



April 22, 2025

Housing Bubble Prediction

Contributor
Kapil Tare

Contributor
Aadit Malikayil





Executive Summary



Over the past few months, we've explored one big question: **Can we predict when the housing market might be headed toward a bubble?**

To answer that, we combined 50+ years of U.S. housing data — including home prices, construction trends, mortgage rates, and key economic indicators — into a single view. From there, we built:



A clean, structured dataset
(One Big Table) bringing all
sources together



Machine learning models to
predict home prices over
time



A custom **risk scoring**
system to detect early signs
of a housing bubble



A fully interactive
dashboard where anyone
can track the market and
generate live bubble risk
scores

Our final product helps users understand how the market is behaving, where it might be heading, and when things could be getting risky — all through a clear, data-driven lens.



Project Recap



When we started this project, our goal was simple: **understand what really leads to a housing market bubble — and see if we can catch it before it bursts**. Here's what we aimed to do:



Dig into the 2008 housing crash and figure out what signals led up to it — like fast-rising home prices, cheap mortgages, or overbuilding.



Bring together messy, disconnected housing data from different sources (FRED, Census, BLS) and organize it into a single, clean table we could use for modeling.



Build machine learning models that can forecast where housing prices might go next — and how much confidence we should have in those predictions.



Design a custom bubble detection system that scores each quarter based on risk factors like price spikes, momentum, and economic pressure.



Turn it all into a simple, visual dashboard that anyone — not just data scientists — can explore and learn from.

At its core, this project is about **making housing data more useful** — so we can spot the warning signs earlier and understand the market better.



Architecture Design



Detailed Architecture

Index	Focus	Purpose
1	End-to-End Project Overview	Big picture — What we built from start to finish
2	Data Ingestion & OBT Creation	Where the data came from and how we built the core table
3	Modeling & Risk Scoring	How we used that data to build predictions + risk signals
4	Dashboard Deployment	How it all came together in a usable tool



End-to-End Architecture Overview

Before diving into the pieces, here's the big picture. We started with messy, real-world data from multiple government sources — everything from home prices to inflation rates — and turned it into a smart, interactive dashboard that tells you when the market might be heading toward a bubble.

Along the way, we built:

- A clean data pipeline using Snowflake
- A combined table with all key housing & economic signals
- Machine learning models to predict home prices
- A scoring system to flag risky market conditions
- A Streamlit dashboard to bring it all to life

Think of it like this: from raw data → to risk insights → to a tool you can actually use.



Data Pre-Processing



Data Cleaning And Transformation

What We Did:

1. **Removed columns with too many missing values**
For example, Square_Feet_Floor_Area_Average had >50% nulls and was dropped.
2. **Imputed selective missing values**
Filled minor gaps in construction-type data with **0**, assuming no activity during that quarter.
Dropped initial rows in price index data that lacked historical context (for lag features).
3. **Standardized column naming**
Renamed columns to snake_case for easier tracking across SQL, Python, and Streamlit.
4. **Date alignment**
Ensured all rows were aligned by **quarter (period)**, our key time axis.
5. **Lag feature prep**
Made sure sufficient history was available before generating lag variables like price_lag_1, mortgage_lag_3, etc.



Data Ingestion & OBT Creation

We brought in datasets from multiple places — FRED, BLS, Census — and loaded them into Snowflake. The hard part wasn't just collecting data — it was **getting them all to speak the same language**

So we decided to:

1. Convert everything to **quarterly granularity**
2. Clean and standardized column formats
3. Finally, join all sources by **Period** to create our **One Big Table (OBT)** — the heart of the project

This OBT included:

1. Housing starts (1-family + multi-unit)
2. Mortgage rates, CPI, unemployment
3. Historical price index values

Everything we needed, clean and connected in one place.



OBT: Collecting Features

Here's what went into it:

1. **Data Sources Joined:**

We merged six different staging tables:

- a. One-family and multi-unit housing starts
- b. Mortgage rates
- c. Housing price index
- d. CPI (inflation)
- e. Unemployment

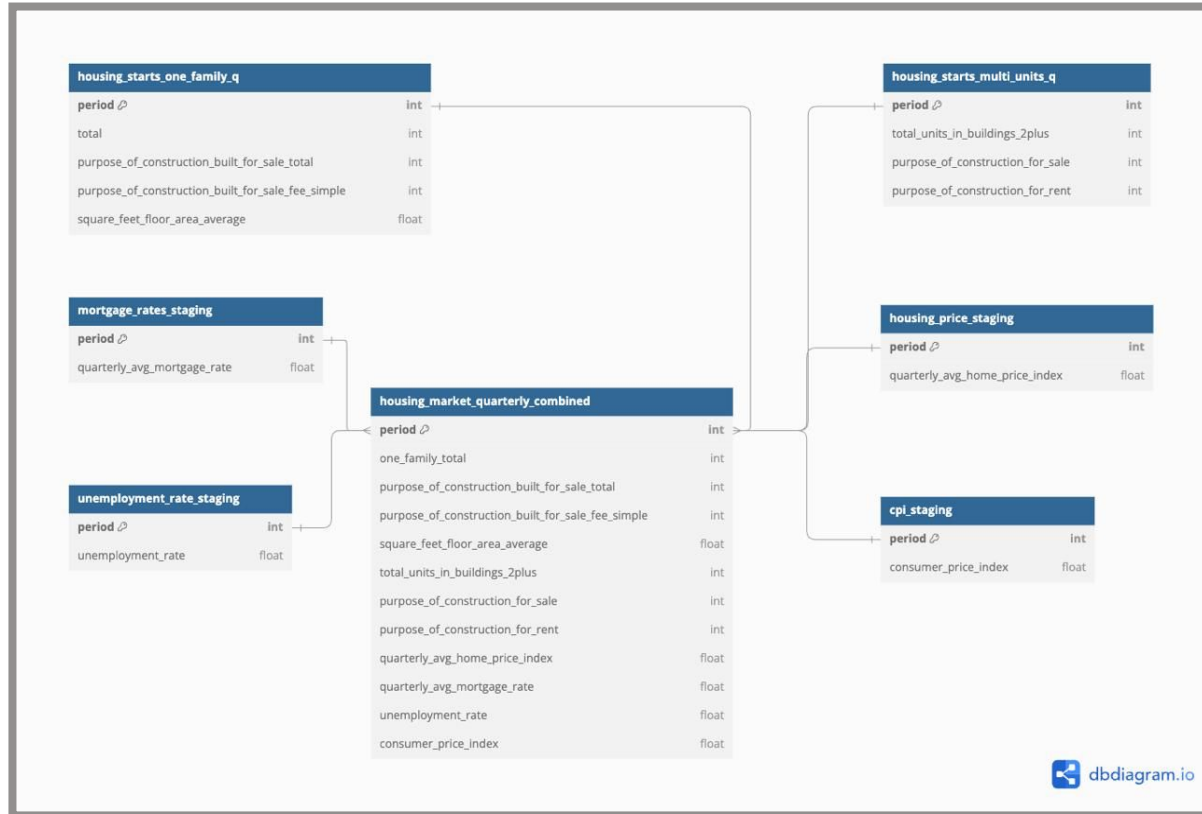
2. **Join Key:** All tables were joined on **Period (Quarter)** to ensure time alignment

3. **Key Fields in the OBT:**

- a. One_Family_Total, Total_Units_in_Buildings_2plus
- b. Quarterly_Avg_Home_Price_Index, Quarterly_Avg_Mortgage_Rate
- c. Unemployment_Rate, Consumer_Price_Index
- d. Purpose_of_Construction_*, Square_Feet_Floor_Area_Average



ER Diagram





Modeling & Risk Scoring



Modeling & Risk Scoring Overview

Once our data was ready, we used it to do two things:



Predict where housing prices were heading

We trained ML models (Linear, Ridge, Lasso) using a technique called walk-forward validation — so the models only saw past data to predict future prices.



Score each quarter for bubble risk

We created a custom scoring engine based on - Price momentum, Year-over-year spikes, Deviation from long-term average and Mortgage rate sensitivity

The higher the score, the more warning signs we saw — just like in 2006 before the 2008 crash.



Predicting Housing Prices – How We Built Our Model

We wanted to understand where the housing market might be heading — not just in general, but quarter by quarter. So we built a forecasting model that learns from the past and makes short-term predictions, just like analysts would in the real world.

What We Used as Inputs:

1. Mortgage Rates
2. Unemployment Rate
3. CPI (Inflation)
4. Housing Starts (1-family + multi-unit)
5. Lag Features (e.g., price from 1 or 3 quarters ago)

Models We Tried:

1. Linear Regression — simple, interpretable baseline
2. Ridge Regression — good for messy, correlated features
3. Lasso Regression — helps automatically filter out weak signals

We Didn't Just Use Random Splits because this is time series — we wanted our models to **see only the past** and predict the **future**, just like in real life. So we used something called **Walk-Forward Validation**, which can be understood in the next slide



Walk-Forward Simulation Logic

Firstly, we'll start by training on the first 80% of data, then we'll predict the very **next** quarter. After this we'll, Slide forward one quarter and train again. Finally, we'll keep repeating until the end

Example:

Iteration	Training Period	Predicting Period
1	2000 Q1 → 2010 Q4	2011 Q1
2	2000 Q1 → 2011 Q1	2011 Q2
3	2000 Q1 → 2011 Q2	2011 Q3

This helps us test how the model would perform **in real-time** — using only information it would've had **at that point in history**. So when we say “our model predicted well,” we really mean it — it wasn't cheating with future data.



Bubble Risk Scoring Logic : Part 1

After preparing our data, we looked at each quarter and ran it through four key checks:



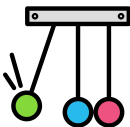
Annual Price Growth

How fast have prices grown compared to the same quarter last year?
We flag anything above 10%, 15%, or 20% as increasingly risky.



Price Deviation (Z-Score)

How far is the current price from its historical average?
A high Z-score means prices are *unusually high* — a classic bubble signal.



Momentum

Have prices been rising consistently across the last few quarters?
Sustained upward momentum adds to the overall risk.



Mortgage Rate Correlation

Are home prices moving in lockstep with falling interest rates?
When rates drop and prices rise quickly, it may suggest inflated demand.



Bubble Risk Scoring Logic : Part 2

Each of these indicators contributes a specific number of points depending on how extreme the values are:

1. Price growth over 20%? → +30 points
2. Z-score over 2.5? → +15–25 points
3. Sustained momentum? → +15 or +25 points
4. Strong correlation with mortgage rates? → +10 to +20 points

Final scores range from **0 to 100**, and are grouped into risk levels:

1. **0–40** → Low Risk
2. **41–60** → Moderate Risk
3. **61–100** → High Risk (possible bubble forming)

Important Note: This model doesn't predict *when* a crash will happen. Instead, it flags *rising risk* so we can see the bubble forming — before it bursts.



Bus Matrix



Bus Matrix

Business Process Name	Fact Table	Grain	Fact Grain Type	dim_date	dim_housing_starts	dim_macro_indicators	dim_mortgage_rates	dim_construction_type	dim_home_price_index
Home Price Index Prediction	model_predictions	One row per model-date combination	Periodic Snapshot	X	X	X	X	X	X
Mortgage Rate Correlation	bubble_risk_scores	One row per quarter	Periodic Snapshot	X	X	X	X	X	X
Annual Price Growth	bubble_risk_scores	One row per quarter	Periodic Snapshot	X	X	X			
Z-Score Price Deviation	bubble_risk_scores	One row per quarter	Periodic Snapshot	X	X	X			
Sustained Price Momentum	bubble_risk_scores	One row per quarter	Periodic Snapshot	X	X	X			
Bubble Risk Score Computation	bubble_risk_scores	One row per quarter	Periodic Snapshot	X	X	X	X	X	X






Dashboarding



Bringing It All Together

We didn't want this project to just live in Python files — so we built a clean, [interactive dashboard using Streamlit](#).

-  Tab 1 shows model predictions vs actual housing prices
-  Tab 2 shows risk scores over time — and lets users generate new ones
-  Every chart is interactive and tied to real data from Snowflake

So whether you're a professor, policymaker, or just curious — you can explore the housing market and see when things start to look risky.



Future Prospects

1. As of date and the amount of data available we were able to predict the trend and also make our model learn the changes in the price indexes
2. A pipeline can be triggered to run the model as and when more data will be available.
3. This trains the model on new data and also helps in predicting the future trend in the price indexes and check for a growing bubble to warn us against a falling market as of real time