Breaking Ground

An analysis of housing development, affordability, and the underlying factors influencing both

Authors: Nathan Cantwell, Sergio Lopez, and Noah Tamminga

Date: Jun 12, 2025

Motivation

There is an apparent housing affordability crisis in the USA, and the "American Dream" of purchasing a home seems less achievable each year. Public perception is that increasing housing prices outpace rising wages, and even with new construction increasing supply, home prices will not decrease. Our team would like to investigate trends in permitting for new housing developments, real estate market data, and how these data differ by US county to gain a sense of the scope of the housing crisis, with a focus on how current development plans may be indicative of a given county's future state in terms of housing supply and affordability.

While investigating prior research our team found two related analyses; the first "*The local residential land use regulatory environment across U.S. housing markets: Evidence from a new Wharton index*" also investigated housing affordability, but focused on blockers to development including zoning, state legislature and court involvement, and density restrictions (Gyourko, Hartley, and Krimmel, 1). The second, "*What Drives House Price Cycles? International Experience and Policy Issues,*" was unable to find consistent housing stock data at the subnational level (Duca, Muellbauer, and Murphy, 38), a possible opportunity for this project.

Based on the economics of supply and demand, we propose that an understanding of historical development trends through permits, in conjunction with related factors influencing supply and demand, including inflation, wages, and the overall cost of living, can be used to predict the future state of housing supply and demand. Developing an understanding of historical trends and relationships will enable us to assess the effectiveness of current development plans. Additionally, it can be used for future predictive analyses in various ways, such as estimating future pricing changes, assessing cost burdens, or predicting the likelihood that planned development meets a given county's needs.

In our report, we will focus on three primary questions of interest:

1. Is home ownership and affordability a systemic issue nationwide in the US, or is it an isolated problem driven by supply and demand?
2. Has housing affordability progressively become worse over the past decade?
3. Are plans for development in the form of permits a leading indicator of increased supply and affordability or are they a demand-driven response to the housing market?

By compiling data from various sources and analyzing the information, we try to uncover the answers to these questions and lay the groundwork for future analytics that may seek to propose new solutions or better explain existing relationships within this data.

Data Sources

Redfin Housing Market Data & Consumer Price Index

The Redfin Housing Market dataset compiles time series real estate sale data from January 2012 to March 2025. For this period, the dataset includes approximately 1.2 million records with 58 features for each observation. Among the many available features, our key features of interest in this dataset include the county listed, sales price, list price, and inventory - which will all be used in our analyses of affordability and development. Access to this dataset can be obtained via the Redfin website, and the tsv file itself can be found here: Redfin tsv file.

This information can be used to understand the history of price and inventory trends across the United States over the past decade. To facilitate useful analysis with other datasets, the Redfin dataset was combined with the Consumer Price Index from the Federal Reserve and a helpful County Federal Information Processing Standard (FIPS) table for consistent references across all datasets. These supplementary datasets will be discussed later in this report in the Data Cleaning section.

Quick Access

- Location(s): Redfin, tsv file
- Format: tsv file
- Variables: County, FIPS, Sales Price, List Price, Inventory
- Records: Approximately 1.2 million records
- Time Period: January 2012 - March 2025 by month

Income

To gather data on income, we used the US Census API on the American Community Survey (ACS) to retrieve median household income data for 2012 to 2023 by county. For these 11 years, our data was approximately 48000 records and six features. The full dataset contained approximately 17.5 million records and 20,000 features - we believe this encompasses the entire ACS dataset. This dataset's primary variable of interest is the median household income, along with the necessary key to join other datasets found in the year and FIPS code. With median household income, we create an affordability index for the cost of housing. Information on the US Census API, including methods of access and available APIs, can be found on the US Census API website. Additionally, our Jupyter file for income provides an example of using the API to extract income data from the American Community Survey API, which includes all the years that are extracted, including data for the year 2023 as an example.

Quick Access

- Location(s): US Census API, ACS 2023 Example Link
- Format: API
- Variables: Year, County, FIPS, Median Income
- Records: Approximately 48000 records
- Time Period: 2012 - 2023 by year

US Census Building Permits Survey

The US Census data from the monthly building permits survey provides imputed and reported permit totals for new privately owned housing by county and month for the period of 2000 to 2025. The data is split into three measures for buildings, units, and value and is organized by four different housing classifications for the number of units it contains (single, double, three-four, five plus). The survey breaks figures into imputed and reported estimates with time aggregates by month, year-to-date, and annual in different files, but for our purposes, we will analyze the data provided on a monthly basis.

For the 25 years of data gathered, we retrieved roughly 650,000 records for both imputed and reported values by county and month. While there are numerous methods to access this data, we chose to extract this data from the US Census' ASCII file repository and specifically their repository for Counties, which provided comma-separated txt files each with a standard format hyperlink for the year, period, and type of report.

Quick Access

- Location(s): US Census' ASCII Repository, County txt files for permits
- Format: Comma separated txt files
- Variables: Date, County, FIPS, Buildings, Units, Value
- Records: Approximately 650,000 records
- Time Period: January 2000 - April 2025 by month

Data Manipulation

Redfin Housing Market Data & Consumer Price Index

The Redfin dataset was a challenge for this project due to the extensive manipulation needed compared to our other sources. This manipulation included three primary areas of focus: basic cleaning steps, joining supplementary datasets for inflation and county data, and applying methods of filtering noise and imputing missing values.

First, various string methods were used to join and align the retrieved data to the standards needed for the other datasets. Among the county names in the Redfin dataset, we found that various county names were inconsistent with the data retrieved from the census. However, after applying a function to map various counties together that were disparate, we were able to join the FIPS identifiers to the Redfin data via matching state abbreviations and county names. The supplementary FIPS county dataset was joined in this manner, and with some additional backfilling steps to create the five-digit FIPS codes, the Redfin dataset was able to merge with our other primary datasets. As a final step, we were able to join the Consumer Price Index supplementary dataset to this final result via matching months. With the Redfin data paired with the supplementary datasets and a subset of columns selected, we were ready to handle additional processing and imputation steps.

The bulk of the data cleaning for our Redfin data involved removing outlier values with z-scores, cross-joining months to add observations in non-reporting periods, imputing values with linear interpolation, and applying an inflation adjustment. For outlier removal, we implemented a z-score filter. Z-scores were computed for a given county and records vastly out of range for a typical record in the county (greater than an absolute value of three) were removed. This resolved the issue of house prices at extremes such as $1 or $1 trillion and other outlier values. Next, because not all counties had reported real estate prices for each period, a pandas series of dates was generated and cross-joined to the Redfin data. The benefit of the cross-join was that it created new records for each missing month in the dataset. With this new dataset filled with all time periods, the observations missing reported metrics needed to be imputed. To accomplish this task, we used linear interpolation to impute missing values. By applying the linear interpolation on a given county group in the dataset, we were able to fill in pricing data that was missing by interpolating increases or decreases linearly across a given county's actual reported value in the Redfin data. We believe this was reasonable given its basis on the Redfin data and applying increases or decreases linearly across the missing months between the actuals reported by Redfin. We also used forward fill for missing records of home inventory; assuming no sales were made for missing values inventory would be constant. Finally, the method for adjusting pricing was relatively simple. With the CPI data provided by the Federal Reserve, we were able to quickly merge the CPI adjustment factor to the primary Redfin dataset by month and calculate the sale and list prices after accounting for inflation.

For more detailed information including a cell-by-cell walkthrough, please refer to "housing_market_data_cleaning.ipynb" in the src repository.

Income & Affordability

To support a detailed analysis of housing affordability, the U.S. Census Bureau's American Community Survey (ACS) 5-Year Estimates API was used to obtain county-level median household income data. The ACS5 dataset offers reliable, standardized, and geographically detailed income figures, making it well-suited for comparison with housing market data. Using the API ensured the data was current, clean, programmatically accessible, and scalable across multiple counties, which was critical for integrating with the aggregated dataset of median home purchase prices.

The home value-to-income ratio (calculated by dividing the median home price by the median household income) was initially considered a potential metric to derive affordability. While it provides a quick snapshot of affordability, it does not account for essential financial factors such as interest rates or loan terms. Consequently, it can oversimplify the analysis by ignoring the real-world costs of financing a home. Given these limitations, the home value ratio was not selected as the primary affordability measure.

Instead, the Home Affordability Index (HAI) was used to evaluate whether a typical household earns sufficient income to qualify for a mortgage on a median-priced home. A fixed mortgage interest rate of 5% was applied to simulate a stable long-term borrowing environment. This allowed for estimating monthly mortgage payments and comparison to local income levels. This data was compiled by combining Census API data with median housing data at the FIPS and year levels, and then the adjustment factor was applied. Incorporating interest rates into the analysis provided a more realistic and comprehensive view of affordability by accounting for the actual financial burden of homeownership under standard lending conditions.

Home Affordability Index (HAI):

$\text{HAI}=(\text{Median Household Income }/\text{Qualifying Income})\times 100$

Qualifying Income Calculation:

$\text{Qualifying Income}=0.28P\times c$

P = Median home price
r = Annual interest rate (5%)
n = Loan term in months (30 years)
c = Monthly mortgage payment calculated as: $c=1-(1+r/12)-nr/12\times P$

For more detailed information, including a cell-by-cell walkthrough, please refer to “home_affordability_index_flow.ipynb” in the src repository.

US Census Building Permits Survey

The US Census website blocks scraping activity, but a list of text file URLs were constructed following the standard naming convention to extract each file and append it to a final dataframe. After extracting the comma-separated values, another helper function applied steps to clean the file contents, including removing odd formatting and setting column names. Following this, the output was split between imputed permits and reported permits. One benefit of the Census data was that issues of missing data and joinable columns were not a concern. The imputed permits already contained a robust imputation process for reporting areas that failed to submit a report. Overall, their methodology imputes a value based on geographic area, history, and type of housing. This imputation specifically takes into account prior annual information for a given geographic area where reporting occurred, as well as current information, to calculate an imputation factor. Then, that imputation factor is applied to non-reporting areas to impute missing values (US Census, BPS).

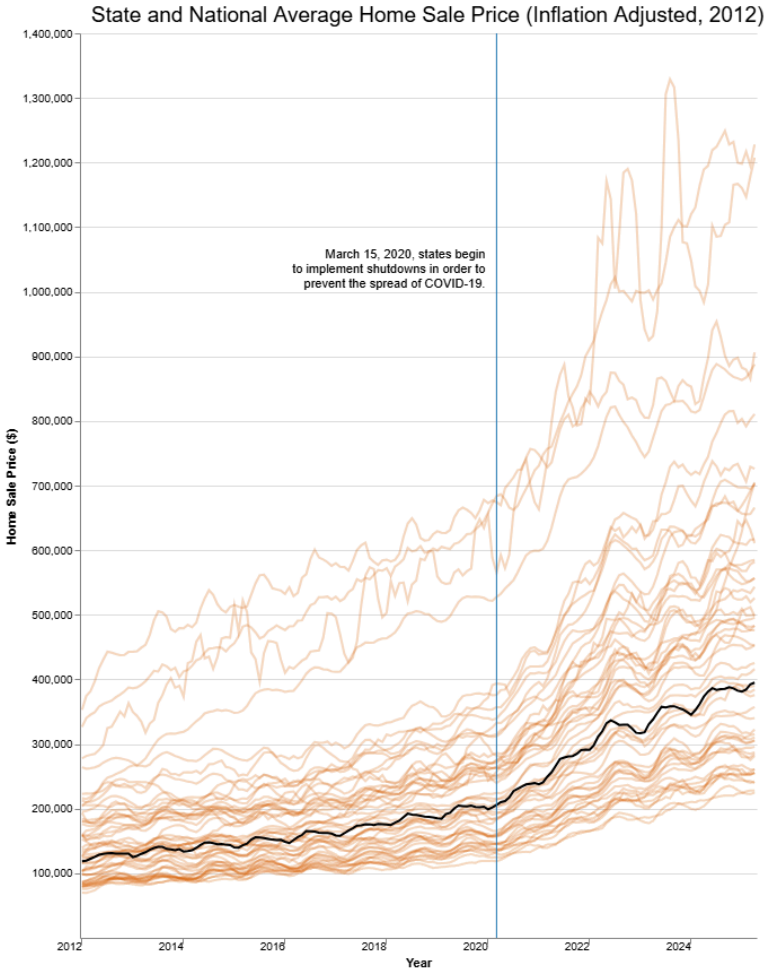
To aid in our visualizations and analyses using permit data, we reorganized our dataframe into a long format from the wide format it came as. Specifically, we wanted three measures: buildings, units, and value. The extra breakout for the number of units is more suited to be a separate categorical column. To accomplish this task, we melted the dataframe, extracted the categorical column, and pivoted the data back to aggregate our three measures of interest. Finally, to join permits and price together, given that both were the two largest tables we worked with, we created an index on the date and county ID codes to merge the datasets. With our two datasets joined and a subset of columns selected, we were ready to progress to the correlation analysis.

For more detailed information including a cell-by-cell walkthrough, please refer to the following notebooks “permit_extract.ipynb”, “permit_cleaning.ipynb”, and “permit_price_cleaning.ipynb” in the src repository.

Analysis & Visualization

Trends

To answer the question of whether housing affordability has progressively become worse over the past decade, we first wanted to understand overall trends in the housing market. With the Redfin dataset, we could observe prices from 2012-2025. During this period, we expected a rise in prices over time, which is verified and communicated through the data visualization below.



National average sale price (in black) rises from $120,000 in 2012, to $200,000 in 2020. By 2025 this has reached nearly $400,000. In descending order three states (MA, HI, CA) and DC appear far higher than the main group, state lines in orange.

As expected, there is an overall rise in home prices, both for the national average (in black) and for each individual state (in orange). This is also generally true when plotting each individual county (over 3,000), but would make this plot too cluttered and harm effectiveness.

Revealed in this plot is how the rise in price accelerated after the beginning of the COVID-19 pandemic. In the eight years up to March 2020, the national average sale price increased by $80,000; in the five years since there has been an increase of $200,000. Over these thirteen years, the average home sale price has more than tripled even when accounting for inflation. Additionally, for state and national trends, there does not appear to be a deceleration in rates of price increases to pre-pandemic levels.

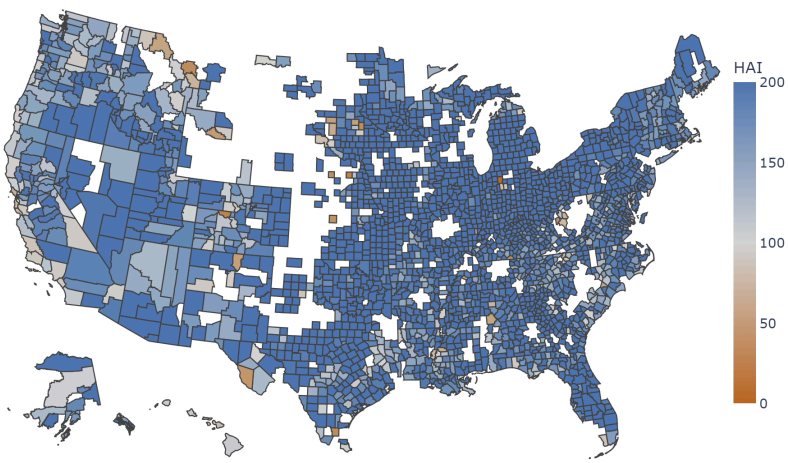
For more detailed information, including a cell-by-cell walkthrough, please refer to “price_trend_plot.ipynb” in the src repository.

Choropleth

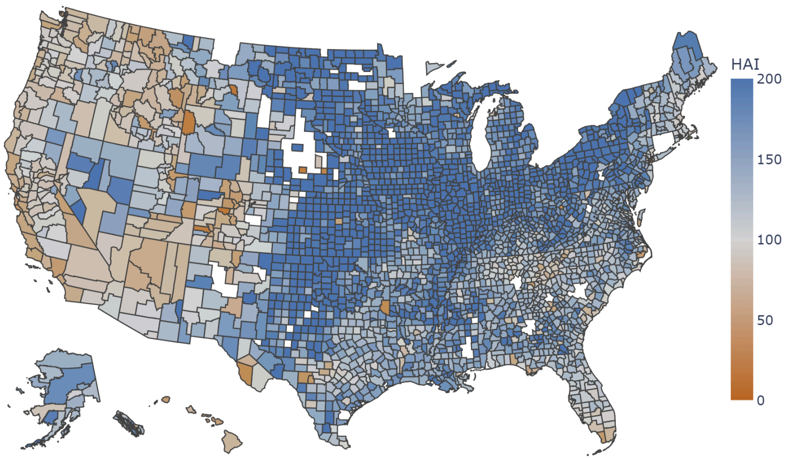
To expand our trending analysis, and to get to the root of the question of housing affordability over time and across geographies, we developed choropleth maps to view differences over time and by geography. Since the data we are analyzing is inherently geography-based, when looking at trends nationwide by county, we create a more effective and granular analysis of housing trends for affordability. Using the Home Affordability Index described earlier, the visualizations below show how affordability has progressed over the last decade. Specifically, although not all counties had good data in 2012, we see a mostly affordable landscape in many geographic regions of the country. In contrast, a decade later, nearly every location has become worse with a significant decrease in affordability found along the West Coast, in the Rockies, and in some Southern states.

For more detailed information including a cell-by-cell walkthrough and visual animations, please refer to “home_affordability_index_flow.ipynb” in the src repository.

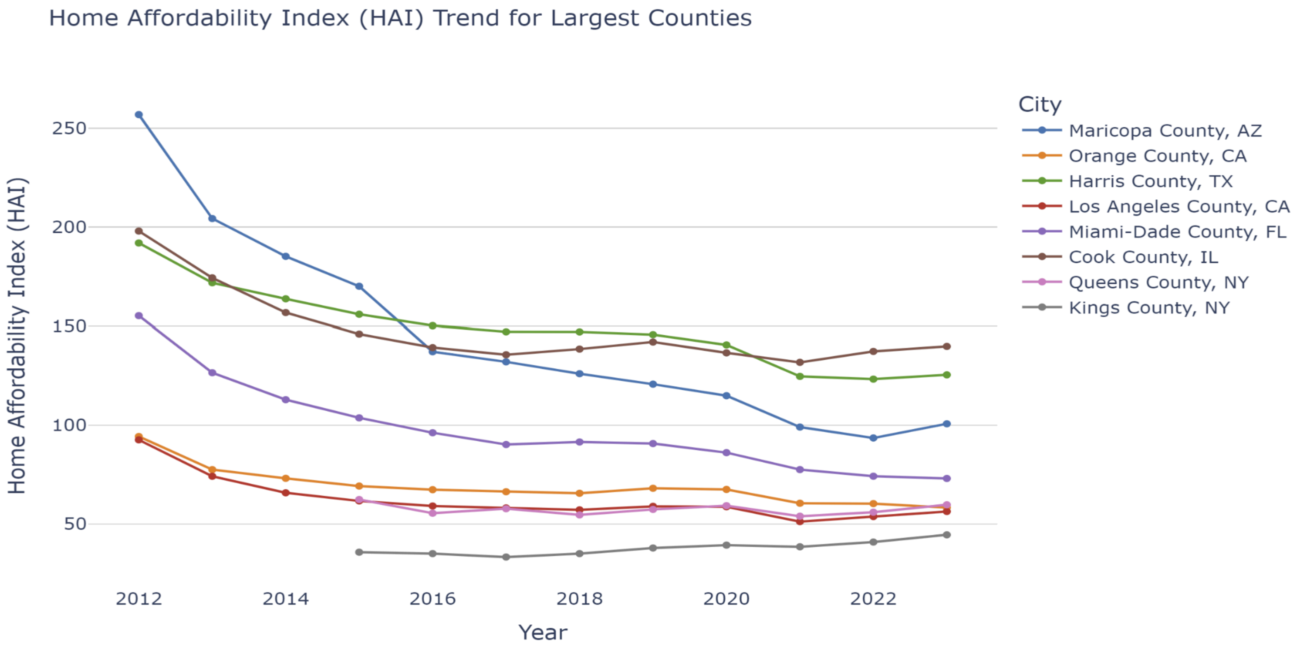
Home Affordability Index (HAI) by U.S. County - 2012



Home Affordability Index (HAI) by U.S. County - 2022



As an additional point of interest, we extracted the trends for the eight most populous counties in the US, which are expressed in the visualization below. Interestingly, whereas some metropolitan areas such as New York have stayed relatively constant in terms of affordability (or even slightly improved), other locations, particularly those in the West and Southwest, have plummeted in affordability over the past decade. In this worsening landscape, finding ways to improve affordability is essential to securing long-term affordable housing.

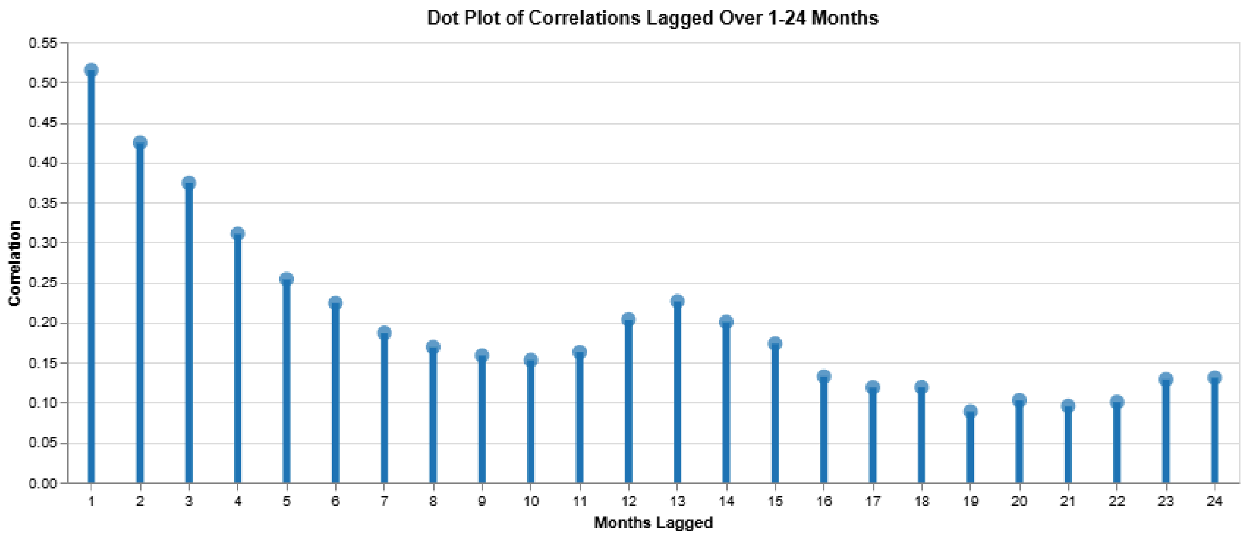


Correlation

In seeking to better understand and explain methods to improve the affordability index, we used various correlation analyses with a focus on lagging 24 months of prices across permits to see any relationship between permits and pricing. We expect that if permits increase supply, we will see a negative relationship between permits and price through negative correlations in future months. However, we might also encounter a situation where there is no discernible relationship or the relationship is the opposite of our expectation due to demand-driven exogenous features that need to be controlled before we can truly understand the relationship between permits and prices.

To start, we will look at the correlation between the normalized results of units and the sale price. Since we are using data from counties with vastly different scales in terms of units and value, we want to make sure everything is evaluated on an equal basis. We want to see if negative correlations exist, and if they do, what their corresponding magnitude is in the negative direction.

In our visual below, we see that among the correlations found via lagging prices across permits by month, there is in fact no negative relationship between permits and prices aggregated by date. Instead, the lagged values only result in positive correlations between permits and prices with roughly a 0.50 to 0.20 correlation for 1 to 6 months out, which is not indicative of a relationship at all in the negative direction.



For more detailed information, including a cell-by-cell walkthrough, please refer to "correlation_analysis.ipynb" in the src repository.

With the relationships being exclusively positive (i.e., an increase in units leading to an increase in price), we encounter the result that many factors are driving up prices that need to be controlled for via causal analysis before we can obtain an understanding of permits and development on price.

As seen in the visual above, the correlations across all price lags are all positive values. A potential explanation for this phenomenon could be prices always rising from demand pressure in high-demand areas (positive correlation) and lowering prices in low-demand areas (negative correlation) rather than capturing any true impact of permits and new developments on pricing.

This seemingly demand-driven environment would indicate the opposite kind of relationship than what we initially expected. Instead of permits and development functioning as a mechanism to improve affordability, it appears that it may be a response to increasing demand along with various other demand-driven factors. In conclusion, although we appear to have uncovered some relationships between pricing and development over time, without conducting extensive research on the causal impact of permits on price, we cannot make a claim about their relationship from what we have discovered in this analysis.

Statement of Work

- All
    - Collaborate in writing the project proposal and defining the project scope
    - Meet on average twice a week to review progress, collaborate on issues, and set project priorities
    - Coauthor project report, reviewing project sections and different visualizations and analysis results
- Noah
    - Author for analysis proposal/sequence in the project proposal
    - Extracted, loaded, and transformed the US Census permit datasets
    - Performed data analysis and cross-correlation analysis on permits and house prices and created dot plot visual
    - Provided feedback/suggestions for other visualizations based on information visualization course content
    - Created outline for final project report and typed up initial project report sections
- Sergio
    - Researched US Census data for income by county
    - Researched API process to extract data from the Census Bureau
    - Researched the Home Affordability Index and required features to perform this analysis
    - Performed Exploratory Data Analysis on this data and joined it with RedFin data to show the Home Affordability Index
    - Investigated possible choropleth plot from multiple modules until one was found to meet our mission requirements
- Nathan
    - Investigate articles for prior research on housing affordability and development
    - Primary author for project motivation, and ethical considerations sections of project proposal
    - Source, clean, and manipulate Redfin housing market dataset
    - Additional datasets for FIPS codes and CPI
    - Create spaghetti plot for home sales prices, by state and national average
    - Primary team coordinator and scheduler
- Drawbacks, what could have gone better?
    - When reviewing our work, we all agreed that using Google Collab helped in organizing our files quickly, but became more difficult to work with as our project grew. With individual workspaces and permissions to each other's datasets, it was more difficult to maintain the correct references to files and have our code run smoothly while simultaneously being shared. We agreed that if we were to start the project again, we would use GitHub instead to create a better-shared code base.

Endnotes

Works Cited

“Consumer Price Index for All Urban Consumers: All Items in U.S. City Average.” *FRED*, 13 May 2025, fred.stlouisfed.org/series/CPIAUCSL.

Duca, J., Muellbauer, J. & Murphy, A. (2021). 'What Drives House Price Cycles? International Experience and Policy Issues.' Journal of Economic Literature, 59 (3): 773-864.

Gyourko, Joseph, et al. “The Local Residential Land Use Regulatory Environment across U.S. Housing Markets: Evidence from a New Wharton Index.” *Sciencedirect.Com*, 4 June 2020, https://www.sciencedirect.com/science/article/pii/S009411902100019X. Accessed 10 June 2025.

Konstantinovsky, Thomas. “Cross-Correlation and Coherence in Time Series Analysis: How to Uncover Relationships Between...” *Medium*, The Pythoneers, 3 Oct. 2024, medium.com/pythoneers/cross-correlation-and-coherence-in-time-series-analysis-how-to-uncover-relationships-between-c83a08990b2d.

“Methodology: Housing Affordability Index.” National Association of REALTORS®, 28 Dec. 2011, www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index/methodology?utm_source=chatgpt.com.

Plotly. “Choropleth.” *Choropleth Maps in Python*, plotly.com/python/choropleth-maps/. Accessed 10 June 2025.

Plotly. “Plotly GeoJSON Map.” *GitHub*, raw.githubusercontent.com/plotly/datasets/master/geojson-counties-fips.json. Accessed 10 June 2025.

“Python - Pandas Interpolate within a Groupby - Stack Overflow.” *StackOverflow*, 2016, stackoverflow.com/questions/37057187/pandas-interpolate-within-a-groupby.

Redfin. “Data Center.” *Redfin Real Estate News*, 3 Apr. 2025, www.redfin.com/news/data-center/.

UMSI. “Spaghetti Plots, A Discrete Alternative.” *Coursera.Org*, 2025, www.coursera.org/learn/siads524/lecture/hLsWt/spaghetti-plots-a-discrete-alternative.

US Census. “American Community Survey 5-Year Data (2009-2023).” *Census.Gov*, 10 Jan. 2025, www.census.gov/data/developers/data-sets/acs-5year.html.

US Census. “American National Standards Institute (ANSI), Federal Information Processing Series (FIPS), and Other Standardized Geographic Codes.” *Census.Gov*, 1 May 2023, www.census.gov/library/reference/code-lists/ansi.html#cou.

US Census. “BPS - How the Data Are Collected.” *United States Census Bureau*, 15 Apr. 2019, www.census.gov/construction/bps/methodology.html.