Management 180 Final Report:
A Brief Examination of the US Housing Market

By: Louis Zhao

UID: 905066048

Datasets:
https://www.kaggle.com/goldenoakresearch/us-acs-mortgage-equity-loans-rent-statistics

https://www.consumerfinance.gov/data-research/mortgage-performance-trends/mortgages-30-89-days-delinquent/

Introduction:

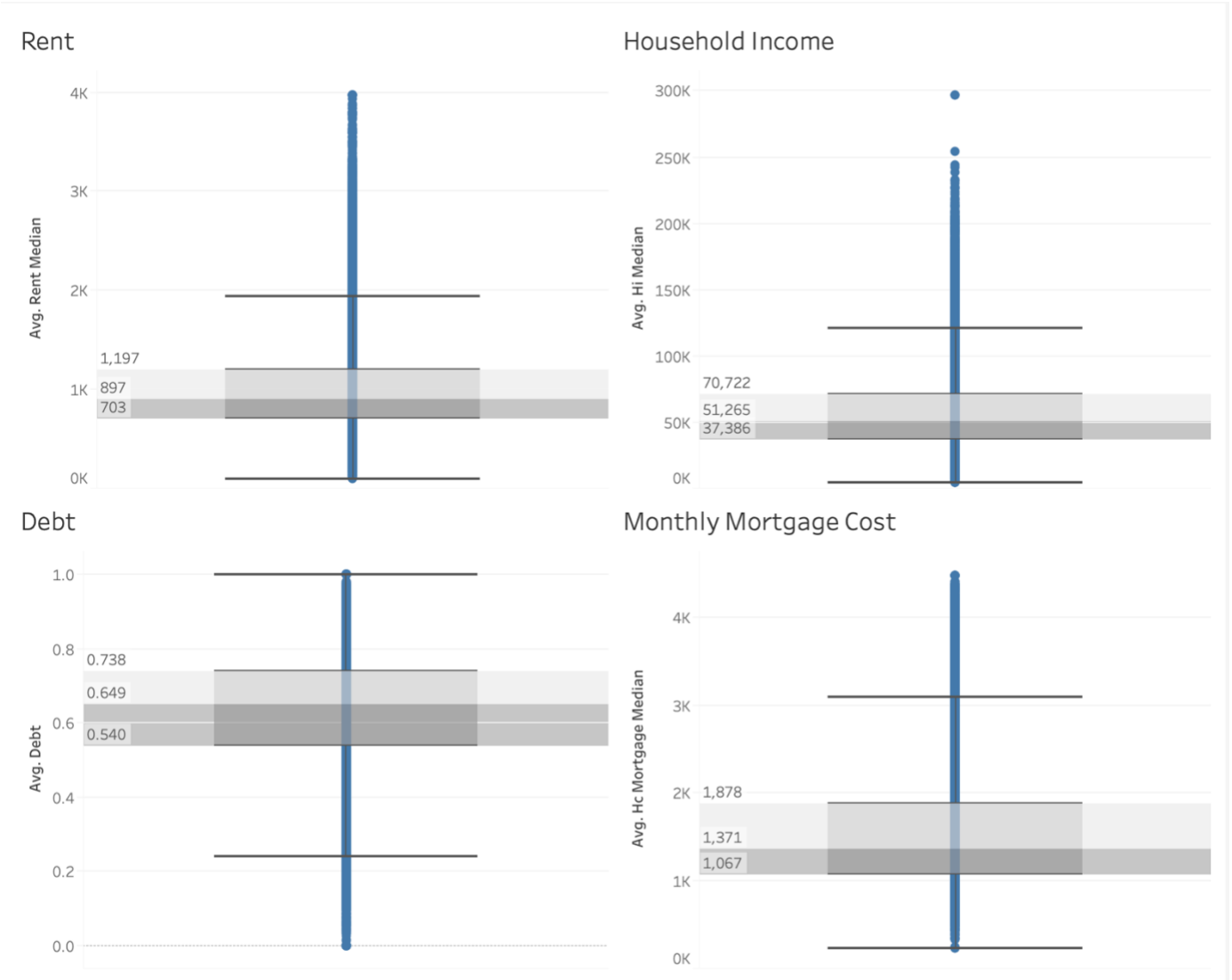
The housing market is an extremely integral part of the US economy. The industry does not just concern the financial processes of buying and selling houses, but it also affects a multitude of other markets from manufacturing to finance. For example, financial institutions such as banks and investment firms rely on exchanging mortgage bonds as part of their business model, while home renovations are an 800 billion dollar industry. Thus, when the housing market sees a downturn, more often than not, the US economy follows. This was shown in 2008 when the US's housing market crashed causing a global recession. What makes things difficult however, is that there are many different factors that affect the health of the housing market and in turn the health of the overall US economy, making it very difficult to analyze and predict. In this project, I wish to examine some of these factors in hopes of gleaning some insight into how the housing market functions as well as the overall health of the market itself in the past couple of years.

Housing Market Explained:

The housing market is really just the supply and demand economics of individuals buying new housing. Of course, the market is more nuanced than that, with it also including home renovations and housing services, but the essence of the housing market is extremely simple. However, when measuring the health of the housing market, it is not simply just how many people are buying new homes and how expensive these homes are. Because of the existence of mortgages and loans, people can technically own extremely expensive houses for a cheap price under the condition they pay the rest of the money over time. However, if they don't have the job security and income to sustain those mortgages, then they lose their homes and the overall health of the housing market goes down. (This was the case in 2008) Thus, when looking at this market, it is arguably more important to analyze the consumers that are actually buying these houses rather than analyzing the houses themselves. That is the direction I wanted to take when tackling this subject matter—what is the economic health of the US population.

Data:

As mentioned before, there is a plethora of factors that affect the conditions of the US housing market. From home sales to household income, they all play a part in deciding the trajectory of where the house market might trend. The main dataset I used in this report is from the 2015 US Census Bureau and focuses mainly on secondary variables that concern with the people that buy the homes rather than the homes themselves. Some key variables that the data provides include: the mean household income, mean monthly mortgage costs, percent of people sampled in debt, and mean monthly rent. The data is spread out between different area codes within different cities all across the United States. Although it is not representative of the entire US population, the dataset includes almost 40000 entries with over 21 million citizens surveyed. A key aspect to note about this data is that it measures averages within communities, not individual people. Thus it makes sure that the data is more resistant to outliers but still keep the overall trends the same. My second dataset analyzes mortgage delinquency rates across US states for the past ten years. I am using this dataset to not only provide context for my first data set but also examine broader housing market trends across the last decade.

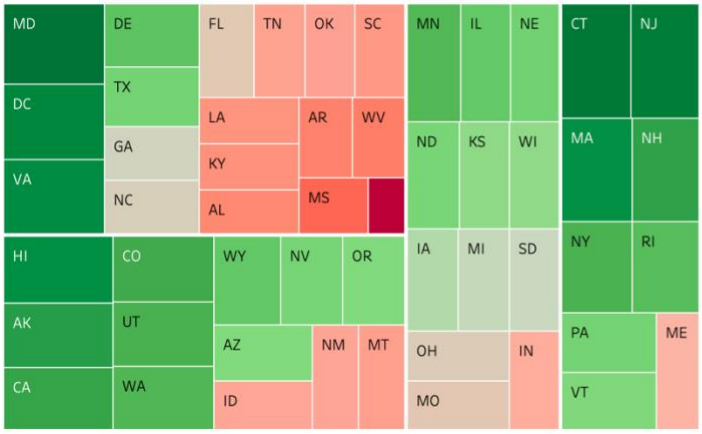


This dashboard is for preliminary analysis of the variables mentioned above. I deemed the four to be the most important and are used in further visualizations so I constructed four boxplots to summarize their distribution and look for any obvious trends or outliers. These boxplots were extremely helpful with creating new calculated field and more complex visualizations. Some interesting aspects to note:

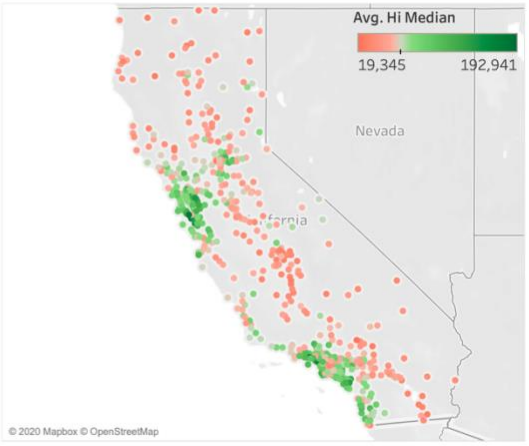
- All the data except for debt seems to skewed right, meaning that the mean will probably be pulled up causing it to be higher than the median
- Over 25 percent of the communities surveyed has a population with 100 percent debt, meaning everyone surveyed in those communities said they had some sort of debt. Also, on average 65 percent of people surveyed reported having debt.
- While rent and mortgage tend to cap out around 4000-5000 a month, mortgage on average is more expensive per month than rent

America's Geographic Trends:

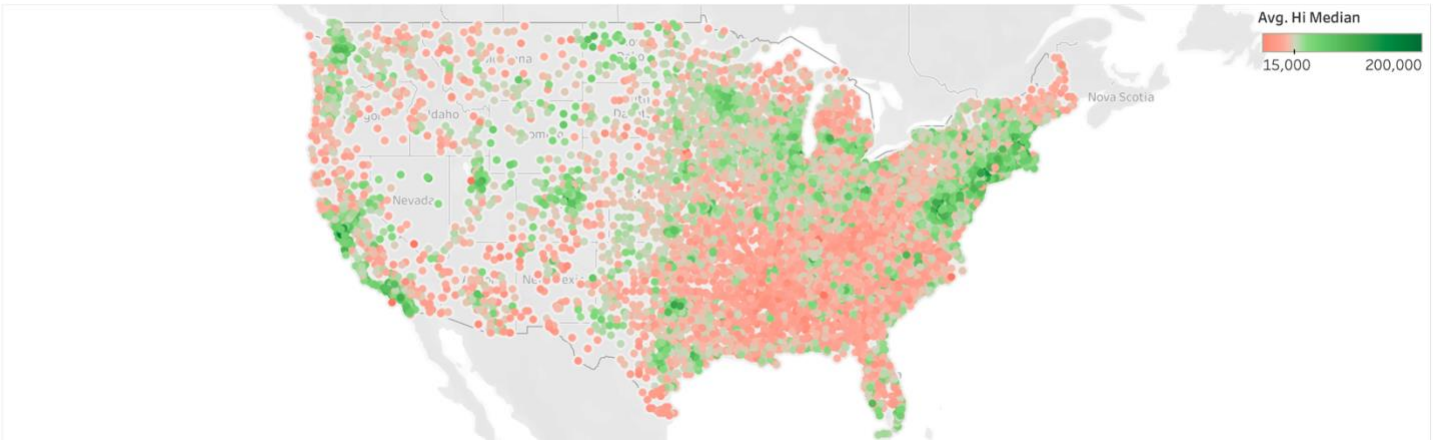
Income Levels By State



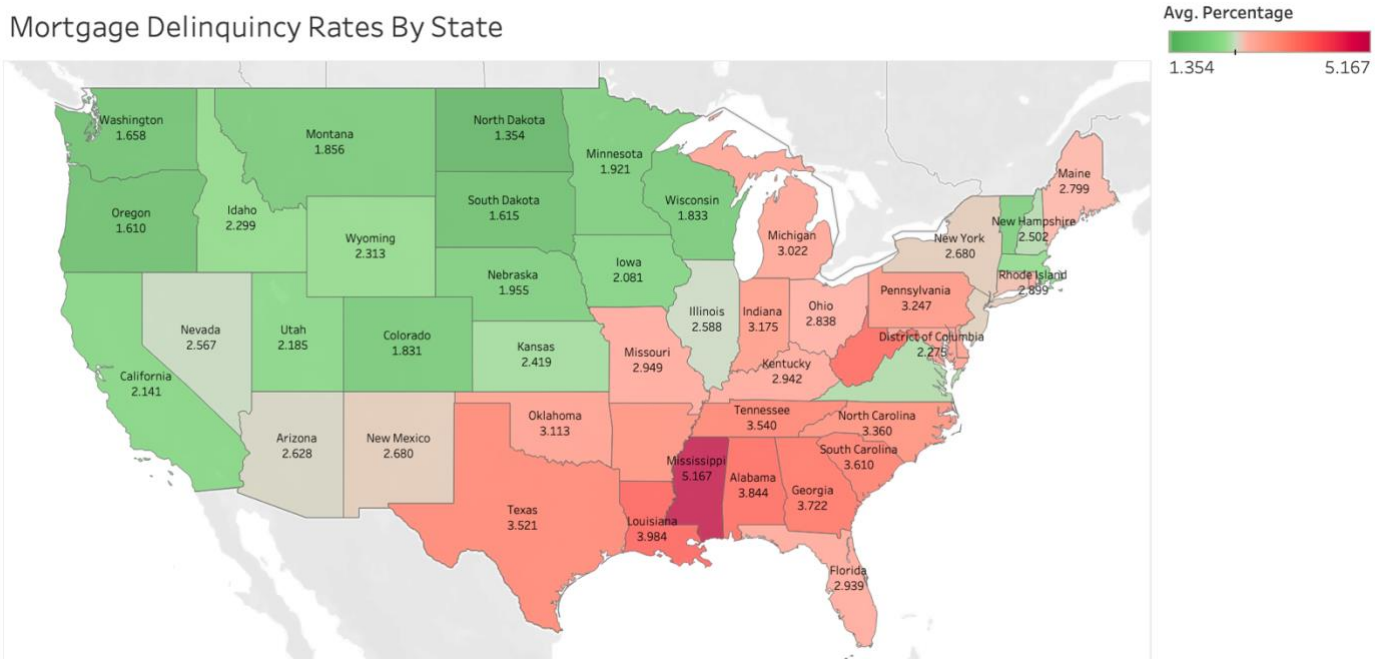
Income Levels Within CA

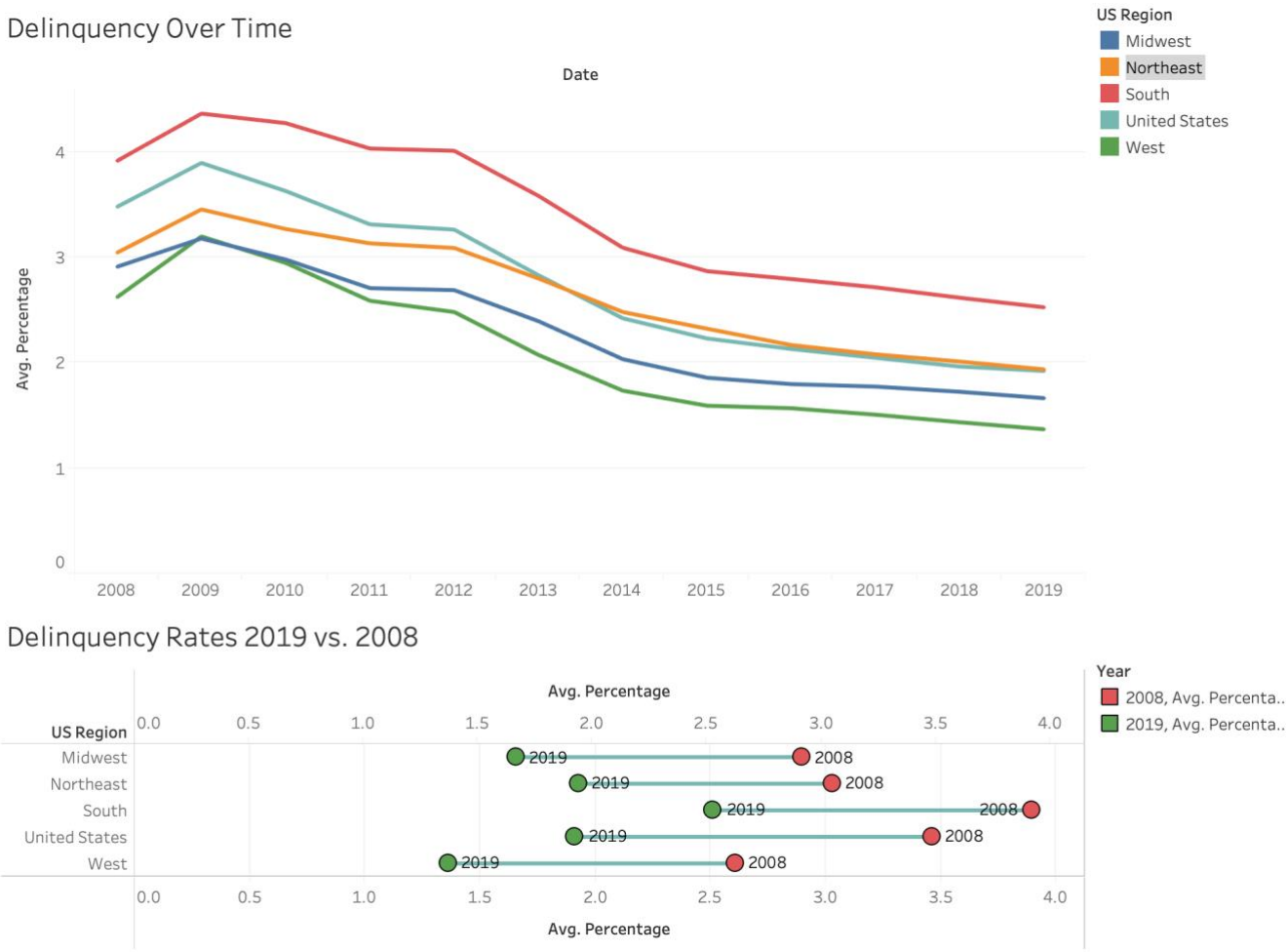


Income Levels By City



Mortgage Delinquincy Rates By State





The above graphs show median household income and mortgage delinquency rates throughout the US by region and state. Some key things to note about the data:

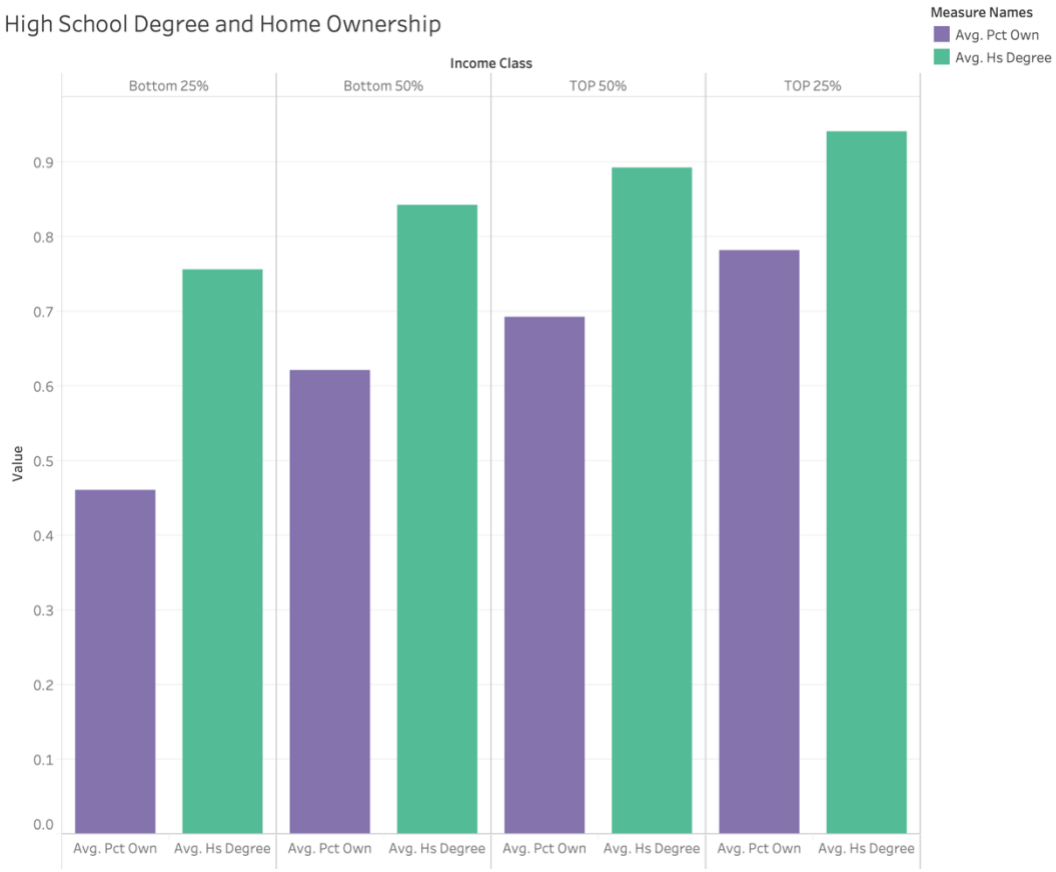
- Mortgage delinquency rates have trended down the past ten years. On average there has been a 1.5 percent decrease in the United States as shown by the barbell graph.
- Mortgage delinquency and income are correlated. From the map data on the page above, it is pretty obvious when comparing the two that the same locations have similar values between the two. This is extremely prevalent in Southern states where there is an abundance of lower income communities and relatively high mortgage delinquencies. However the two is not a one to one correlation so there must be other variables that affect their outcome

Income Levels:

Next I wanted to analyze the data in the context of each community's income level. I chose three variables to relate together, average income, average percentage of debt and average divorce rate. I determined their income level based off what percentile they were in within the population.

- As you can see from the left bar graph, the lowest level had a median income of around 28000, while the next levels were 44200, 60000, and 98000 respectively. The key thing to note about their incomes is the drastic increase from the top 50 percentile to the top 75 percentile. This is representative of overall trends within the United States where there is a substantial income gap between the rich and the poor.

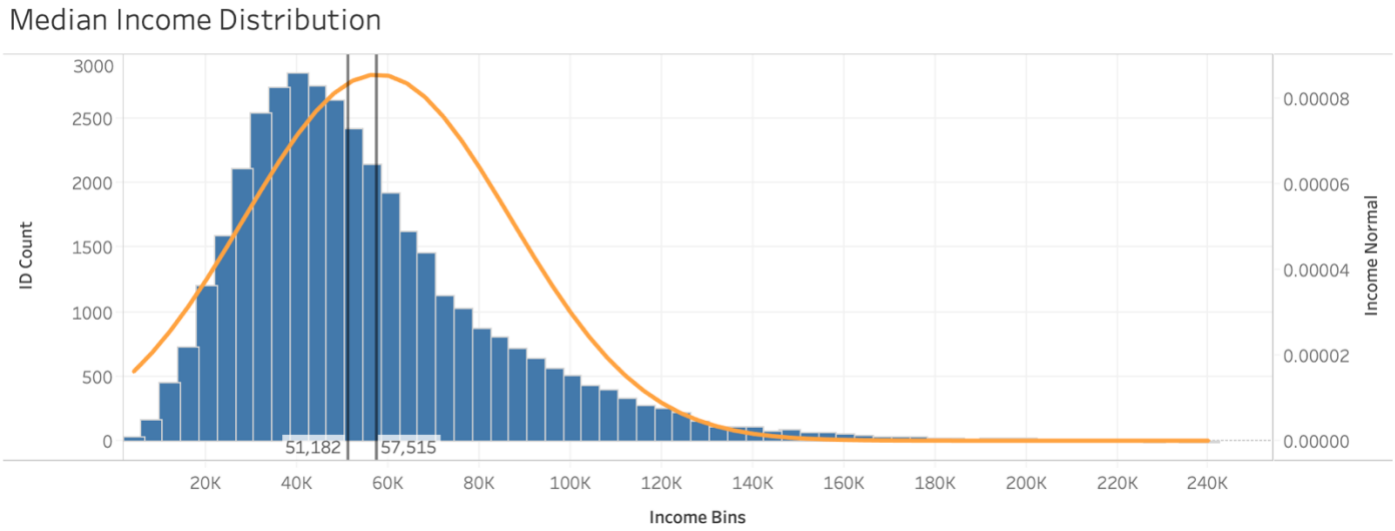
- The next graph analyzes the prevalence of debt within the four income classes. The general consensus is that debt is bad. Whether it be student debt or housing debt, nobody wants to have to owe money for something. However, this data paints a different story. As income levels increase, the number of people that have debt tend to increase. This is because of various confounding variables and societal trends that are not shown in this data. For example, people in the higher income categories tend to have a college degree. However, most people who go through college often come out with student loans. They might have really good jobs and earn quite a bit due to their college degree but they still have a debt to pay off. Also, people who earn more money can actually think about and afford putting down a mortgage on a home, rather than living in an apartment. Thus this is another reason why they might have debt hanging over them as they are in a contract pay off their mortgage or housing debt over the span of many years. Evidence for this claim can be seen in the following chart:

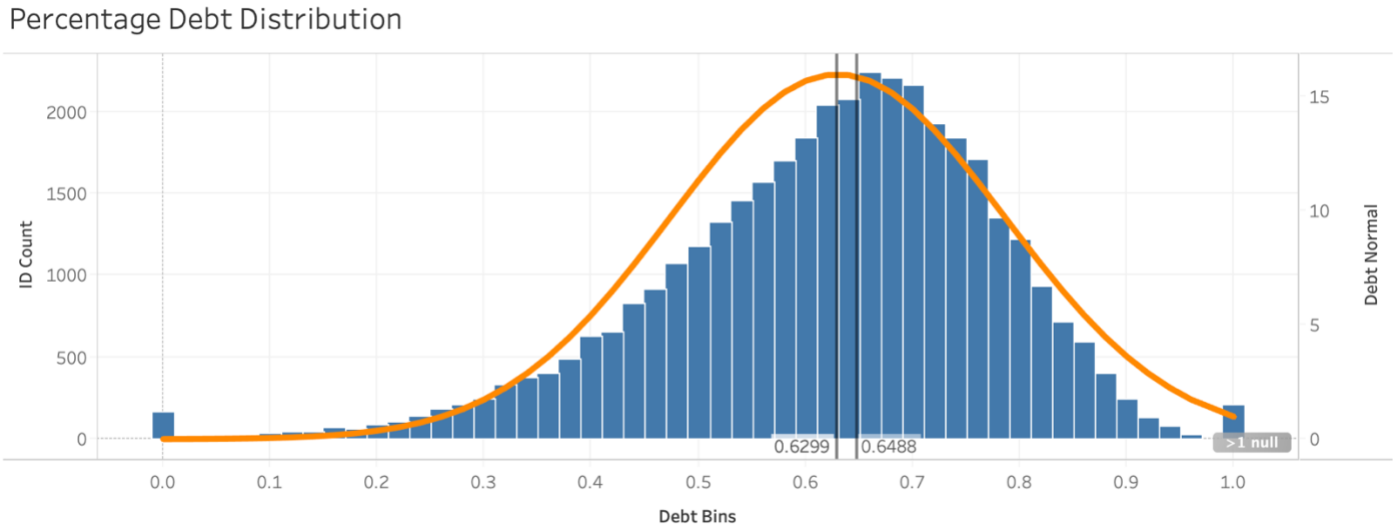


As income levels increase, there is a subsequent increase in percentage home ownership and percentage high school degree. As a result there should be a corresponding increase in terms of the prevalence of mortgage debt and college debt, leading to the results seen above.

- The last graph shows the effect income level has on divorce rate. I chose divorce rate because I wanted to give more context to this data outside of economic and finance variables. Divorce rate serves as a great secondary attribute because of its relationship with happiness. As you might expect, higher income leads to lower divorce rates. This can be extended to a higher income leading to overall higher happiness due to income and job security thus resulting in more stable and happier marriages.

Normal Distributions of Key Variables:





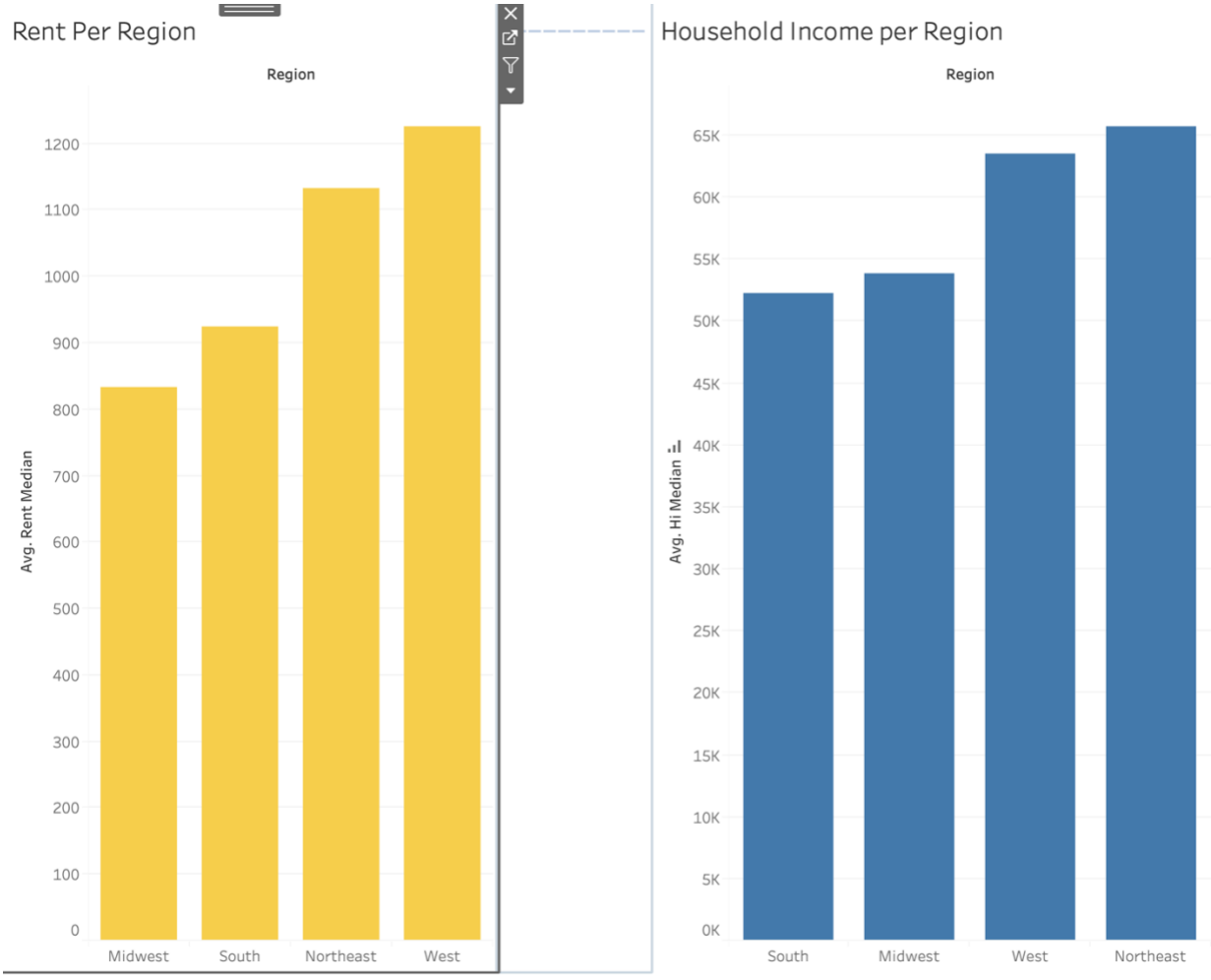
The data above shows the distribution of income and debt within the dataset. Some key elements to note about their distribution:

- Both are roughly normally distributed. The histograms plotted match the imposed normal curve. This makes sense in the context of the dataset as you would assume a large percentage of the population to be concentrated around the mean and median when it comes to their yearly income and percentage debt. This also means that a myriad of statistical tests (Hypothesis Tests, Confidence Intervals, etc.) can be conducted on these variables without worrying about having too high of a margin of error.
- As supported by the boxplots from the beginning of the report, income seems to be heavily skewed right while debt seems to be skewed left. This results in the mean being pulled away from the median. The mean for average income is higher than the median while the mean for percentage debt is lower. This is due to the shape of the distribution and the prevalence of outliers pulling on the center measures.

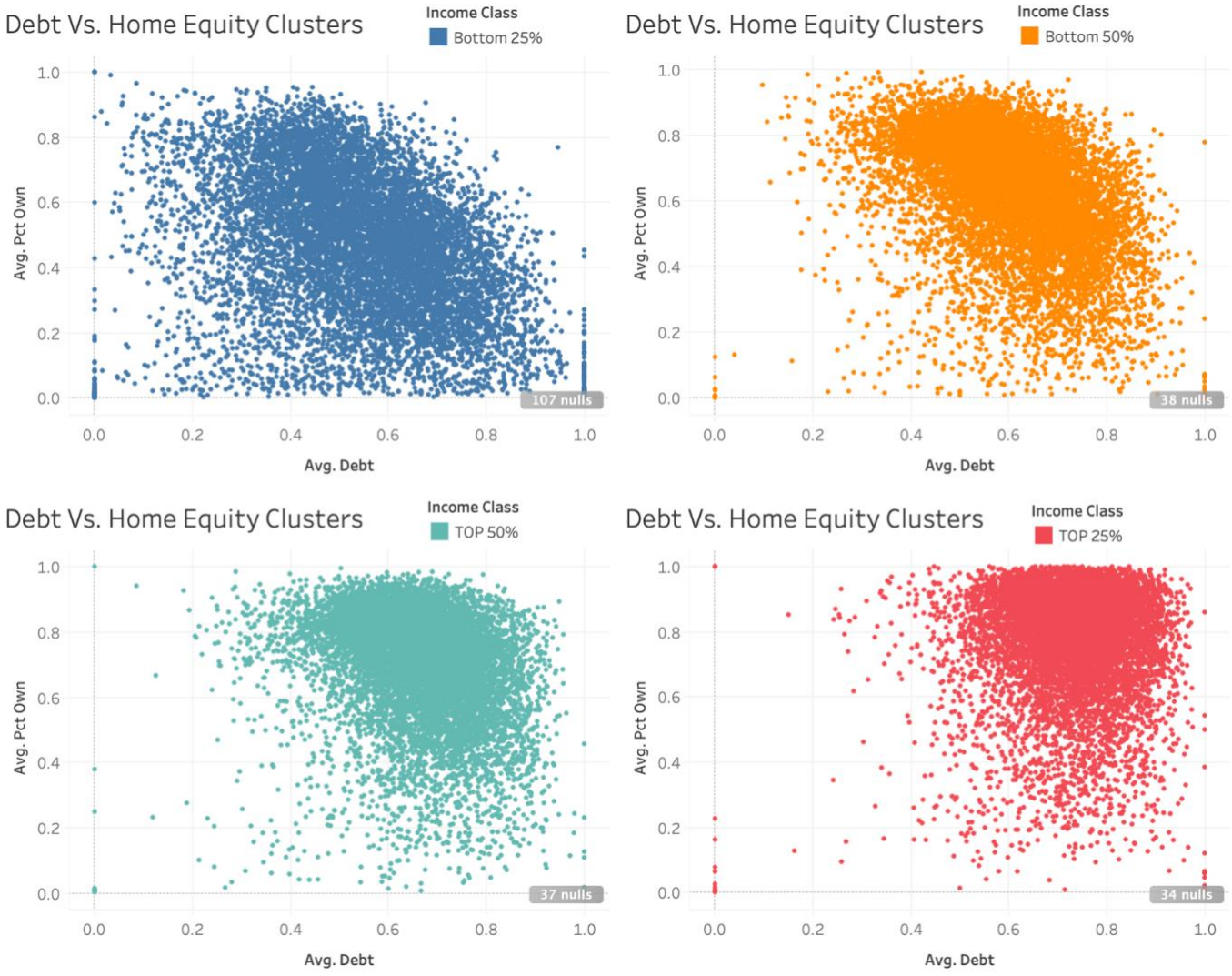
Machine Learning and AI:

Regression:

- In terms of regression, it's mainly used to determine if there is some sort of mathematical or statistical correlation between two variables. It can be easily applied to my data set. For example: take a look that this linear regression plot between a community's household income and the average rent. Although this example is a fairly simplistic comparison, the graph shows that there exists a linear correlation between the two values. Their R squared value which basically measures how correlated the two are is around .42 which means although it is not particularly strong, there is definitely some linear relationship between the two variables. This is shown the by slope of the regression line, which signifies for every $1 increase in median income of an area, there is on average, a corresponding increase of .8 cents on the areas average rent price. This relationship can also be represented by a bar chart, however, the exact mathematical correlation will not be evident:



Clustering:



- Another example is classification and clustering. On the dataset above, I have mapped the percentage of people who own a home against the percentage of people with debt. Not only is there somewhat of a linear correlation but when you examine each income category independently, you can see trends and clusters that define each section. As you move above income classes, the more concentrated the data becomes. Also ,each income class has their own distinct region where they are most defined. This defines trends and traits that are inherent to the communities within each income level.

Machine Learning and AI Applications:

I want to highlight three main applications this dataset has in terms of Machine Learning and AI. They are in real estate investment, finance, and government:

1. Real Estate Investment: Within real estate investing, the most important element is knowing the future worth of a property. Although most houses tend to trend up in value over time, getting the most out of your investment might be difficult considering all the different factors that might affect property value. Thus, if you can use this data set to train a machine learning algorithm to factor in all of these different attributes, from median income to prevalence of home equity loans, then you can accurately predict the future values of communities and invest in the right locations. Machine learning algorithms can also optimize your investments based on predetermined factors and criteria. For example if you only have a certain amount of capital to invest or have a bad credit rating, the algorithm can factor in these conditions to help you obtain a higher return on investment.
2. Finance: The other big implication is within the financial sector. Financial institutions can predict future mortgage delinquency rates by examining an area's current economic conditions and then take suitable actions in response. For example, if a machine learning model detects that mortgage delinquencies will increase in the upcoming years, banks can give stricter loans to ensure there will be less defaults. If the opposite is true, banks can give riskier loans to improve their profit margins.
3. Government: Lastly, this dataset can be used by the US government, both national and local, to determine which areas to focus on in terms of financial aid. If they can forecast the future economic circumstances within a city or area, they can take the necessary precautions to ensure the worst does not happen. In addition, machine learning algorithms can help the government optimize their aid, making the best use of their limited resources. Lastly, the data can help identify geographic clusters of economic data that might help with future plans for new housing or commercial developments.