Mortgage Data Description

Summary The dataset contains a sample of conforming mortgage loans purchased by Fannie Mae and Freddie Mac over 2000–2006. To focus on a homogeneous product, we restrict our sample to single family, 30-year fixed rate, purchase mortgages for primary residency purposes.

Loan-level variables

Variable	Description	Value
source	The government entity that purchases the loan.	FN: Fannie Mae; FD: Freddie Mac
loan_id	Unique identifier assigned to each loan.	12-digit string
Year_orig	The year of note origination.	numeric, no missing values
Quarter_orig	The quarter of note origination.	numeric, no missing values
delinquent30	Dummy variable which is 1 if the loan has ever been at least 30 days past due from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
frst_dte	The date on which the first full month of interest begins to accrue.	numeric, no missing values, MM/01/YYYY
orig_rt	The original note rate as indicated on the martgage note.	numeric, no missing values
orig_amt	The unpaid balance of the mortgage on the note origination date, rounded to the nearest $1,000.	numeric, no missing values
oltv	Original loan-to-value, dividing the original mortgage loan amount on the note date by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	numeric, in percentage points, no missing values.

ocltv	Original combined loan-to-value, dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	numeric, in percentage points, no missing values.
dti	Original debt-to-income ratio, dividing the sum of the borrower's monthly debt payments by the total monthly income used to underwrite the loan as of the date of the origination.	numeric, in percentage points, no missing values
cscore_b	Credit score at the origination date.	numeric, no missing values
mi_pct	Mortgage insurance coverage, the percentage of loss coverage on the loan in case of default provided by a mortgage insurer. Usually non-zero for loans with LTV greater than 80.	numeric, in percentage points, no missing values
fthb_flg	First time homebuyer flag, an indicator for whether the borrower is an individual who had no ownership interest in a residential property during the three-year period preceding the date of the purchase of the mortgaged property.	Y: Yes; N: No
num_bo	Number of borrowers, categorical variable, the number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property.	1 if there is 1 borrower; 2 if there are more than 1 borrowers
num_unit	Number of units, denotes whether the mortgage is a one-, two-, three-, or four-unit property.	numeric, 1 to 4, no missing values
state	A two-letter abbreviation indicating the state within which the property securing the mortgage is located.	2-digit string, no missing values

Macro variables

Data file	Description	Unit of observation
hpi_state.csv	The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties.	state-by-quarter level
rate.csv	The dataset downloaded from St. Louis Fed FRED contains 30-year fixed rate mortgage rates and 3 month treasury bill rates.	monthly level

Notes:

- In the mortgage dataset, we observe the quarter of origination and the month of first scheduled payment, but not the month of the origination. On average, there is a 45-day or two-month gap between mortgage closing and the first payment. When using macro variables, you might want to be careful about what information is already known at the time of prediction.