TRACKING AND FORECASTING HOUSING AFFORDABILITY
DS 6600

Janice Guo
School of Engineering and Applied Science
Department of Computer Science
University of Virginia
Charlottesville, USA
vdq8tp@virginia.edu

1 INTRODUCTION

The housing market is one of the most important sectors of the economy, influencing both personal financial decisions and broader economic outcomes. Housing affordability, in particular, has emerged as a critical concern in recent years as many households struggle with the combined pressures of rising mortgage rates, escalating home prices, stagnant income growth, and shifting housing demands. When housing becomes unaffordable, it not only affects individual households' ability to build wealth and achieve stability but also has broader implications for economic conditions.

Understanding affordability requires looking at multiple, interacting factors such as income, interest rates, housing supply, and demographic trends. By building a data pipeline that integrates housing market data with mortgage rates and economic indicators, I aim to create an interactive dashboard that helps users explore affordability trends across U.S. regions. The final project will not only combine several data sources but also support descriptive analyses and simple predictive models, making it possible to visualize both historical trends and potential future scenarios. In doing so, I hope to contribute a flexible tool that could be valuable for researchers studying affordability, policymakers evaluating housing interventions, or the general public seeking to understand how housing market conditions have shifted over time and where affordability is most pressing.

2 BACKGROUND

Housing affordability has become one of the defining policy challenges in the United States, shaped by the interaction of rising home prices, mortgage rates, income inequality, and local supply constraints. The traditional metric of affordability, housing costs exceeding 30% of household income, has been widely used for decades, but researchers have increasingly criticized its limitations. For instance, Airgood-Obrycki et al. (2023) argue that this threshold often masks hardship, as many households considered "affordable" under the 30% rule still lack sufficient residual income to cover essential non-housing needs. Their study demonstrates that adopting a residual income framework significantly changes our understanding of the affordability crisis, showing that a far larger share of renters face real economic strain than official statistics suggest.

An additional perspective is provided by Iqbal et al. (2023), who analyze the determinants of affordability across U.S. states. Their findings show that factors such as household size, property taxes, vacancy rates, and dwelling characteristics meaningfully shape affordability outcomes. Importantly, they highlight that affordability challenges differ by region and demographic group, and that economic shocks, such as changes in mortgage rates or property values, disproportionately impact households already under financial strain. This suggests that affordability is a multidimensional issue requiring both local and national analysis.

Another study, by Petach (2022), extends this understanding by linking long-run affordability trends to income stagnation and inequality. Using Census microdata and counterfactual simulations, Petach finds that for households in the bottom quintile of the income distribution, stagnating incomes since 1980 explain nearly the entire decline in affordability, whereas for higher-income households, hous-

ing market frictions (such as zoning restrictions driving up costs above construction prices) play a larger role. This finding underscores that affordability crises have different causes for different groups of people.

Together, these strands of research suggest two critical insights. First, affordability must be understood using better measures than simple ratios and incorporating multiple features, capturing residual income and regional cost differences. Second, affordability drivers vary across households and geographies, with low-income families more affected by wage stagnation and inequality, and higher-income families more sensitive to housing supply and regulation. Building on these insights, my proposed project aims to integrate both descriptive analytics and machine learning to explore affordability dynamics across U.S. regions, highlighting where affordability pressures are most serious and what economic factors are most predictive.

3 DATA

Several sources for raw data have been found:

- Zillow (csv): https://www.zillow.com/research/data/
- Redfin (csv): https://www.redfin.com/news/data-center/
- FRED (API): https://fred.stlouisfed.org/
- US Census (API): https://www.census.gov/data/developers/data-sets.html
- ACS (CSV): https://www.census.gov/programs-surveys/acs/data.html

All listed sources offer data to use for free in a personal or non-commercial setting. However, all generated content from these sources must have proper references and citations, stating explicitly where the data came from.

4 POTENTIAL ANALYSIS

The dashboard will emphasize affordability, understood as the relationship between housing costs, financing conditions, and household incomes. Variables of interest include the following:

- Zillow: home value index, observed rent index, for-sale inventory, sale-to-list ratio, market heat index, new homeowner/renter income needed
- Redfin: median sale price, home sales, price per square foot, weekly/monthly market data, home price index
- FRED: income, average housing salse price, housing affordability index, mortgage index
- US Census: rental housing financial survey, housing vacancies and homeownership, new residential construction
- ACS: median household income, household size, demographic composition

Some potential analyses include the following:

1. Exploratory Data Analysis
    - Time series of median home price, rental indexes, median income, mortgage rate per metropolitan area, and affordability metrics
    - Regional comparisons (e.g. East vs. Midwest vs. West) of affordability ratios, possibly including map visualizations of cost burdens
    - Historical visualization of how interest rates affect monthly affordability
    - Correlation matrices and cross tabulations of features (mortgage rates, incomes, vacancy, prices)
    - Scatter plots of mortgage rate vs. price growth or vacancy rates vs. affordability, coloring by income brackets or region

2. Machine Learning
    - Predict the next period's median home price or forecast affordability ratios through regression models using features such as mortgage rates, prior prices, income growth, supply indicators, and demographics, and visualize predicted vs. actual prices in a scatter plot
    - Create a binary classifier to categorize affordability and examine feature importances to reveal strong predictors of burden
    - Use unsupervised learning to find clusters of different types or brackets of housing markets and display on a map

5 CHALLENGES

Several challenges may arise in building this project, with the main challenges being data integration, cleanliness, and comprehensiveness.

Combining data from different sources may cause issues due to differences in geographic granularity, such as ZIP code versus city versus state, and differences in time frequencies such as weekly versus monthly versus yearly. Wrangling these different datasets and plethora of features into one consistent schema will require careful selection, cleaning, and preprocessing to ensure that only variables with meaningful values are left. Some geographical areas or time periods may also have gaps in coverage, so I will have to filter through all datasets to find a subset of areas or years if I were to conduct historical or time series analyses. Additionally, given that some variables are part of multiple datasets, there may be conflicting values from different sources, which I will have to resolve when merging data into one final dataset. Another major challenge lies with understanding the documentation for all chosen data sources, as sometime variables are not properly and APIs may not provide sufficient descriptions on how to call the resource to obtain the needed variables.

An alternative plan in the event that this housing idea is not feasible is working with stock market information to analyze trends such as interest rates vs bond returns, inflation in short-term stock returns, sector comparisons and correlations, etc. A multitude of stock market information can be found from Yahoo! Finance, Alpha Vantage, and FRED for other macroeconomic information.

6 GENERATIVE AI USAGE

I used GenAI to narrow down my final project topic and pick and alternative idea based on a list of sources. None of the written information present in this report was generated by GenAI.

REFERENCES

Whitney Airgood-Obrycki, Alexander Hermann, and Sophia Wedeen. "the rent eats first": Rental housing unaffordability in the united states. *Housing Policy Debate*, 33(6):1272–1292, 2023. doi: 10.1080/10511482.2021.2020866. URL https://doi.org/10.1080/10511482.2021.2020866.

Javed Iqbal, Jeff Brdedthauer, and Christopher S. Decker. Determinants of housing affordability in the usa. *International Journal of Housing Markets and Analysis*, 18(1):158–177, 08 2023. ISSN 1753-8270. doi: 10.1108/IJHMA-05-2023-0071. URL https://doi.org/10.1108/IJHMA-05-2023-0071.

Luke Petach. Income stagnation and housing affordability in the united states. *Review of Social Economy*, 80(3):359–386, 2022. doi: 10.1080/00346764.2020.1762914. URL https://doi.org/10.1080/00346764.2020.1762914.