Joshua Shleifer

Data 205 CRN 22017

Final Report

Done with help from the CFPB

The Changes Behind Real-Estate Pricing Trends

Introduction

The overarching goal of this project is to analyze single-family property values and how the surrounding factors are influenced by changes in value. This is a particularly important topic to me because a major part of the America dream is for a person to be financially self-sufficient, but due to rising real-estate prices, it is looking less and less likely that every hard-working individual will reach a point where they can buy a house. As Alexander Hermann and Peyton Whitney found, the ratio of the median sale price of a single-family home to the median household income has consistently risen over the past few years and in 2022 was sitting at a record high: a whopping 5.6. This trend is slowly bringing homeownership, a goal that used to be a standard achievement, out of reach. Due to this, I decided to research what is causing these changes and how policies can be shaped to help make this dream a reality again for all those who reach for it.

The Data

To do this, I used three datasets: the HMDA dataset, the ‘citizens connect’ county dataset from the Bureau of transportation statistics, and residential construction permits dataset. The origin of the HMDA public dataset comes from the Home Mortgage Disclosure Act which requires lenders to submit records of their mortgage loans, and some of that information is released to the public by the CFPB. Note that there was a significant policy change in 2018, and the list of variables collected and released was altered with it. Due to this, I only used the data from 2018 and on as property value was not released for earlier years. There are also several different versions of the data for each year (snapshot, 1-year, and 3-year). My understanding is that the latest one is generally most correct as corrections are made over time. Therefore, I used the latest version available for each. The variables I will be working with are as follows: property value, loan purpose (home purchase vs improvement), county code, derived race, derived ethnicity, action type (filter for only applications to avoid double counting), and number of units (to separate apartment buildings, multifamily dwellings, and single homes when looking at property value).

The other two datasets were linked to add an additional point of analysis. The Residential Construction Permits dataset contains the number of final permits for residential home improvement by county, is created from the State of the Cities Data Systems (SOCDS) and is maintained by arcgis.com. I am interested in the columns for Geo ID and the number of permits by dwelling type for the years 2018-2022 to match the HMDA years that have the information I need. The specific variable I used was the single-family property building permits, as I filtered the HMDA dataset for the same to avoid comparing apples to oranges (such as the value of an apartment building with that of a small home).

The Citizens Connect dataset contains estimated commuter numbers per county and how they commute (e. g. by car, bus, etc.), as well as estimated demographics and median household income by county. I used the commuter data to attempt to model if people are moving into towns and driving up prices due to jobs in nearby areas. Median household income was used to show the ratio of loaned property values to income in each county. Unfortunately, this version of the dataset appears to be missing information for demographics. Their website says that they will be updating and redoing that work soon and adding an interactive dashboard so that may be a worthwhile avenue for further research.

Tools Used

For the analysis of this data, I used a variety of tools including some that do very similar things but I prefer different ones for different situations. For loading the datasets, I used the bash scripts from the HMDA data science kit for downloading and setting up a SQL server and modified them to only load from 2018-2023 and to load the most recent version of each dataset (as of November 2024). The other two datasets are included in my repository for this project as I wanted to preserve the exact versions I was using so my results could be replicated. All the analysis and visualizations were done in python. The packages I are were pandas for handling the datasets, ast to handle dictionaries (because pandas stores dictionaries that are in a dataset as a string), numpy and scipy for statistical analysis, and matplotlib, seaborn, and plotly for graphing.

Data Cleaning and Processing

The HMDA dataset is a list of loan actions taken with each row having all of the information pertaining to that loan. To get the information that I wanted to look at by county, I pulled each of the events where the property value is given (which excludes where the loan does not depend on property value), where the property is a single family home, and where the action taken is loan originated and derived the following variables grouped by county code: mean property value, median property value, the count of each race who applied for a loan, the same count normalized, and the number of entries for the county. From there, I removed the counties that have fewer than 25 entries as I had found many counties that seemed to be outliers to due small sample sizes. I calculated the median increase as a percentage of the previous year for each county (e. g. the median increase for a county in 2019 is its median value in 2019 divided by its median value in 2018). I then used a z-test to evaluate if each median is statistically significantly different from the rest and a chi-squared test to evaluate if the loaning demographic changed from the previous year. I also used the same test on the citizens-connect commuter information to see if there was significant change there. With that, I had everything I wanted to look at so I moved forward with my analysis and graphing.

Overview Analysis

The first thing I looked at is how real-estate values changed from year to year. These values are a proportion of the previous year, so 1.1 means a 10% increase over the year.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Standard Deviation | 25th percentile | 75th percentile |
| 2019 | 1.080033 | 0.087094 | 1.041667 | 1.117647 |
| 2020 | 1.153442 | 0.097561 | 1.098361 | 1.193548 |
| 2021 | 1.068355 | 0.084272 | 1.00 | 1.115942 |
| 2022 | 1.075672 | 0.093195 | 1.00 | 1.131148 |
| 2023 | 1.029258 | 0.075076 | 1.00 | 1.067308 |

Two things immediately jumped out at me from this table: an extreme increase in 2020 and a much lower increase in 2023. 2020 seems easier to explain to me as it coincides with two rather large events: the election of Joe Biden the outbreak of COVID-19. The fact that this led to an increase in prices is a bit surprising as the rest of the economy was in recession and historically housing prices usually fall in a recession (McCord), but this may indicate that due to the pandemic, fewer people wanted to move so people weren’t selling houses. This is shown by the fact that in May of 2020 right by the start of the pandemic, home sales hit an all-time low (Applewhaite). The lower increase in 2023 may be the logical conclusion of the previous price jumps during a recession: with record high prices and high interest rates, most people can’t afford to buy, leading to lower demand.

To avoid cluttering this section with graphs, I will describe some of the graphs I had and what they may indicate. The full graphs can be seen as part of my repository for this project.

The next thing I looked at was if certain areas experienced significant change in their lending demographics. Most of the United States seemed to have a fairly small change in their lending groups with a few single counties exhibiting change but there were four larger areas that looked notable to me: namely Southern California, Northern Minnesota, South Florida, and the New York/New Jersey/Delaware area. Interestingly they also shared the trend they were following. It appears that they experienced huge changes in demographics in 2018-2020, but the changes slowed down after that with both the effected areas growing smaller and the changes becoming less significant. To me, this indicates that there was a change in the lending demographic, but the areas are now settling into a new normal.

One more visualization that I think helps give context to the property value problem is the ratio of median property value to median household income by county. Below is the graph for 2018. (Note that I only have the graphs for 2018 and 2019 due to limitations with the citizens-connect dataset.)

A map of the united states

Description automatically generated

This looked pretty much exactly like I expected. The more typically rural farmland has a lower value to income ratio, the stereotypically urban areas sit higher at closer to 5-7, and the vacation spots were generally between 6 and 8. Only the cities that are generally known as extremely expensive deviated significantly from this, such as New York City at close to 15.

Final Product

My original plan was to find which of the variables I though may be significantly contributing to the price change actually were and then create a model with those variables. However, despite testing a multitude of variables, I did not find any that significantly correlated with either the statistical significance of the change in value or with the change in value itself. There were a few individual years that had a very low p-value for the correlation, but the r-squared and the correlation coefficients were too low to show any meaning in the relationship. I think this exhibits two things: that over 85 percent of data science projects are thrown out before achieving a final model, and that sometimes there is meaning in the lack of a relationship.

(I don’t see much value in displaying the results here as they were all insignificant. If you are curious, they can be found in the project repository.)

The variables I tested for significance included the p-value of the change in the lending demographic, the current value, previous increase, the number of construction permits, and the p-value for a change in commuters from an area. These are all things that I would have expected to have at least some relationship, but the fact that they don’t indicates that they are not good descriptors for the likelihood of coming property value jumps. I found this very surprising, as these are many of the things generally pointed at when talking about gentrification, which is a community being forced to move out due to the members no longer being able to afford living in the area. This shows that the causes of price jumps what we can do to prevent them is not as simple as controlling a few numbers, but instead requires research into how changes may effect a community.

Potential Next Steps

There were some things that I was unable to look into for various reason, but may be good avenues for further research. For one, the number of residential building permits is very heavily skewed by outliers due to natural disasters. There is a dataset by FEMA that could be used to filter out those that qualified for rebuilding to exclude those from the calculations. Another interesting thing would be to look at if the lending demographic matches the demographic that lives in the area. Unfortunately the citizens connect dataset was mislabeled and missing information, but that information may be able to be derived from the census tract data or may be fixed in the future. Lastly, my mentor at the CFPB said that one major contributor to real-estate prices is zoning laws and other legislations that effect the supply. If that could be adequately geographically modeled, it may show how different laws effect prices, which is something that can be directly controlled by those creating policies to manage housing and living costs.

Acknowledgments

I would like to take a moment at the end of this paper to thank the people who I would have been unable to do this without. My mentors at the CFPB were invaluable in helping me make sure that I was addressing the problems with the correct information. David Roell was a huge help meeting with me every week ad making sure I had direction with the project and that the final product is accessible to those not necessarily already familiar with the field. Le Quyen and Alex Nongard were on call to help me with anything I needed. My classmates’ feedback helped me shape the project into something that answers an essential question. And Professor Valentine, who taught me the skills that made this whole project possible and taught us how to create something that can help change society as a whole for the better.

Sources

Datasets:

HMDA public: <https://ffiec.cfpb.gov/data-browser/>. Note that this can also be loaded using their data science kit found at https://github.com/cfpb/HMDA\_Data\_Science\_Kit

Citizens Connect - County Data: https://data.bts.gov/Research-and-Statistics/Citizen-Connect-County-data-live-/t3kh-5nek/about\_data

Residential Construction Permits by County: <https://hudgis-hud.opendata.arcgis.com/datasets/HUD::residential-construction-permits-by-county/about>

Articles:

Applewhaite, Kenneth. “Redfin Reports Homes Sales Fell to One of the Lowest Levels on Record in May.” *Redfin Corporation*, 14 June 2024, investors.redfin.com/news-events/press-releases/detail/1120/redfin-reports-homes-sales-fell-to-one-of-the-lowest-levels.

Hermann, Alexander, and Peyton Whitney. *Home Price-to-Income Ratio Reaches Record High*. 22 Jan. 2024, www.jchs.harvard.edu/blog/home-price-income-ratio-reaches-record-high-0.

McCord, Roisin. “The Housing Market and the Pandemic.” *Richmond Fed*, [www.richmondfed.org/publications/research/econ\_focus/2020/q4/district\_digest](http://www.richmondfed.org/publications/research/econ_focus/2020/q4/district_digest).