

Case Study of Top 10 Metropolitans of USA: Relation of Macroeconomic Indicators (Inflation, Interest Rates, and Income) & Housing Affordability

MUHAMMAD AJLAL KHAN, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

As a result of globalization and interconnected economies, we are witnessing an enormous outflow of skilled workers from developing countries towards developed economies like Germany and USA. Usually, the ground realities of developed economies are not transparent from the home countries of expats, hence it becomes difficult for the skilled workers to make informed decisions while relocating to developed countries. The author particularly wants to analyze housing affordability over the last 10 years in the top 10 Metropolitans of the US. Housing affordability is the most important thing during relocation for expat-skilled workers because rental prices constitute a disproportionate chunk of their disposable incomes as compared to other expenses like food, etc. Moreover, the condition of the housing market can be seen as a proxy measure to measure the health of an economy if analyzed side by side with Macroeconomic indicators. This study will provide an outlook on the US economy for people interested in moving there.

1 INTRODUCTION

Research Question: Analyzing Relationship of Selected Macroeconomic Indicators & Housing Affordability in the USA.

The general intuition behind this parallel approach to explaining housing affordability with Macroeconomic indicators is as follows.

1. Consumer Price Index (Inflation): Inflation eats up the purchasing power over time so if there has been increased inflation it means people will find it more difficult to meet their expenses on the same income, which in turn triggers a series of events, e.g., the housing prices/rentals will increase, causing pressure on other expenses as well.

2. Real Disposable Personal Disposable Income (RDPI): RDPI gives an overall trend of people's income after taxes. If it is increasing over time it shows a positive trend, i.e., people are not experiencing an ever-increasing tax burden and they can have more disposable incomes year after year hence, they can improve their quality of life. If the percentage increase in home prices is greater than the percentage increase in RDPI over time then it means affording a house is getting harder and harder over time.

3. 30/15-Year Fixed Mortgage Rates (Interest Rates): Interest Rates are not only good indicators to judge the ease of buying houses but also have an indirect effect on housing rental prices. If Interest Rates are lower then it means more people will prefer to buy a house on lease rather than a rental house. A high trend for buying houses might have a positive effect on the customers of the rental market as less demand for rental houses might drag the rental prices down.

2 DATA SOURCES

We will utilize two categories of data in this project, which are given below.

Macroeconomic Data: It is collected from the Federal Reserve Bank of St. Louis. This data contains four sub-categories for which

the definitions have already been provided before and for each subcategory, we have one dataset, these sub-categories include.

- (1) Consumer Price Index (Inflation)
- (2) Real Disposable Personal Income
- (3) 30-Year Fixed Mortgage Rates
- (4) 15-Year Fixed Mortgage Rates

Housing Data: It is collected from Zillow Group, a Real Estate tech giant in the USA. This data also includes four sub-categories as given below with definitions.

1. Zillow Home Value Index (Home Prices): A measure of the typical home value across a given region and housing type.

2. Zillow Observed Rent Index (Home Rentals): A measure of the typical observed housing rent across a given region and housing type.

3. Zillow Observed Renter Demand Index (Rental Demand): A measure of the typical observed rental market engagement across a region and housing type.

4. The Housing Market Heat Index: A measure that aims to capture the balance of for-sale supply and demand in a given market for all types of homes.

For each of the first 3 sub-categories, we have collected two types of data i.e., single-family-homes and all-types-of-homes. We have only one dataset for heat index which makes the total number of housing datasets 7.

2.1 DATA LICENSE

Macroeconomic Data: The Federal Reserve Bank of St. Louis encourages the use of FRED data, and associated materials, to support policymakers, researchers, journalists, teachers, students, businesses, and the general public.[[FRED 2024](#)]

Housing Data: All data accessed and downloaded from Zillow is free for public use by consumers, media, analysts, academics, and policymakers, consistent with our published Terms of Use.[[Zillow 2024](#)]

2.2 DATA QUALITY & FORMAT: RAW DATA

Macroeconomic Data: For our project we have 4 different macroeconomic datasets from FRED, which FRED sources from specialized bureaucratic departments, e.g., CPI from the US Bureau of Labor Statistics, etc. All macroeconomic data is in a long/time-series format, each row representing an individual date. CPI and RDPI data are already with monthly intervals where each row represents one month's data. However, the interest rate data is with weekly intervals.

Housing Data: For our project we have 7 different housing datasets from Zillow Research, a specialized research wing of Zillow Group. All the housing data is in a table/wide format with monthly intervals, where each row represents a metropolitan area and each column represents a month in addition to some other columns.

Required Transformations: Using Pandas/matplotlib EDA methods we found that these datasets do not have any significant missing values or other data anomalies. Although we have high-quality data, there are a lot of technical transformations that our pipeline needs to implement before it is analysis-ready. Highlights are as follows.

In the Macroeconomic data, both interest rate datasets are with weekly intervals so our pipeline would take care of these two datasets and convert them into monthly intervals so that, later on, we can combine all four Macroeconomic datasets into one time series format using the DATE column as reference.

In the Housing Data, all datasets are in the same table/wide format with monthly intervals hence our pipeline needs to convert them into long format and then join all these datasets into one housing dataset using DATE and Region columns as reference. Moreover, Zillow started to maintain different datasets from different years, i.e., Heat Index, and Rental Demand data starting from 2018 and 2020 onward respectively. So, our pipeline will use a Gradient Boosting Regressor (GBR) to get the values of missing years so that we can have complete 10 years of data.

3 DATA PIPELINE

In this project, we have used Python to develop an automated pipeline with classic Extract, Transform, and Load (ETL) architecture (Fig. 1.), details are given below.

3.1 Data Extraction:

We used the requests module to send an HTTP GET request to the provided URL using the requests.get method. If the request is successful, it proceeds with extracting data. If not, it implements counteractions, e.g., retries it. Finally, we used pandas library's `pd.read_csv` method to read the CSV data from the text content of the response.

3.2 Data Transformation: Feature Engineering & Data Cleaning:

We are only interested in keeping the data from 2015 onward because we plan to perform analysis over the last 10 years. So, all the datasets have been filtered so that they contain data from 2015 onwards only. Dataset-specific transformation steps are given below.

1. CPI: In raw data we only had the CPI values for each month so our pipeline calculates the Inflation (year-over-year percentage change in CPI). Then the pipeline renames the `CPIAUCNS` column to CPI and also adds the Inflation column to the dataset.

2. Interest Rates: Then our pipeline processes fixed 30/15-Year mortgage rates data. As we had weekly data for them so our pipeline calculates the monthly average of the interest rate based on weekly data, then creates a new column for 30/15-Year rates as `MORTGAGEUS30_MonthlyAvg/MORTGAGEUS15_MonthlyAvg` and then transforms data to monthly intervals by dropping weekly data.

3. RDPI: No feature engineering was needed as it was already in monthly intervals, our pipeline only renames the original column for Disposable Income, from `DSPIC96` to `RDP_Income`. Note that, at this moment after transformation, CPI, RDPI, and Interest Rate data are saved as separate CSV files in the data directory.

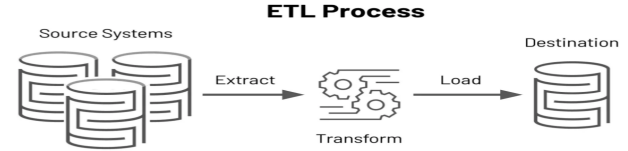


Fig. 1. ETL Pipeline Architecture

4. Housing Data: Then our pipeline uses a common function defined to transform all the Zillow Housing data. As all the 7 housing datasets were in wide/table format and had the same structure except for the target variable e.g., rental prices, rental demand, etc. For all the Housing Datasets our pipeline filters the top 10 largest metropolitan areas. Then to unpivot the table it sets 'RegionName' as the index, and drops unnecessary columns. It then transposes the data into a long format, where each row represents a date & region, and columns correspond to the target indicator, e.g., prices/rentals. The pipeline saves all 7 transformed housing datasets as separate CSV files in the data directory.

5. Dataset Merging: In this step our pipeline merges all the CSV files and saves them in the data directory, details are given below.

A. Merging Macroeconomic Data: At this point all four macroeconomic datasets are standardized as they have monthly intervals they have data from 2015 onward and they are all in time series format. Our pipeline combines all of these four datasets into a single dataset by using the Inner Join method and the DATE column as the reference to join.

B. Merging Housing Data: At this point, we have a total of 7 housing datasets in the time series format for the top 10 metropolitan areas of the USA. Our pipeline uses the first available housing dataset as a starting point and then performs Left Join on it, with DATE and Region columns as a reference, to join all remaining housing datasets.

C. Merging Macroeconomic & Housing Data: Now we have both the macroeconomic and housing data in two separate CSV files in time series format with monthly intervals so our pipeline, using left join, joins Macroeconomic data to the Housing Data, with the DATE column as a reference.

6. Final Transformation: At this point our merged CSV file has a long format with 14 Columns/features where each row represents a region and a month. Now our pipeline will fill the missing values using Gradient Boosting Regressor (GBR), and then finally the pipeline will save the cleaned data as a single CSV file that is ready for typical time-series analysis. We specifically used GBR because it is very lightweight, can automatically learn interactions between features, and is capable of capturing complex nonlinear relationships between features and the target variable. It achieves this by building an ensemble of decision trees, which are inherently nonlinear. This flexibility allows it to model intricate patterns in complex tabular data. We didn't rely completely on pandas' built-in methods to fill in missing data because of their limited scope, i.e., we had a series of data missing where, e.g., `bfill/ffill` methods do not work and using mean or median would have poor statistical significance.

4 DATA QUALITY & FORMAT: TRANSFORMED DATA

We used Histograms and Kernel Density Estimation (KDE) plots to analyze the quality of our transformed data. Histograms (Fig. 2.) suggest that our data is perfectly balanced across Regions, Years, and Months. In the Years and Months graph, the last bars are shorter because the data from the last few months of this year is still missing. The KDE plots for the continuous variables show that all the

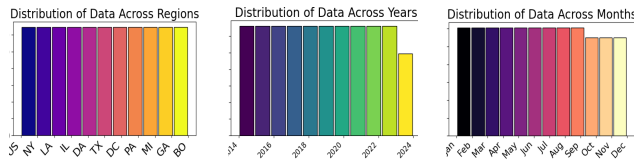


Fig. 2. Categorical Features from the Transformed Dataset

features of Macroeconomic (Fig. 3.) and Housing (Fig. 4.) data are normally distributed with well behaved bell-shaped curves, suggesting a real life practical data. There is a weak positive skew in most



Fig. 3. Continuous Macroeconomic Features from the Transformed Dataset

of the normal distributions suggesting that the features, i.e., rental prices, demand, and home values have been dominated by the lower end of the spectrum of the spread which suggests that either the prices/demand increase for a shorter period or have increased in recent years.



Fig. 4. Continuous Housing Features from the Transformed Dataset

5 RESULT & LIMITATIONS

Output: The output of our pipeline is one CSV file in time series/long format composed of both Macroeconomic and Housing data saved in the data directory. The dataset is complete with no missing values and structured with temporal, categorical, and continuous variables, ensuring a broad spectrum analysis. Our pipeline performs actions in small steps thus making debugging easy in case of errors. Our pipeline is robust against data source changes, let's say, if for some reason the first extraction operation fails for a dataset our pipeline retries to download it three times before moving to the next dataset. Moreover, if any dataset out of 11 gets removed from the source website, our pipeline will not fail, and we will get a final dataset minus the missing dataset.

Reasons for this Type of Output: We have a temporal problem at hand and we needed final data in time series format, that is why our pipeline performs an extensive transformation on both datasets, i.e., applying ETL on 11 datasets then joining them into a single file with long format. So we needed a pipeline that is robust against data source changes & would do the job in small steps.

Critical Reflection on Limitations: The main limitation is the lack of availability of Macroeconomic data specifically for the individual Metropolitan areas of the USA that we are covering in our analysis that is why to analyze the relationship between our selected macroeconomic indicators and housing affordability in the top 10 Metropolitan areas of the USA we will compare baseline/average USA Inflation and Real Disposable Personal Income levels over time. However, the interest rates are by default applicable to every one in USA irrespective of their residence area.

The Most important features of Housing Data for our analysis are Home Price and Rental Price data and both of them are naturally available without any missing values from 2015 onward. So, another limitation could be the use of artificial intelligence to get 5 years' worth of Rental Demand data (2015-19), and 3 years' worth of Heat Index data (2015-17) as these datasets were only available from 2020 and 2018 onward respectively. But both the Rental Demand and Heat Index are secondary features for our analysis, hence the use of artificial intelligence to predict some historical data for them is not that problematic for our analysis because we use them as optional features for the sake of completeness to have a comprehensive look at the housing sector trends.

After trial & error we decided to use only Home Price data (SF/All) as main predictors for our regression because it gives statistically closer predictions as compared to available data. Due to this dependency we have limitation to always successfully extract Home Price data to meaningfully regress the missing values of target columns. Because, if our pipeline fails to extract Home Price data then the regression model would partially fail but would still regress values based on non-significant predictors and these predictions would not be statistically viable.

References

- FRED. 2024. *Federal Reserve Economic Data (FRED)*. <https://fred.stlouisfed.org/legal/#:~:text=FRED%20provides%20data%20and%20data,and%20conditions%20of%20the%20service>. Accessed: 2024-11-13.
- Zillow. 2024. *Zillow Research*. <https://www.zillowgroup.com/developers/api/public-data/real-estate-metrics/> Accessed: 2024-11-13.