Data in Brief
Open access
Article template

ARTICLE INFORMATION

Article title

A Curated Dataset for Analyzing Housing Affordability Trends in U.S. Metropolitan Areas

Authors

Janice Guo*

Affiliations

University of Virginia, 1919 Ivy Rd, Charlottesville, VA 22903

Corresponding author's email address and Twitter handle

vdq8tp@virginia.edu

Keywords

Affordability index; Regional economics; Real estate markets; Time-series data; Socioeconomic indicators; Mortgage rates; Income and financial statistics

Abstract

This project presents a curated, longitudinal dataset designed to support analysis of housing affordability trends across U.S. metropolitan areas. The dataset integrates multiple publicly available data sources to provide a harmonized view of household income, housing market values, and mortgage interest rates over time. Specifically, median household income data are drawn from the American Community Survey (ACS), housing value metrics are sourced from Zillow's Home Value Index (HVI), and mortgage rate information is obtained from the Federal Reserve Economic Data (FRED) database. Together, these sources offer complementary perspectives on the economic and financial factors influencing housing affordability.

The data collection process involved downloading raw data files from each source, followed by systematic cleaning and transformation to ensure consistency across geographic units and time periods. Metropolitan statistical areas (MSAs) were standardized using a canonical naming scheme to address changes in geographic definitions, naming conventions, and coverage across years. Monthly housing value and mortgage rate data were aggregated to annual values to align with ACS reporting frequencies. The final dataset spans the years 2010 through 2024, with the exception of 2020, and includes variables for metropolitan area identifiers, state, income levels, housing values, mortgage rates, and derived affordability measures. Both annual and monthly versions of the data are retained to support analyses at different temporal resolutions.

The resulting dataset is structured in a tidy, long-format representation to facilitate reuse in statistical analysis, visualization, and machine learning workflows. It is suitable for integration into dashboards, exploratory data analysis, regression modeling, and clustering applications. By combining widely used public data sources into a single, consistent framework, this dataset lowers the barrier to entry for researchers, students, policymakers, and practitioners interested in housing market dynamics.

The transparent data processing pipeline and reliance on reproducible public sources further enhance its potential for extension, replication, and adaptation to related research questions.

SPECIFICATIONS TABLE

Subject	Social Sciences
Specific subject area	Housing Affordability in the United States
Type of data	Table (.csv format)
Data collection	Data were collected from publicly available sources: ACS for income statistics, Zillow for housing value indices, and FRED for mortgage rates. Raw files were cleaned and standardized by resolving inconsistent metropolitan area names, aligning geographic definitions, and removing incomplete records. Monthly housing and mortgage data were aggregated to annual values to match survey reporting frequencies, and all variables were normalized into a unified, long-format structure.
Data source location	Data was collected across different metropolitan areas in the United States. The three primary sources of data can be found here: • ACS (for each year): api.census.gov/data/{year}/acs/acs1?get=NAME,B19013_001E&for=metropolitan statistical area/micropolitan statistical area:*&key={key} • Zillow (ZHVI All Homes Seasonally Adjusted Metro & US): https://www.zillow.com/research/data/ • FRED: https://fred.stlouisfed.org/series/MORTGAGE30US The secondary compilation of this data can be found on GitHub.
Data accessibility	Repository name: DS6600-project Data identification number: N/A Direct URL to data: https://github.com/janiceeguo/DS6600-project/tree/main/data
Related research article	N/A

Data in Brief
Open access
Article template

VALUE OF THE DATA

- The dataset combines income statistics from the American Community Survey, housing value indices from Zillow, and mortgage interest rates from the Federal Reserve Economic Data database into a single, harmonized structure. This integration reduces the need for researchers to independently collect, clean, and align data from disparate sources with differing temporal resolutions and geographic definitions, lowering technical barriers to entry and improving efficiency for housing market analysis.
- The dataset spans 2010–2024 (minus the year 2020) and includes both annual and monthly representations of key housing and financial variables. Annual aggregation aligns housing values and mortgage rates with survey-based income measures, while retained monthly data allow for higher-frequency analysis. This dual structure supports a wide range of analytical approaches without requiring additional preprocessing.
- Data are provided in a tidy, long-format structure that is compatible with common statistical, econometric, visualization, and machine learning pipelines. Researchers can readily apply regression models, clustering algorithms, time-series analysis, or interactive dashboards without restructuring the dataset, facilitating reuse across methodological contexts.
- The dataset is suitable for reuse in academic research, policy analysis, and instructional settings. Its reliance on publicly accessible sources and transparent preprocessing steps makes it appropriate for replication studies, classroom assignments, and extensions that incorporate additional socioeconomic or geographic variables.

BACKGROUND

Housing costs and household income are central components of economic well-being and are closely monitored by policymakers, researchers, and market participants. In recent years, changes in mortgage interest rates, housing prices, and income growth have increased interest in understanding how these factors evolve across metropolitan areas and over time. Publicly available data sources provide extensive information on these dimensions, but they are often distributed across separate platforms, reported at different temporal frequencies, and subject to inconsistent geographic definitions.

The motivation for compiling this dataset stems from the need to bring together these complementary sources into a single, standardized framework. Income data from the ACS are reported annually, while housing values and mortgage rates are typically available at monthly or weekly intervals. Additionally, metropolitan area names and boundaries may change across reporting periods, complicating longitudinal analysis. Addressing these challenges requires systematic data collection, normalization, and aggregation procedures. By assembling income, housing value, and mortgage rate data into a harmonized structure aligned at the metropolitan level, this dataset provides a consolidated resource that reflects key dimensions of housing affordability dynamics over time.

DATA DESCRIPTION

Original ACS data files pulled from the API are stored in the *acs_income* folder, which are later combined into the *acs_income_all_years.csv* file. The other two files in the outermost directory are original files downloaded from Zillow and FRED. In the *clean* folder, these three files are cleaned into separate files named *acs.csv*, *zillow.csv*, and *fred.csv*. The clean Zillow and FRED data are further aggregated at an annual level with the *annual* keyword in the file name in the same folder. Next, files are combined at an annual level into *acs_zillow_fred.csv* and at a monthly level into *zillow_fred.csv* in the same folder. Finally, these two combined files are transformed into a long-format table and saved as new files in the *final* folder. Figure 1 below shows the relationship between the two final long tables.

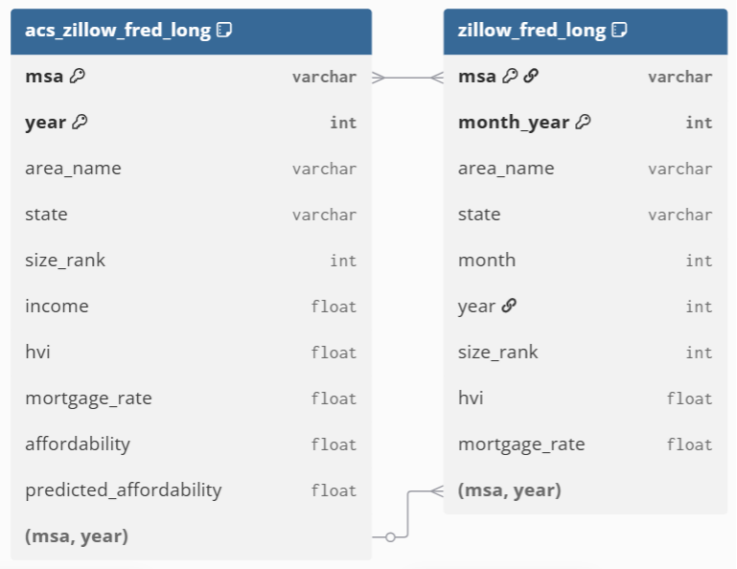

Figure 1: ER Diagram

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

This project followed a reproducible data pipeline design to acquire, clean, normalize, and integrate housing affordability–related data from multiple public sources into a unified dataset at the U.S. metropolitan area level. All steps were executed using open-source software and publicly accessible data repositories. Data processing scripts were written in python and executed locally within a Docker container to ensure reproducibility and consistent dependency management. Three primary data sources were used:

1. American Community Survey (ACS): Median household income data were obtained from the U.S. Census Bureau's American Community Survey 1-year estimates (Table B19013). Data were accessed programmatically using the Census Bureau API for all available years from 2010 through 2024 at the metropolitan and micropolitan statistical area level. API queries were issued using Python scripts, specifying year, summarization level, and table identifiers. Raw responses were returned in JSON format and converted to tabular structures.

2. Zillow Home Value Index (ZHVI): Housing value data were downloaded as CSV files from Zillow's public research data portal. The dataset provides monthly home value indices for metropolitan regions. Columns included region identifiers, region names, geographic classifications, and monthly observations spanning multiple decades.
3. Federal Reserve Economic Data (FRED): Weekly 30-year fixed mortgage rate data were downloaded as CSV files from the Federal Reserve Bank of St. Louis FRED database. Each observation consisted of a date and the corresponding interest rate.

ACS income data were cleaned by converting reported values to numeric types and identifying incomplete records. Metropolitan area names were standardized by removing suffixes (e.g., "Metro Area," "Micro Area") and extracting canonical area and state identifiers. Alternate naming conventions were preserved in auxiliary fields to support longitudinal merging. For example, an original value of "Allentown-Bethlehem-Easton, PA-NJ Metro Area" would be separated out into the main identifier column "Allentown, PA" along with an area name of "Allentown", state of "PA", alternate name of "Bethlehem-Easton", and alternate state of "NJ". These alternate names and states were used to help merge rows that had different naming conventions between different years, so that the cleaned ACS data has one row per MSA with multiple income columns for different years.

Zillow and FRED data were filtered to the 2010–2024 period to align with ACS coverage. Due to COVID-19 having an adverse impact on data collection and causing significant non-response bias, the US Census Bureau did not release a 2020 ACS, so this year was similarly filtered out from the Zillow and FRED data. As Zillow data is provided on a monthly basis, the weekly mortgage rate observations were aggregated to a monthly average. Monthly Zillow and mortgage data were then aggregated to annual values using calendar-year means.

After each of the three data sources were cleaned, they were merged into an annual dataset (combining ACS, Zillow, and FRED data) and a monthly dataset (combining Zillow and FRED data). When merging ACS and annual Zillow data, only MSAs that were common to both datasets were kept, and any MSAs that had missing income or housing data for any of the years were also filtered out. This resulted in 454 final MSAs, which was deemed as more than enough data to conduct analysis on. Afterwards, annual FRED data was cross merged, since the mortgage rate was originally calculated as a national average, and is thus the same for each MSA for a given year. Similarly, monthly Zillow data was also cross merged with monthly FRED data.

After selecting only the relevant columns of interest show in Figure 1, the annual and monthly merged datasets were reshaped into a long-format structure, meaning that instead of having separate income, housing, or mortgage data columns for each time period, there is one column that defines the year or the month/year pairing, and one column for the income, housing index, or mortgage rate. This long format makes it easier to conduct analysis, produce consistent joins across sources, and reuse the data in future experiments.

Since a main goal of this project is about housing affordability, an affordability ratio was also calculated for the annual merged dataset, which was simply just the quotient of the housing value index and the income. This metric measures how many dollars of home value correspond to one dollar of income in a given MSA. Some machine learning techniques were also applied to calculated a predicted affordability value based on income and with linear regression, and k-means clustering

was also used to categorize housing across all MSAs. The affordability ratio, predicted affordability, and cluster group were all added to the final annual merged dataset.

The two annual and monthly merged datasets were loaded into PostgreSQL tables using SQLAlchemy. All scripts used for data acquisition, cleaning, aggregation, and database ingestion were stored as standalone Python files. Configuration parameters (API keys, database credentials) were managed using environment variables. This workflow allows the dataset to be regenerated or extended by rerunning the pipeline with updated source data.

Finally, a dashboard was created to help visualize the data. For a singular MSA chosen from a dropdown, the dashboard shows a tab for annual trends, monthly trends, and machine learning insights. Annual trends include an income over time line chart, a housing value over time line chart, an affordability index over time line chart, and a housing value vs income scatter plot colored by year. Monthly trends include a housing value over time line chart and the housing value overlayed with the mortgage rate over time line chart. Machine learning insights include an actual vs predicted affordability scatter plot with the linear regression line equation displayed above the plot and a housing value vs income scatter plot colored by MSA cluster. This dashboard can be accessed at https://jguo.pythonanywhere.com/dashboard/

LIMITATIONS

The dataset is subject to several limitations related to data availability, coverage, and standardization across sources. Income estimates from the American Community Survey are based on multi-year survey samples and may reflect sampling error, particularly for smaller metropolitan or micropolitan areas. Additionally, ACS income data are reported annually, while housing values and mortgage rates are collected at higher temporal frequencies, requiring aggregation that may obscure short-term variation.

Geographic definitions of metropolitan areas are not fully consistent across data providers or over time. Changes in metropolitan area names, boundaries, and classifications required manual and rule-based reconciliation, which may introduce residual inconsistencies despite normalization efforts. Some metropolitan areas appear in certain years but not others, resulting in discontinuous time coverage that required removal, resulting in a smaller amount of feasible data.

Housing value data from Zillow are based on proprietary modeling techniques, and coverage may vary by region and time period. Mortgage rate data represent national averages rather than region-specific rates, limiting geographic specificity. Finally, the dataset excludes years prior to 2010 due to inconsistent availability of harmonized metropolitan-level income data, along with the year 2020, restricting the historical scope of longitudinal analyses.

ETHICS STATEMENT

The author has read and followed the ethical requirements for publication and confirms that the current work does not involve human subjects, animal experiments, or any data collected from social

media platforms. Any humans involved in the questionnaire process were volunteers for the US Census Bureau.

CRediT AUTHOR STATEMENT

Janice Guo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing, Visualization, Supervision

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTERESTS

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] W. Airgood-Obrycki, A. Hermann, S. Wedeen. “The rent eats first”: rental housing unaffordability in the United States. Housing Policy Debate, 33(2023) 1272–1292. https://doi.org/10.1080/10511482.2021.2020866.

[2] J. Iqbal, J. Brdedthauer, C.S. Decker. Determinants of housing affordability in the USA. International Journal of Housing Markets and Analysis, 18(2023) 158–177. https://doi.org/10.1108/IJHMA-05-2023-0071.

[3] L. Petach. Income stagnation and housing affordability in the United States. Review of Social Economy, 80(2022) 359–386. https://doi.org/10.1080/00346764.2020.1762914.

[4] U.S. Census Bureau. American community survey 1-year estimates [dataset]. https://data.census.gov/table/ACSDT1Y2024.B19013?q=median+household+income

[5] Zillow Group, Inc. ZHVI all homes (SFR, condo/co-op) time series, smoothed, seasonally adjusted ($), metro & US [dataset]. https://www.zillow.com/research/data/

[6] Federal Reserve Bank of St. Louis. 30-year fixed rate mortgage average in the United States [dataset]. https://fred.stlouisfed.org/series/MORTGAGE30US

[7] J. Guo. DS6600 final data [dataset]. GitHub. https://github.com/janiceeguo/DS6600-project/tree/main/data