Paper TITLE

Summary/Intro

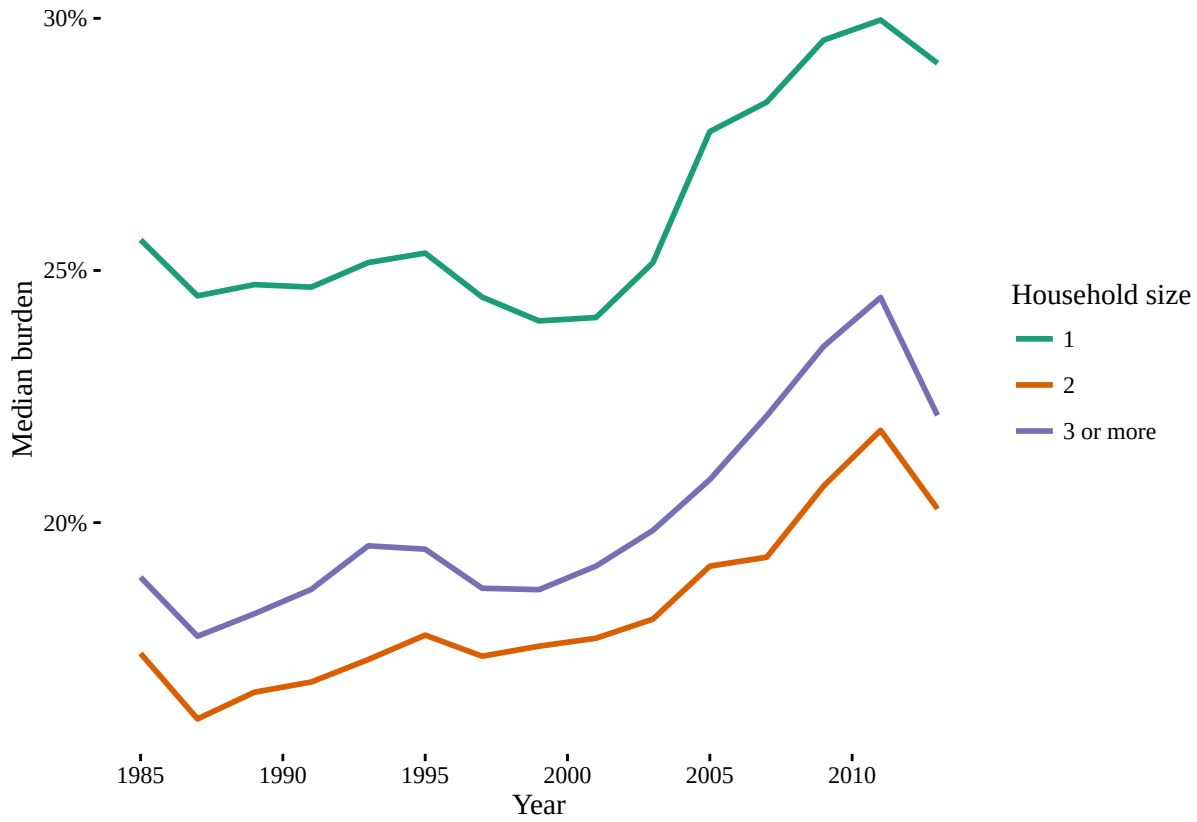
We chose the dataset from the American Housing Survey, which derives sets of data from 1985 and on. The data set includes categorizing each housing unit by factors such as affordability, income through Fair Market Rent and the Adjusted Median Income. Taking into account variables such as Burden (housing cost divided by its monthly income), certain households with a zero or negative income are input as -1 within the data. The data source was recorded from the American Housing Survey. Certain factors, such as Poverty Income, is based on the Census Bureau's official thresholds.

We chose to analyze the data and ultimately view determines whether housing affordability has increased or decreased over the last few decades. Our initial hypothesis assumes that housing affordability within the United States has decreased due to key, significant events that have occurred within the last few years. The United Stateâ s growing income inequality can be attributed to the current state of housing affordability, and is a clear indicator of the shifting income brackets between who is poor and who is rich. Another major event that could have potentially affected the current state of housing affordability, is the 2007 Housing Market crash that affected the entire country, and even the world. Between all of the foreclosures, loan defaults, and stock market crash, we will closely examine the years between 2007 and 2011 to capture a better sense of the housing market affordability status.

Cleaning data

Cleaning data Initially, the dataset is split into fifteen datasets, one for each odd year between 1985 and 2013. Due to differences in how the survey was conducted over the years, each dataset has varying column names and positions, and, for surveys performed earlier than 2003, missing columns. To standardize the data, we used the format of the more modern datasets, i.e the datasets from 2003 to 2013. Thus, we begin by appending a column onto each dataset to mark which year that data was gathered, remove an extraneous column from the 2009 dataset, and then simply combine the datasets. For the older years, the datasets must again be split into two groups as some columns are named differently and do not fit with our standard. Once we rename the columns, we add the missing columns that dictate the raw, adjusted by growth, and adjusted by people income limits for each household. These income limits specifically determine what the extremely low income bracket is for the area each household is in. For each list of datasets, we reorder the columns to exactly match that of our base, the datasets gathered from 2003 onward. Finally we combine every list of datasets into a single dataset, and then store it in a comma separated file called combined_years.

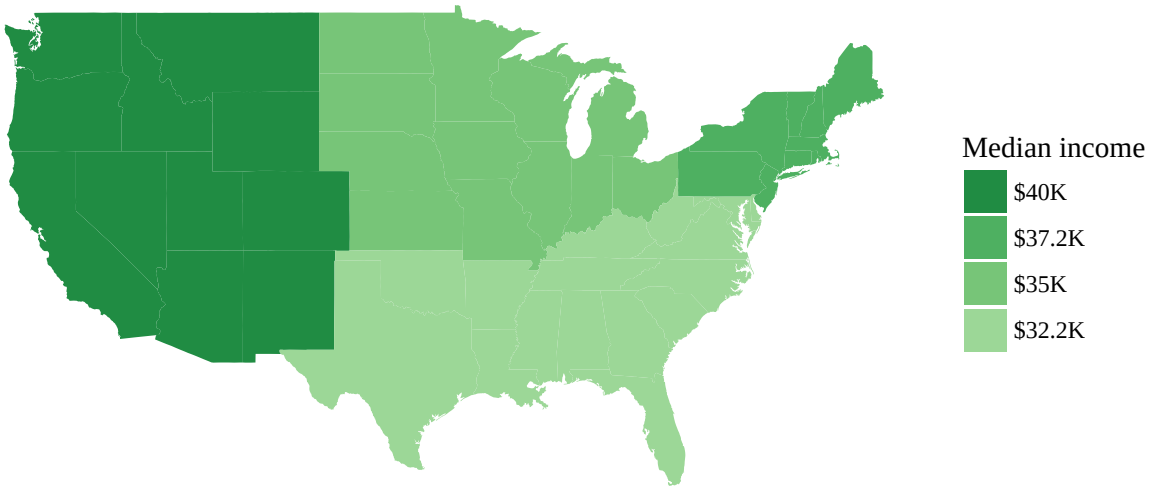
The "combined_ years" data was read into a dataframe using the "read_ csv" function. Within the read_csv function, the column types were explicitly stated as a string variable (c = character, i = integer, etc.), and the nonsensical integer date (negative numbers) were converted to NAs in the dataset. Then, the combined years dataset was cleaned using the piping methods in dplyr. All the column names were set to lowercase, and all the quotations were removed from character values. Then, an additional column â ownrentâ was added which labeled levels 1 and 2 to own and rent as factors. Then, the data was written to a clean_years csv file.



The graph shows the development of median burden (proportion of housing costs relative to household income) from 1985 to 2013. The median burden varies dramatically between households of different sizes. The highest burden through the years has been on households of one person. This is understandable, since people living alone cannot split any costs between other earners.

Surprisingly, households of two people have the lowest burdens overall. Our hypothesis is that this is due to a large proportion of two-earner couples in this group. As housing cost burden is simply a householdâ s monthly housing cost divided by its monthly income, households where all members are earners have a higher monthly income and a lower resulting burden. In households of 3 or more people, the burden picks up again. This is likely due to a larger amount of children and / or retirees in larger households.

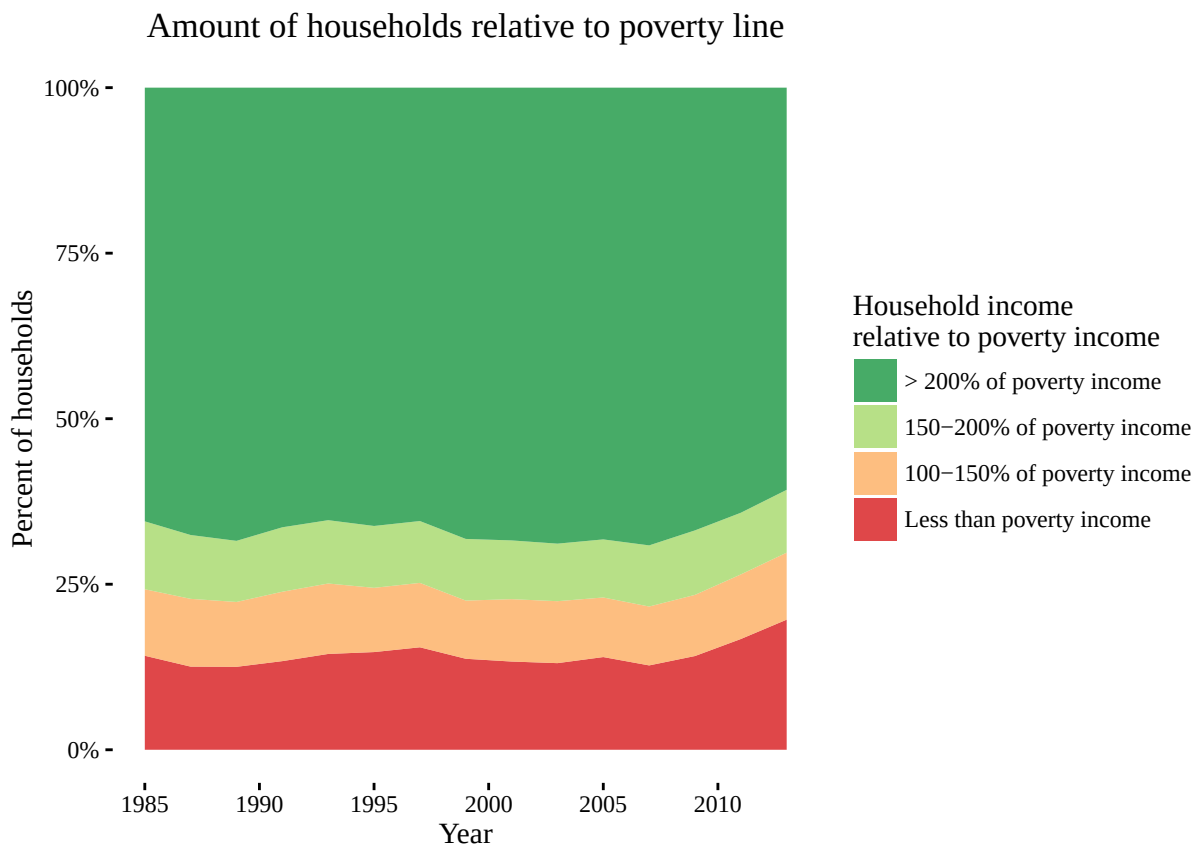
Another interesting thing is the dramatic increase of burden during the buildup to the financial crisis 2000-2007. The most probable exlanation is the housing market price bubble. Once the crisis hit in 2007, the effect of dropping incomes continued to increase burdens until 2011. Only in 2013 data did the burdens start to fall from the unprecedently high levels.



To gauge income levels in different areas we used the four census regions in the dataset and joined them with a scraped table from mapsoftheworld.com. Using the table we could connect census regions to state names and thus colour a map of the continental US with census region incomes. We used median incomes instead of mean incomes to avoid skewing the results with outliers.

The household median monthly incomes vary between different census regions from 32200 dollars in the South to nearly 25% higher, 40000, in the West. The Northeast trails closely behind the West with 37267.50 dollars, and the Midwest is closer to the South with a median income of 35000.

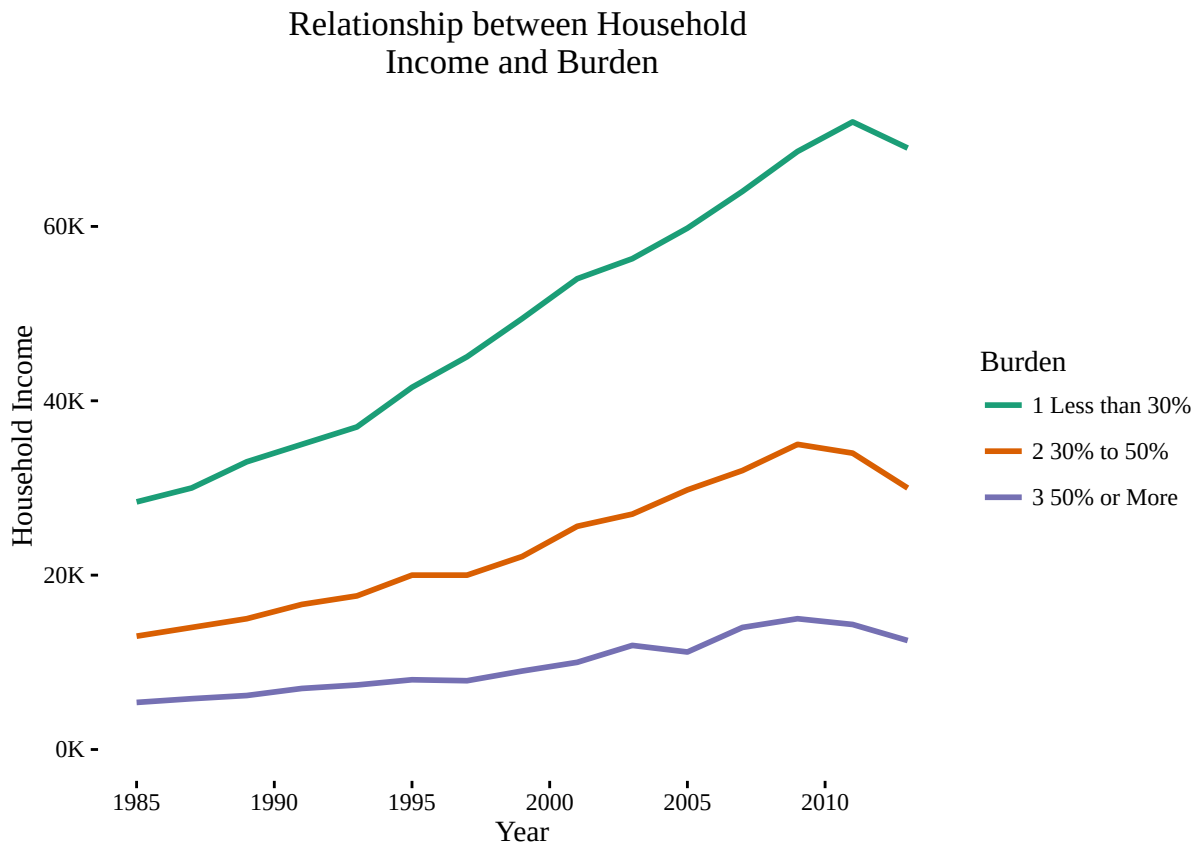
While the dataset doesn't have more granular location data, even these four regions are very distinct from each other. The large high-income cities in the Northeast and the West likely contribute to the higher respective median incomes, while more rural Midwest and South are poorer.



This plot shows the amount of households in a given income group in a given year. Each household has been calculated a poverty treshold according to the Census Bureauâ s official thresholds, based on family size and number of children.

As can be seen from the plot, the amount of people in different income groups stayed relatively constant between 1985 and 2007. The majority of people earned over twice the poverty income, while little under a fifth of the population was in poverty. This stable relationship broke noticeably in the financial crisis of 2007 and the following recession, when the amount of people in poverty started a quite steep and steady climb upwards.

Interestingly, the income brackets of 100%-200% stayed almost exactly the same size as before, while the increase in poverty was in effect solely the result of the dwindling of the highest (over 200% of poverty income) income bracket. A reasonable explanation would be the effect of increasing unemployment, with previously employed well-off people dropping to unemployment and thus small or no income. The reality can be more nuanced, with people dropping to the middle income groups and a simultaneous equally large movement from these brackets to poverty, but the big picture of large impacts in poverty following the financial crisis stays the same.



As pictured in the graph, the relationship between household income and burden is seen as a positive correlation. From the graph, you can see that higher income household have less burden to deal with, which makes sense because those with more money are able to pay off monthly costs more easily. Those with 50% more burden are near the poverty line. From this graph, you can see a spike in households that have burden. Households with higher incomes started to pay more burdens, which can be attributed to 2007 financial housing market crisis.

From this graph, you can see that the Northeast and West have higher rates of burden at higher income levels, therefore illustrating how housing affordability in these regions are less than that of the Midwest and South.

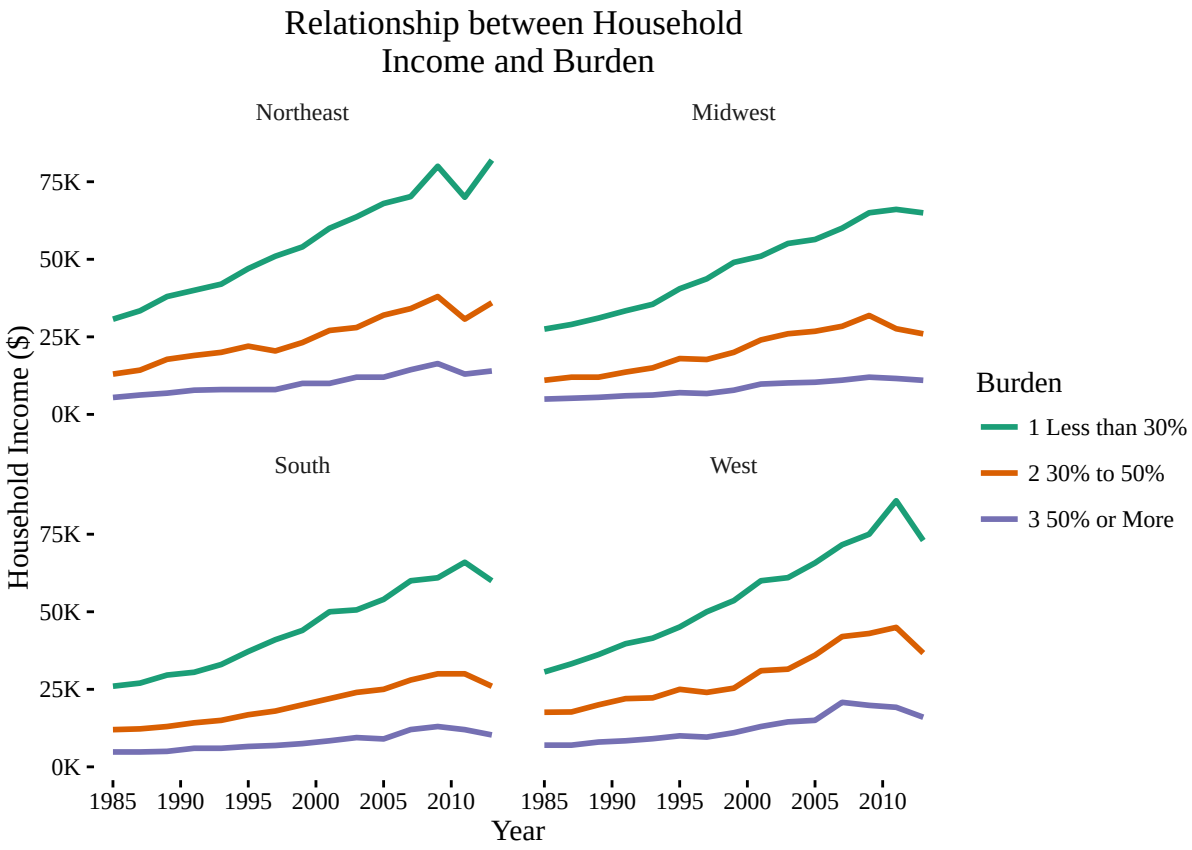
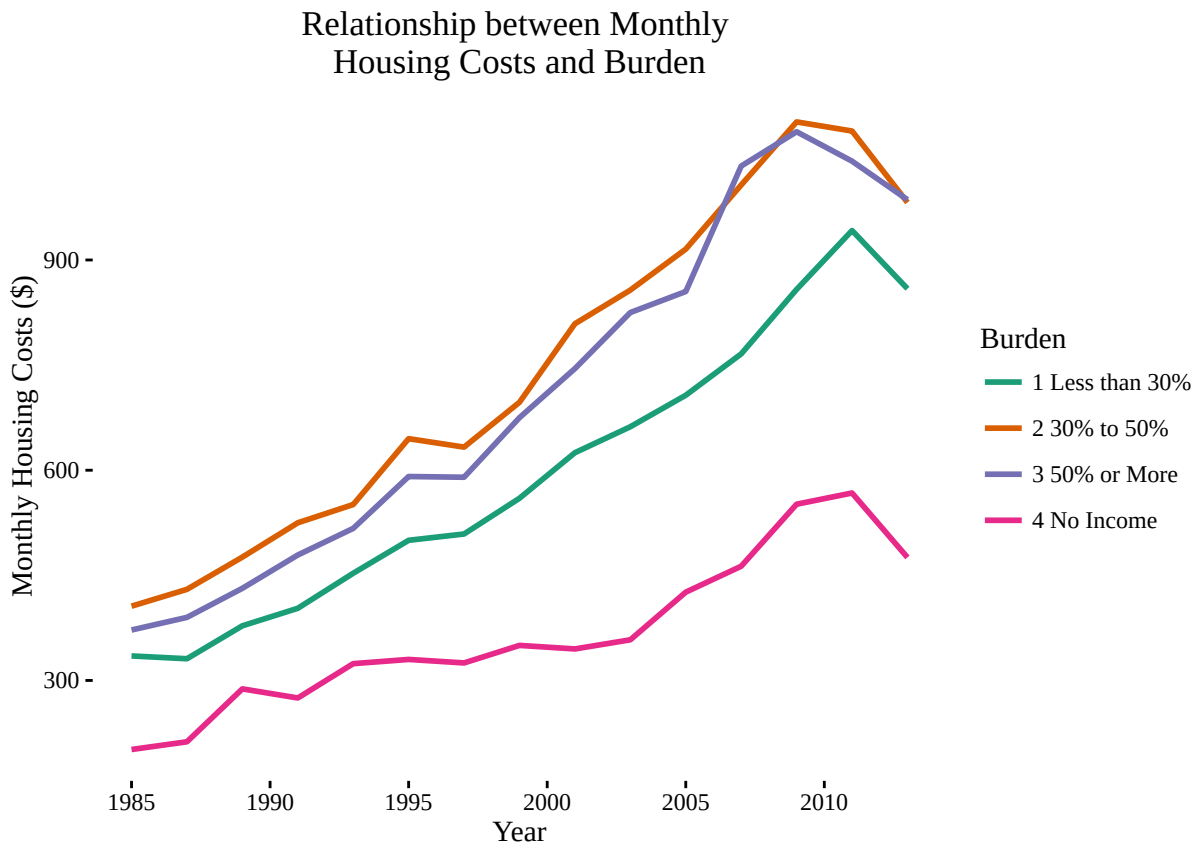
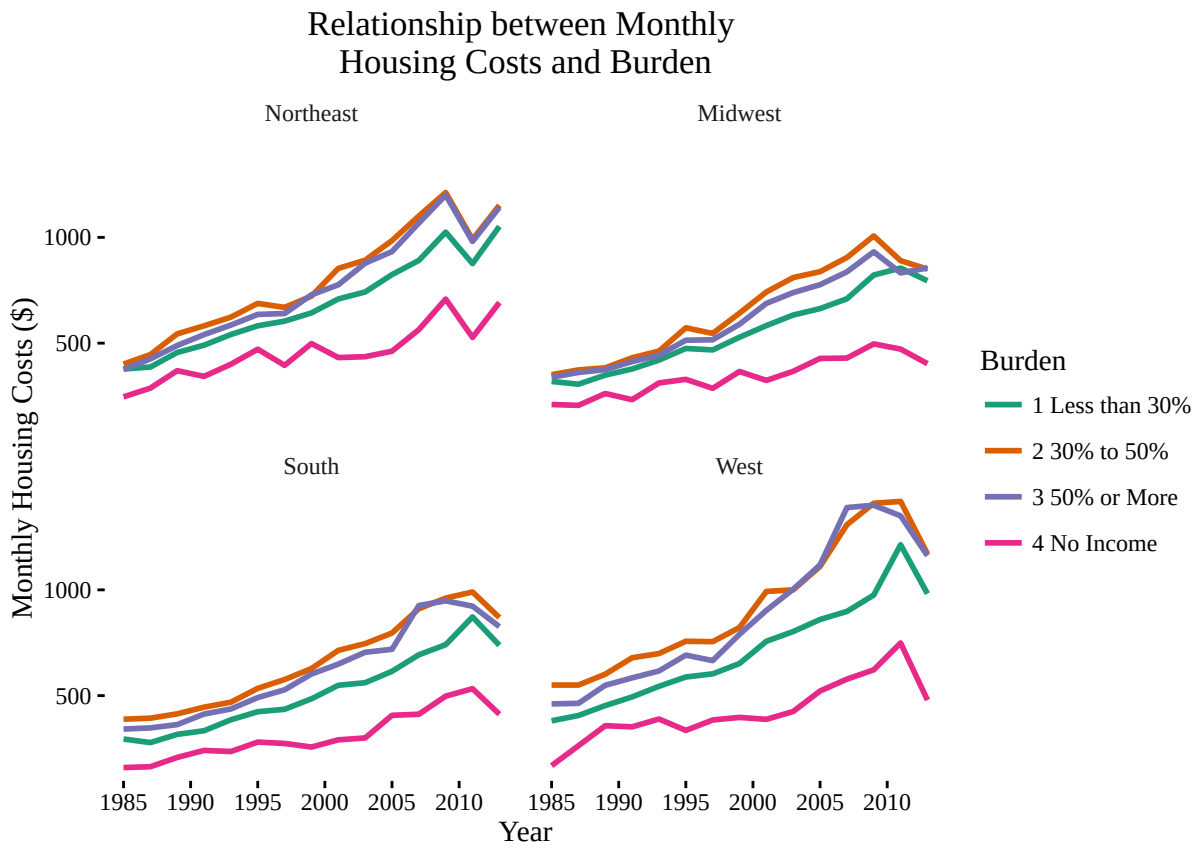

Relationship between Household Income and Burden
Northeast
Midwest
South
West
Household Income ($)
75K
50K
25K
0K
1985
1990
1995
2000
2005
2010
Year
Burden
1 Less than 30%
2 30% to 50%
3 50% or More

Figure 1:



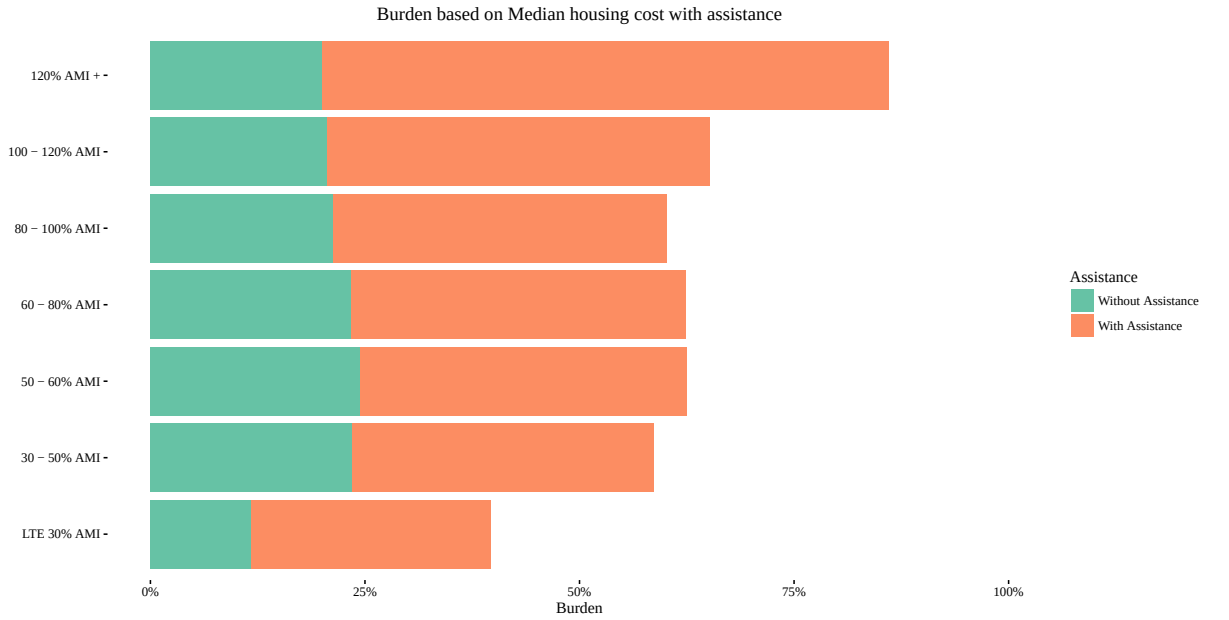
From this graph, you can take a more indepth look at the relationship between housing costs and burden. It's interesting to see that all three burden lines are close to one another, which illustrates the principle that those with more money aren't buying more expensive houses. From the previous grpah, we saw that those with higher incomes have less burden, and those with lower incomes have higher burden. Each level of burden has similar housing costs, therefore reinforcing the idea that people are buying similar priced housing among the lower and middle classes. Again, a spike around 2010 is prevalent, which can be caused by the 2007 housing crisis.



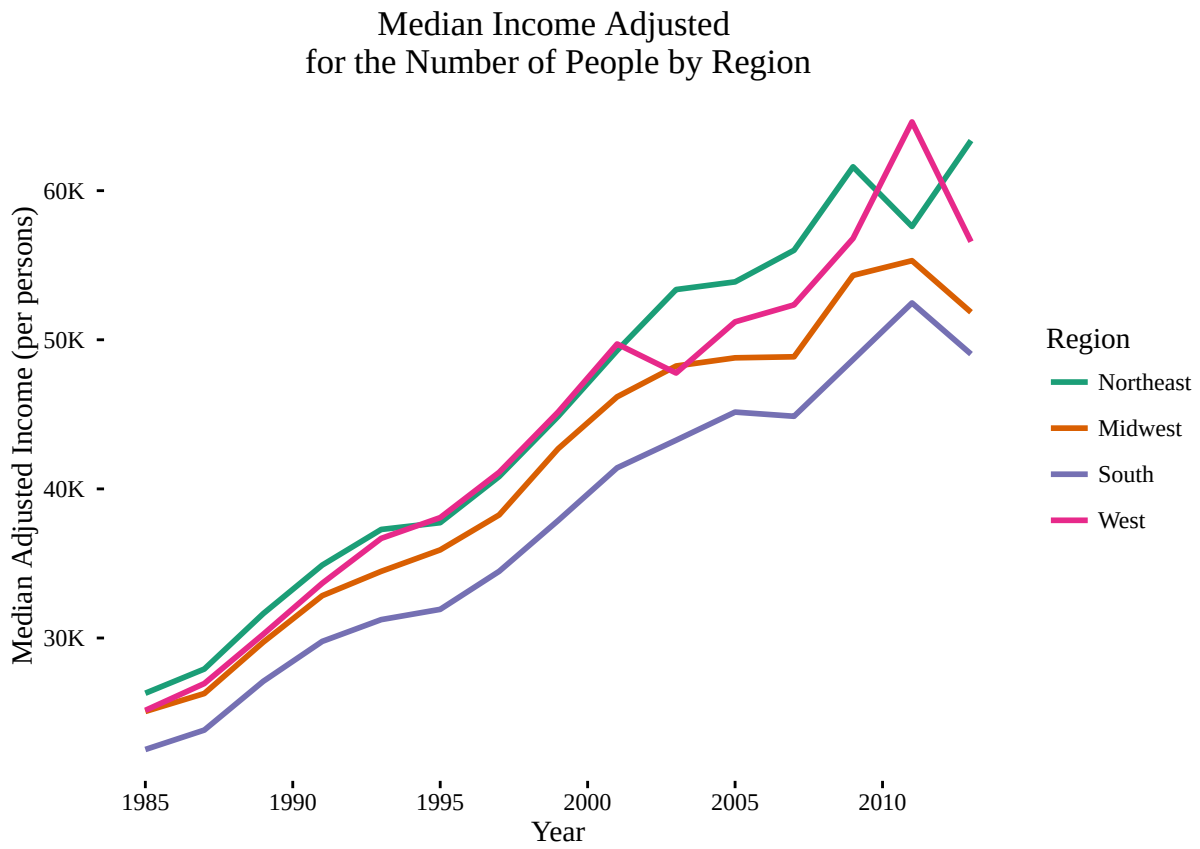
From this graph, you can see that the Northeast and West have higher rates of burden at higher income levels, therefore illustrating how housing affordability in these regions are less than that of the Midwest and South.

This plot charts a time series showing how the typical income for various types of households such as single family or apartment buildings changed over the years. One major trend is that typical income of all housing types except single family homes compared to their area has fallen, accelerating after 2007. However, incomes

of single family homes, barring the housing crash in 2007, have been centered on the area's median income. There are a couple of possible explanations for this disparity. The first is that, with the growth of wages, more people could afford single family homes and thus leave the buildings with higher numbers of units. As wage grew, the area's median income also grew, and so, combined with higher waged workers moving into single family houses, single family homes remained constant while all other types shrunk proportionally. This would also explain the acceleration after the housing crisis, as single family homes became much cheaper temporarily and thus allow more workers to obtain single family housing. The second explanation is that as time went on, supply of housing increased, driving prices down and allowing lower waged laborers to obtain homes. Due to the purchasing power of higher wage laborers however, single family houses, even with their falling prices, were more difficult to obtain as time went on, and thus lower wage workers had to settle for other types of housing.

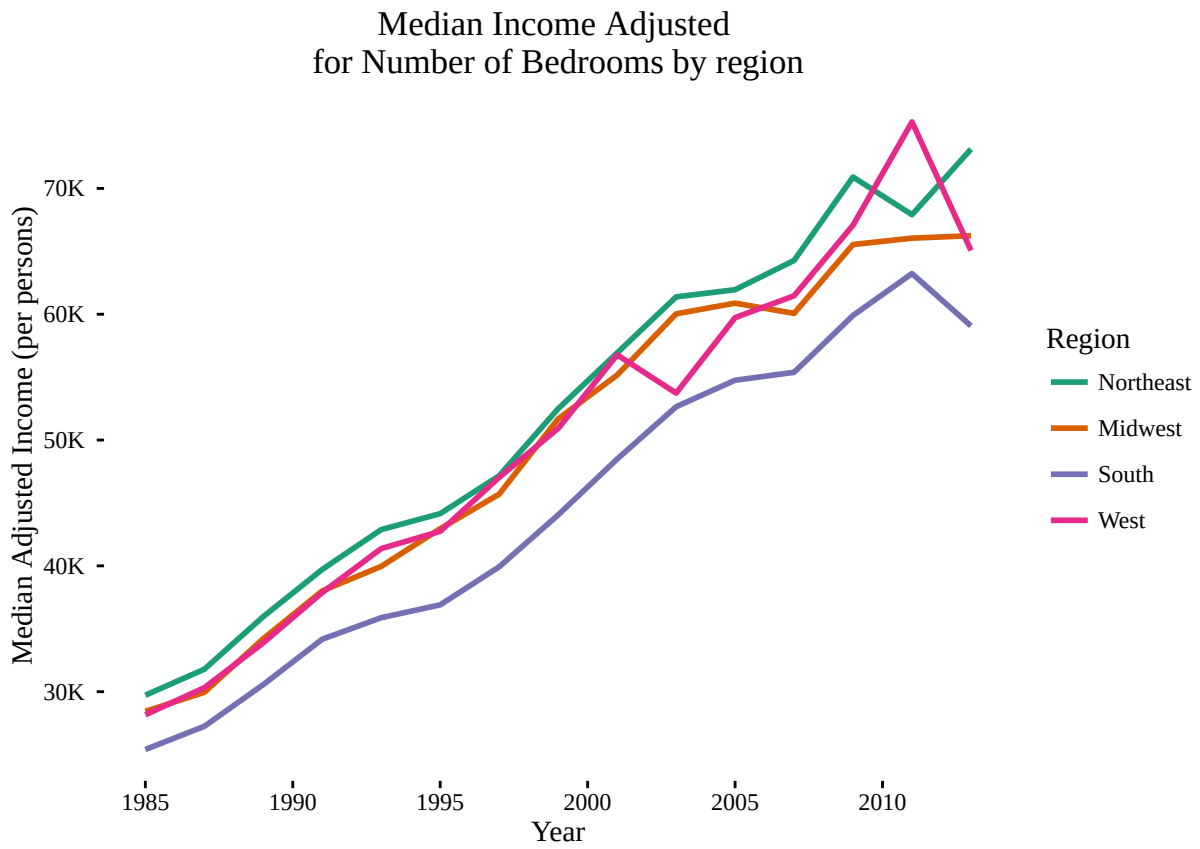


The plot shows how burden changes as a result of assistance based on what percentage of the area median income is their monthly housing costs. At initial observation, the burden of those who recieve more assistance is higher than those who recieve no assistance. This clearly makes sense, as those who need assistance are typically people who've lost their job or had some monetary crisis that would reduce their total salary. However, although assistance has around the same effect for most cost brackets, the households that have excessively high costs for their area have vastly higher burdens with assistance than all other brackets. One explanation for this is that those who purchase expensive houses in their area typically have higher salaries, and if they lose those salaries and require assistance they have to support the cost of the house. Also, single family homeowners have significantly less mobility so if they have a financial crisis they cannot easily transition to cheaper housing as opposing to those with rental properties typical of a larger number of units.



This plot compares APLmed by year for the four regions in the United States (Northeast, South, Midwest, West). It is clear that the Northeast and the West are the most expensive regions. This is makes sense given that the northeast and West are by far the most productive (by GDP) regions in the United States.

The median adjusted income increases steadily (roughly linearly) over time. There is a decrease in income around 2010, correlated to the economic recession with a 2 year lag between the Northeast and the rest of the regions.The reason for this lag may be due to the fact that the financial center of the United States lies in the northeast, around New York City. So, when the 2008 crisis occurred the repercussions were immediately felt in the Northeast (Wall Street and Main Street- which refers to the Washington D.C. economy) while the other regions took some time to correct for the downturn in the financial market.



The trends of the graphs here are similar to the ones in APLmed, however ABLmed signifies the median adjusted income for the number of bedrooms in a household's living quarters.

It is interesting to note that the disparity between the incomes increases between the Northeast/Western regions and the Midwest/Southern Regions. This is liekly because those living in the Midwest and South ten dto live in larger homes, housing more bedrooms on average, relative to their incomes due to the higher housing affordability. However, this should mean that their incomes go 'further,' as they have more purchasing parity given the lower cost of living compared to the Northeast and the West. This is not shown in the graphs, which indicate that the income levels are significantly different. However, if housing affordability was factored into the calcualtion, the income disparity would be far less if not almost nonexistent.

The lag in the downward trend after 2008 still occurs in the plots.

Conclusion:

Blah blah blah blah blah