Another possibility is that houses with 0 income from wage or salary had an alternative income source such as social security, investments, etcetera, not distributed in a traditional manner.) When summing the incomes, we used an if statement to exclude incomes of 0 and also computed a separate count to get a more accurate average.
9. The reducer emitted the following for each of the eight sub-regions: number of households rated, average rating, and average income.

The table below shows the output for the 2013 HADS data. Due to the large number of households analyzed the relatively low magnitude of our rating value, a column was added to include rating \* 104, to normalize the values to whole numbers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Region | City / Suburb | # Homes Rated | Average Rating | Average Rating (x 10^4) | Average Income |
| 1 | City | 5597 | 0.0031 | 31 | 42333.86 |
| 1 | Suburban | 9943 | 0.0018 | 18 | 61866.58 |
| 2 | City | 4942 | 0.0035 | 35 | 36299.53 |
| 2 | Suburban | 11276 | 0.0016 | 16 | 50061.6 |
| 3 | City | 5114 | 0.0034 | 34 | 38109.1 |
| 3 | Suburban | 12518 | 0.0014 | 14 | 45703.45 |
| 4 | City | 4026 | 0.0045 | 45 | 48353.62 |
| 4 | Suburban | 6681 | 0.0027 | 27 | 50537.85 |

We expected our rating to correspond with household income in a parabolic manner. That is, we would expect medium scores --middle aged household head and middle number of residents-- to have higher incomes than those with old household heads or with young head of households and many occupants. Below is a scatterplot of the normalized average rating versus average income for each sub-region.

As shown by the plot, aside from the outlier, which corresponds with the data for cities in the western region, the rating and income data correspond approximately as we predicted. We predict with better criteria to breakdown location we would see a stronger curve.

The graph below is a bar graph showing average income for each of the eight sub-regions. The bars are colored differently for city and suburban sub-regions.

It is interesting to note that average city income is lower in all four regions. This is not necessarily unexpected though, as the “Suburban” category covers a great deal more territory than the “City” designation. It is also interesting that the western region has the lowest discrepancy between sub-categories, while the northeastern region’s is the greatest.