Housing Trends in America

DSC550 Data Mining

Shaquiel Pashtunyar

Fall 2022

Currently we are facing a large financial issue in the US economy. We have inflation that has gone loose, interest rates rising, external effects of supply chains and war, and more issues occurring across the globe. All these problems are leading to instability in the economy but also instability in the housing market. Leading up to the year 2022, we saw the housing market go through incredible amounts of gains year after year. Now we have a recession looming, the question becomes, where will the housing market go next. Will it continue to thrive as the demand rose during the pandemic or will it come down to more controlled levels? This is an important question to answer for people interested in buying a property, struggling with all the instability we are currently seeing. Using data mining and regression modeling we want to look at what factors affect this trend.

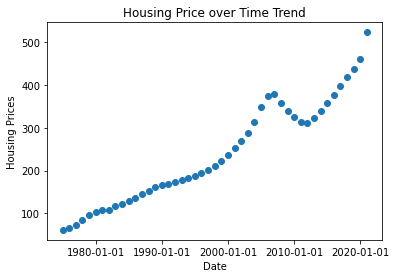
To take a look at the housing sector we first need to data that we think could be good predictors of the trends we see in the housing market. I searched Kaggle, an online site of datasets, to find economic trend data found over the last couple of years. Using this link: <https://www.kaggle.com/datasets/faryarmemon/usa-housing-market-factors> we see a list of factors and their trend over the last 46 years. Here is a snippet of the dataset and the various fields it contains.

Graphical user interface, text, application

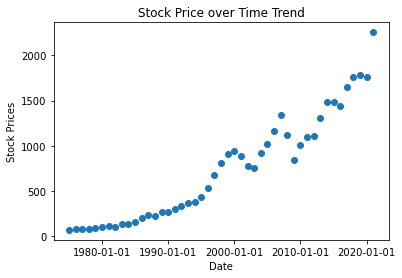
Description automatically generated

The main variable we want to look into is the house price index of America. Some interesting factors that are included into this dataset are disposable income, mortgage rates, GDP, stock price, consumer price index, unemployment, and population for the year. All of these are time series variables that also change over time.

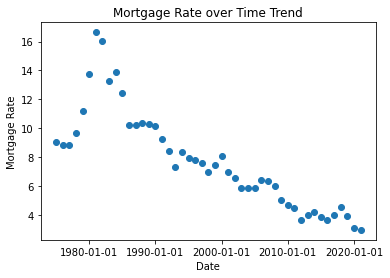
One more dataset that I went ahead and obtained was California specific housing data. I wanted to take my analysis a step further and also look at trends in California, a state with very high housing demands. This dataset was also found on Kaggle here: <https://www.kaggle.com/datasets/camnugent/california-housing-prices>

In order to start my analysis I needed to understand the data a little better. I went ahead and made some plots showing the time series data in action. The first was looking at the house price index in the US over time. What this showed me was that the housing prices generally went up over the years. It was largely stable until 2008 when we had a housing recession and we saw a 20-30% drop in the price index. 

I also wanted to check if the stock market, another commonly watched investing field, had the same pattern as the housing market. What I saw there was that it was a little more unpredictable but still largely similar, with the exception of the two drops, one during the housing recession as well as the dot com bubble in 2001.



The last variable I went ahead and visualized was the mortgage rates year over year. The assumption being that if the mortgage rate has an affect on housing prices, we would see it have a similar linear pattern. Instead we saw it go up, then down, and have little changes during the two recession events I mentioned above. This was a little surprising as I expected it to play a bigger role in housing prices.



Now with the charting out of the way, I wanted to find out with had the greatest correlation with the housing index. I would assume stock prices, and not mortgage rates. I ran a correlation study and here are the results.

Graphical user interface

Description automatically generated with low confidence

What we found was that the variables that had the largest correlation, or number closest to one, included the stock price, consumer price index, GDP, and leading the pack, real disposable income. Here is one last chart showing real disposable income in relation to the other variables.

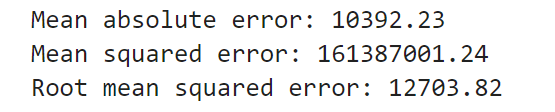


I did a similar study with the California data and decided to select the proximity to the beach as well as disposable income as my variables to study in relation to the housing price index.

With very high correlation values, and trends that arguably looked linear, I decided I would use a linear regression model to study my dataset and see if it can make predictions for housing prices using the variables present to me. I first had to do a bit of data wrangling before that. I removed all the variables with low correlation values from both data sets. I then cleaned the data by removing any rows with invalid or missing data points. I also did some clean up on geographic locations in my California dataset. Once the data was cleaned, I kept all the variables left and created dummy variables in the California data. The US dataset didn’t require this as I was using time series data for all of my inputs.

Once the data was cleaned, I split the dataset into training and testing data tables, with my target variable being the US housing price index. The training dataset would be what I would build my model off of, and the testing dataset I would use for verification. I chose the variable of real disposable income as my variable of interest from the main dataset. My assumption was that this was the most linear because it had the highest correlation value. Upon completion of this step, in python the Logistic Regression package was imported, and it created a model from the cleaned data.

I then took a score of the model by looking at the R2 value showing how good the model was. I got a score of 95% which indicates the model was quite good and had a 95% success rate. I also took other analysis into consideration like the Mean Absolute Error, which came back with 10,000. What this tells me is that I was able to predict the price of the home using the algorithm within 10000 for the year, this was great.



I repeated the study with the California data. The model had an exceptionally low rate of 8%. This may be due to me creating a model that wasn’t the most accurate or it could have spoken to the unpredictability of the California housing market and the location of the house. Either way, the second dataset was more of a learning experience and I got most of the information I needed from the first regression I ran.

So, what does this model tell me about the US housing market? Well with the high correlation of various factors, we do see some linearity across several of these variables. The one we see the highest correlation to housing is the disposable income. I want to say this makes sense. When purchasing a house, the most important thing you are thinking about is your monthly payment. This is what banks base their approval or rejection by in your loan. If you have more disposable income or come from a more affluent area, you will see the house price go up as you can take on a bigger mortgage in that area. In times of recessions we see individuals have to cut back on their extra income and just have enough to buy the essentials, this correlates to a drop in home prices, as individuals can’t take on larger monthly payments.

Another curveball is the 2020 pandemic. We printed a ton of money during that year and handed it out to individuals, this free money increases the disposable income of all Americans by a large amount. This is where we see the largest spike in home prices. Now in the year 2022, we have inflation on the rise, so we are taking money out of the Americans pockets, translating to a drop in home prices. All of these real world examples speak to the truth of using disposable income as a home price identifier.

As far as my model goes, I would say it is nearly ready for deployment. It had a 95% accuracy rating. Another test I would like to do would be to check the disposable income levels by zip code, and then the home prices in relation to this. Since my linear regression was accurate for the median home price in America, I would say there are significant differences by zip code, and it could lead to a better model. My theory is that more affluent neighborhoods with higher home prices would also have more disposable income and vice versa.

This entire project was a huge learning experience for myself. The ability to find and clean data, prepare it for modeling, and then perform a model was the first full length project like this that I have done. If I were to go back I would do similar models using the other high correlation variables like GDP, Stock price, or the consumer price index. My guess is that all of these would also have high R2 values, but I doubt any would be as high as disposable income. It also tells us a lot about the greatest lever we have on the housing market prices, and the other economic trends in question. I am sure economists across the country can replicate this basic study and use this regression model to make insights on the spending and bills they choose to pass in any given time frame. It is unfortunate the California dataset didn’t yield any compelling results, but I am sure if I redid the calculations by geographic location, I would have a model similar to the one I got for all US data.