Projecting US Home Costs

(COMP3125 Individual Project)

Anthony Ferrucci
School of Computing & Data Science

Abstract

Housing costs in recent years seem to have skyrocketed far beyond ordinary levels. This has stirred unrest among homeowners and renters alike, who must front these necessary expenditures. Existing research into the current housing crisis fails to consider the immediate needs of consumers, instead relying on sentiment and day-by-day price movements. This report aims to use market data to reveal the extent of the crisis, predict movements in the next several years, and offer solutions for consumers looking to purchase now.

I. Introduction

Housing costs in the United States have increased drastically in the post-Covid market, seemingly far beyond the rate of decades prior. This report seeks to determine whether recent home prices are, in fact, abnormal compared to past decades. Regardless, astronomical home prices have left many Americans wondering if, and when, the market will rebound [1]. Analyzing real estate data from recent years may offer insight into future projections for home costs, as well as predicting anomalies in the market. Home prices have also differed greatly between regions, causing millions of Americans to move. This report aims to identify areas with the cheapest home costs, and how these regional housing markets compare. Prior research conducted by Bankrate concludes that the rate of increase for shelter in the US has been plateauing, although 77% of Americans hold negative opinions about the current market [2]. This overwhelming sentiment demands a conclusive analysis of how uncommon these prices are, how they will behave in the future, and immediate affordability.

II. Datasets

A. Source of datasets

The Federal Housing Finance Agency (FHFA) offers a large dataset of Home Price Indexes (HPIs) via data.gov [3]. Records are from all 50 states between 1975-2022, just after the pandemic. Data was collected through the transactions of the nation's two largest mortgage handlers, Fannie Mae and Freddie Mac. Indexes are calculated over this broad period by using the Weighted-Repeat Sales technique. This method tracks the cost of an individual home over the course of several sales, providing accurate nominal figures.

In order to gain insight into relative historical costs, these indices must be adjusted for inflation. This requires a Consumer Price Index (CPI), which tracks the affordability rather than the investment value of a property. CPI data is collected by the U.S. Bureau of Labor Statistics, made available by the Federal Reserve Bank of St. Louis. [4]

Trends in the recent market require more detailed data, spanning each region. Realtor.com, one of the most popular real estate listing sites, maintains up-to-date information about millions of homes across the US. This data is made available by a researcher from the University of Waterloo via Kaggle [5].

B. Character of the datasets

HPI data obtained from the FHFA spans a 49-year period, with over 128,000 quarterly insights into the value of single-family homes. It contains seasonally and non-seasonally adjusted (SA/NSA) indexes. The SA feature is particularly useful, since it is resilient to cyclical changes in the housing market that exist year-over-year, exposing annual trends rather than seasonal peaks. A special characteristic of these indexes is that they are relative. In this dataset, an HPI of 100 is arbitrarily chosen for the year 1991 – this serves as the baseline comparison for all other rows. For example, an HPI of 58 in 1970 would indicate that prices in 1970 were 42% lower than in 1991, whereas an increase in HPI indicates an equal percentage increase in price. A single null value was replaced with a linear interpolation for the 2022 index, so that projections can make use of the most recently available HPI data.

CPIs are observed from 1953-present, with updates made monthly. This dataset considers the average affordability for all urban shelter, so it does not use individual properties for adjustments. For this reason, there are only 867 rows – one for each month since 1953. Like the HPI, this index is relative – it arbitrarily chooses the period from 1982-1984 as a baseline for the 100 value. This dataset is also seasonally adjusted, meaning it is compatible with HPI trends. No cleaning is necessary for this dataset.

The realtor dataset contains over 2.2 million rows of listings, with each specifying a price, details about the home, and location. Some major outliers exist in the key features of this data, such as homes with an upwards of 80 bedrooms. These points were removed from analysis by dropping the 99.5^{th} percentile from the set, which enabled some reasonable high-end homes to remain. Properties lacking vital information (i.e. a null value for number of bathrooms) were also dropped – this ensures that most datapoints are residential homes and allows for seamless analysis.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

III. Methodology

A. Line Chart

Indexes must be normalized to the same epoch, such that they measure the same relative change. To normalize the data, all rows in the HPI column were multiplied by a constant scale factor (1). This adjustment considers the relative change of any given year to the desired baseline. 1984 was chosen as the desired baseline, as this will normalize the HPI data to the CPI data.

$$HPI_{Normalized} = HPI \times \frac{100}{HPI_{Baseline}} \quad (1)$$

HPI and CPI trends were then graphed to the same figure using Matplotlib. Any dates outside of the overlapping range were dropped, such that only common years for the two datasets were being compared. An advantage of this approach is that any correlation between HPI and CPI trends will be exposed, and they will be appropriately aligned for direct comparison. Specifically, the normalization will allow consumer affordability to be compared to market rates for any given year. A drawback of this approach is that the direct comparison is only valid for comparing urban shelter affordability against national housing price averages, since these are the populations that the datasets sample for. Regardless, trends here will indicate broader movements in the market.

B. Linear Regression

A linear regression model is a simple method used to predict a dependent variable based on an independent variable. In order to use linear regression to project future indexes, it must be assumed that the year (independent variable) and index (dependent variable) are in a linear relationship, and that each datapoint is independent.

The advantage of this model is that it will, if the relationship is truly linear, predict market trends over the next several years. By constructing lines of best fit for each index, the ability of consumers to keep pace with prices will also be predicted. Furthermore, a residual chart can be used to test the validity of the experiment. If the difference between predictions and actual datapoint is normally distributed when plotted, then the regression model is valid. Numpy's polyfit function provides the functionality necessary to implement this regression model. Residual plots can be created by comparing sample data to the model, and plotting the difference between expected and actual indexes using PyPlot.

C. Map

Maps are useful graphics for visualizing geographical information across an area. The realtor dataset contains millions of properties along with their state, city, and ZIP code. Simply mapping housing prices is not enough, however. An auxiliary feature, such as cost per square foot, is often used to compare the value of properties, since it relates the overall cost to an aspect of that listing. This is a useful metric, since homes in an urban area may cost significantly more per square foot than in rural regions, but the overall price between a small and large home might be the same.

A benefit of using a map is that it will identify regions with the cheapest housing costs, and it will identify the extent of the urban-suburban-rural pricing disparity.

Creating the map requires geopandas, a version of pandas for parsing geographical data, and the latest version of Matplotlib. Since the dataset contains ZIP codes rather than coordinates, region boundaries must be defined by geojson files – these exist open source for each individual state. For this analysis, only a map of Massachusetts is used.

IV. Results

A. Line Chart

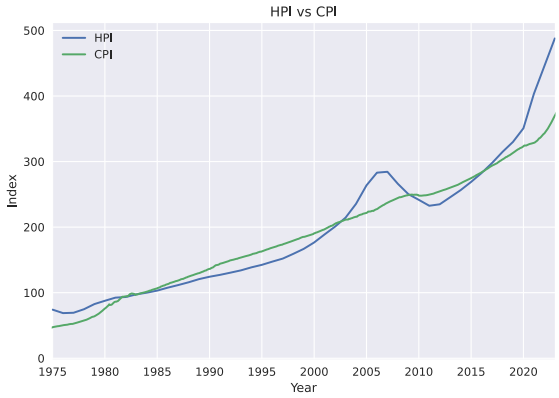

Figure 1: HPI and CPI growth, 1975-2023

HPIs and CPIs between 1975 and 2023 follow a clear, upward trend (Figure 1). The two indexes generally increase together, although one occasionally overtakes the other. The rate at which they increase varies greatly; at several points in this period, a rapid change in one index may be accompanied by only a modest change in the other. There are several notable areas of the graph where index behavior stands out:

1) In the 1980's, both indexes grew at a slow, steady rate.
2) In the early 2000's, HPI began to rapidly overtake the still steady CPI.
3) HPI underwent a significant decrease between 2005-2010, while CPI remained largely unchanged.
4) Since 2016, HPI has been increasing far faster than CPI.

These regions correspond to major economic events, and offer insight into what the current index position means. When indexes were growing steadily in the 1980's, the US was recovering from a recession, and general economic growth favored real estate as both an investment opportunity, while still being an affordable option for many Americans. In the 2000's, a bubble began to form in the housing market, and real estate investment surged far beyond the abilities of consumers – resulting in a HPI crash when the bubble finally burst. Finally, since 2016, HPI has returned to rapidly outpacing CPI, far beyond what was seen leading up to the 2008 Mortgage Crisis. These events indicate that when housing

investment and housing affordability are equal, the market is healthy – but when these diverge, it is unfavorable for consumers. Based on this graph, there is evidence to indicate that the US market is currently seeing abnormal prices compared to the past, since the difference in HPI and CPI is great.

B. Linear Regression

Figures 2 and 3 depict the linear regression model over historical data for HPI and CPI, respectively. Using the linear regression models derived from historical data, it is projected that the HPI and CPI of 2030 will be 405 and 357, respectively. The slope of the HPI regression is 2 index points per year greater than that of the CPI model – indicating that the investment value of the housing market will continue to be better off than consumer affordability. Unfortunately, this means that prospective homeowners may continue to face harsh prices for the next several years, unless another anomaly occurs in the market.

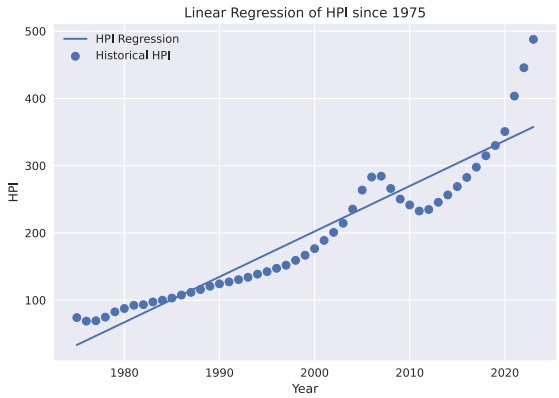

Figure 2 – Linear Regression of HPI

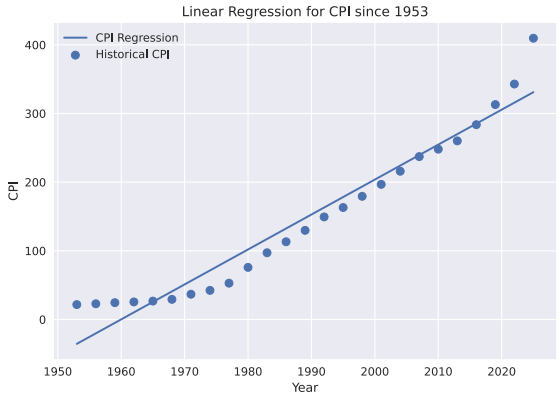

Figure 3 – Linear Regression of CPI

C. Map

The cost of housing is more expensive in urban areas compared to rural areas (Figure 4). Major cities, such as Boston and New Bedford, have among the highest unit costs for living space in Massachusetts, as more people are concentrated in these places. Further west, where there are fewer population centers, the cost declines. For consumers, this means that living far away from bustling cities is usually going to be cheaper. Some cities, however, offer lower prices – Worcester, the second largest city in Massachuesetts, has a cost per square foot of under $750, which is similar to prices seen in remote areas. Prospective homeowners should also avoid areas where tourism is popular: Nantucket and several ZIP codes on Cape Cod have a relatively high cost per square foot of real estate compared to the rest of the state.

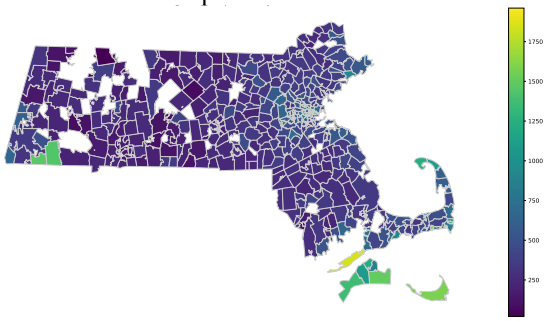

Figure 4 – Housing cost per square foot by ZIP Code, Massachusetts

V. Discussion

Linear regression may not be the best option for predicting the future of the housing market, since the data used shows key signs of non-linear or dependent behavior. This is revealed by the residual plots (Figures 5, 6), which have distinct patterns. An accurate linear regression model would have a random distribution for residuals. In this case, however, clear shapes exist. This indicates that a first-degree line of best fit may not be the best method for predicting future movements.

The validity of the linear regression model may also be diminished by the existence of external factors – patterns exist in past indexes where major economic movements have happened. These may not always be centered in housing investment, like in 2008 – for example, the pandemic induced rampant price hikes and inflation. This led to a rise in CPI for all commodities, not just housing. HPI and CPI alone cannot predict anomalous events occurring that may disturb the market.

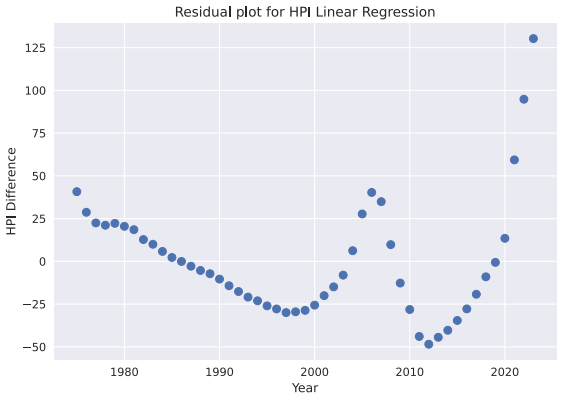

Figure 5

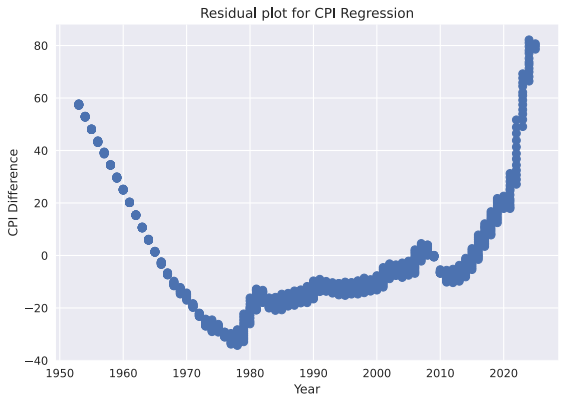

Figure 6

VI. Conclusion

Americans seeking to become first-time homeowners and those looking to rent are facing a difficult market. Real estate prices are at abnormal highs, and trends show that they will continue to remain high. Over any period, the cost of housing is expected to increase steadily, even when the market is relatively quiet. Outside events, such as a pandemic or recession, have a massive impact on how affordable shelter is for consumers. Currently, market indicators suggest that not only are costs incredibly high, but the market as a whole may be unstable. A large disparity between the affordability and investment value of real estate has been seen in previous crashes, and the disparity has never been as wide as it is now. At present, consumers can look to communities outside of major metropolitan areas and popular vacation spots for relatively cheap housing.

References

[1] N. Pisano, “Consumer sentiment in 2025: Strong concerns over inflation, tariffs, government cuts,” Clever Real Estate, https://listwithclever.com/research/consumer-sentiment-2025/?utm_source=press%2Brelease&utm_medium=pr&utm_campaign=consumer_sentiment_2025 (accessed Mar. 30, 2025).

[2] A. Dehan, “The housing market and inflation,” Bankrate, https://www.bankrate.com/real-estate/inflation-housing-market/ (accessed Apr. 12, 2025).

[3] “Federal Housing Finance Agency - FHFA house price indexes (hpis),” Catalog, https://catalog.data.gov/dataset/fhfa-house-price-indexes-hpis-948c6 (accessed Apr. 13, 2025).

[4] “Consumer price index for all urban consumers: Shelter in U.S. city average,” FRED, https://fred.stlouisfed.org/series/CUSR0000SAH1 (accessed Apr. 13, 2025).

[5] A. S. Sakib, “USA Real Estate Dataset,” Kaggle, https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset (accessed Apr. 13, 2025).