**Data and Code for:**
**"Why is Intermediating Houses so Difficult? Evidence from iBuyers"**

*Greg Buchak (Stanford GSB), Gregor Matvos (Northwestern Kellog),*
*Tomasz Piskorski (Columbia GSB), and Amit Seru (Stanford GSB)*

This file contains the (pseudo) data and code necessary to replicate "Why is Intermediating Houses so Difficult? Evidence from iBuyers." This document describes the structure of the replication package, the data, and the steps necessary to recreate all Tables and Figures in the body and appendix of the manuscript.

Questions should be directed to Greg Buchak (buchak@stanford.edu).

## 0. Certification

- I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

## 1. Structure of the replication package

The replication package contains the following directories:

- *data/*
  - *raw/ → contains raw (sampled) data*
  - *processed/ → contains processed (sampled) data*
- *out/*
  - *tables/ → contains tables*
  - *figures/ → contains figures*
- *data_prep → creates processed data from raw data*
- *reduced_form → does reduced-form analysis (and creates tables and figures)*
- *structural_model → does structural analysis (and creates tables and figures)*

## 2. Data sources

*Public data*: Our analysis uses (public use) data from the census for basic geographical demographic information, and market-level housing data from both RedFin and Zillow. The

replication package includes these full raw data under "*data/raw/census,"* "*data/raw/redfin,"* and *"data/raw/zillow":*

- Census: *data/raw/census/ nhgis0023_ds201_20135_2013_zcta.csv*
- Redfin: *data/raw/redfin/redfin_data.csv*
- Zillow: *data/raw/zillow/Metro_mean_doz_pending_uc_sfrcondo_month.csv*

*Census data*: Publicly available with free registration.

> Jonathan Schroeder, David Van Riper, Steven Manson, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 20.0 [dataset]. Minneapolis, MN: IPUMS. 2025. http://doi.org/10.18128/D050.V20.0
>
> The census data is at zip-code level, from the 2013 ACS. The raw data includes
>
> - Age by sex
> - Race
> - Hispanic or latino population
> - Means of transport
> - Travel time
> - Educational attainment
> - Median household income
> - Employment status
> - Rent as a % of household income

*Redfin data:* Publicly available without registration by scrolling to "Redfin Monthly Housing Market Data" and clicking the "Download" tab.

> Redfin. "Downloadable housing market data." Redfin News. Accessed November 11, 2025. https://www.redfin.com/news/data-center/.

*Zillow data*: Publicly available without registration by scrolling to "Days on Market and Price Cuts" and downloading the data for the Metro and U.S. regions.

> Zillow. "Housing Data – Zillow Research." Zillow. Accessed November 11, 2025. https://www.zillow.com/research/data/.

*Proprietary data*: Our analysis makes extensive use of two proprietary data sources, which we cannot share publicly due to licensing requirements. As described below, researchers can procure this data from the vendors under the licensing requirements. Data from CoreLogic contains property transaction records merged with property-level data from tax assessments. Data from the Multiple Listings Service (MLS) provided by ATTOM contains information on listings posted to the MLS.

*2.1.CoreLogic Deeds & Tax Records*

***Description.*** CoreLogic aggregates county recorder (deeds/mortgages) and assessor (tax/characteristics) records into a near-national property-level panel compiled from local government sources. Coverage spans multiple decades and includes both arms-length and non–arms-length transfers, with temporal and field-level completeness varying across jurisdictions. The dataset covers recorded transactions and tax/assessment records across U.S. counties and independent cities. Records can be linked within and across years using standardized addresses, enabling longitudinal analyses of property transfer histories.

***Main fields.*** Full property address; property use/type; structural characteristics (beds, baths, living area, lot size, year built); assessment values and tax amounts; transfer details (deed and recording dates, sale price); buyer and seller names; and, when available, financing fields (lender, mortgage amount, loan type).

***Variables used in this paper.*** We use recorded sale prices and property/structure characteristics, along with buyer and seller fields to identify transaction counterparties when names are available. Fields in the data are self-explanatory.

2.2. Multiple Listing Service (MLS) data via ATTOM

***Description.*** ATTOM licenses listing-level MLS data that provide real-time property listing information (e.g., price, status, dates) along with standard property attributes. Our MLS sample covers the selected study markets examined in the paper. These data capture the lifecycle of each listing—from initial activation through status changes such as pending/under contract to either closing or withdrawal/cancellation—together with associated timestamps, allowing construction of listing-hazard and time-on-market measures. Because MLSs differ in their business rules and field definitions, ATTOM applies a normalization layer; nonetheless, some cross-market heterogeneity remains in status codes and time metrics.

***Main fields.*** MLS/listing identifiers; property address; list and close prices (with price-change history); listing dates (list, contract/pending, close, withdraw/cancel); listing status (active, pending, closed, cancelled, etc.); days on market (DOM) and core property features. DOM generally measures the number of days a listing is in an active-for-sale status.

***Variables used in this paper.*** From the MLS data we use listing outcomes and dynamics, including time on market and status transitions, together with listing prices and their updates for the study markets. Where applicable, we rely on ATTOM's standardized status and date fields to ensure comparability across MLSs. Fields in the data are self-explanatory.

2.3.Access and acquisition of proprietary data:

**CoreLogic** (rebranded as **Cotality**):

- Data citation:

  Corelogic Deeds Data. “Our Data.” Cotality. Accessed November 11, 2025. https://www.cotality.com/our-data.

- Cotality offers bulk property, transaction and tax datasets for academic research. Several universities and regulatory institutions have already contracted with Cotality (via universities, higher-ed research centers, government agencies, Federal Reserve system) and have access to the data used in our paper.
- To inquire, the first step is to fill out the “Data Inquiry” or “Research Inquiry” form on the Cotality website (a form at the bottom of this page: https://www.cotality.com/our-data ). Pick “Education” as industry. A representative will respond.
- Academic pricing (“university bulk property data solutions”) is available but needs to be negotiated on an individual institutional basis and its terms are not disclosed publicly.
- Typical arrangements for institutional bulk licensing with Cotality may involve contract negotiations of several weeks—once the scope, usage rights, and extraction/delivery format are agreed.
- Usage rights, costs and contract terms are institution-specific; researchers should clarify whether they need raw bulk data, enriched/delivered slices, and what geographic/time coverage.
- Licensing typically requires an institutional agreement and imposes restrictions on redistribution and public disclosure of record-level information.

**ATTOM** (MLS)

- Data citation:

  ATTOM MLS Listings Data. “ATTOM Bulk Data Licensing.” ATTOM. Accessed November 11, 2025. https://www.attomdata.com/solutions/bulk-data-licensing/how-it-works/

- ATTOM provides the MLS data via bulk licensing, APIs or cloud delivery.
- To start, one submits a “Contact a Data Expert” or “Bulk Data Licensing Inquiry” form on the ATTOM website: https://www.attomdata.com/solutions/bulk-data-licensing/how-it-works/ the team typically responds within a week and sets up a call, followed by a test file and works through delivery format and usage specifications.
- Academic pricing is available but needs to be negotiated on an individual institutional basis.

- Researchers should ensure compliance with ATTOM's licensing terms (for example: data cannot always be redistributed, and may require on-site or secure environment hosting).
- As with Cotality, specify geographic scope, time period, variables (listing date, transaction price, tax assessment, etc.), and intended usage (e.g., academic non-commercial) to obtain a tailored quote.
- Licensing typically requires an institutional agreement and imposes restrictions on redistribution and public disclosure of record-level information.

2.4: *Sampled and randomized proprietary data included in the replication package*

To facilitate the use of the replication code, we provide randomized and scrambled raw data. The code that creates this data can be found as follows:

- *data_prep/1_prep_corelogic_data.r → create_sharable_raw_data()*
- *data_prep/3_prep_mls_data.r → generateSampledRawData()*

As these functions rely on the (unsharable) raw data, they cannot be run, but we provide them so that our data-creation process is clear. Briefly, they sample down the raw data (to allow for more convenient code execution) and randomly permute the data in each column. Thus, when running executing the code on this data, users should expect code that runs but produces different output from the published manuscript.

These functions generate, in their respective directories, a subdirectories called */share/* which contain sharable data. The CoreLogic data is accessible to us on a local server, and so the raw (sampled and permuted data) is provided in the "raw" directory. Additionally, since the MLS data analysis relies on merged "listing spells" it is useful for us to provide sampled and permuted spell data post-merging. Other functions included in the replication package will by default reference this sampled and randomly permuted data.

*Corelogic data included in replication package:*

- data/raw/corelogic/share/
  - *MaricopaOpendoorBase[YYYY].csv*: Phoenix
  - *32003OpendoorBase[YYYY].csv*: Las Vegas
  - *12095OpendoorBase.csv*: Orlando
  - *48113OpendoorBase.csv*: Dallas
  - *13135OpendoorBase.csv*: Atlanta

*MLS data* included in replication package:

- *data/raw/mls/share/*
  - *split_raw_1.csv*
  - *split_raw_2.csv*

    - ...
    - *split_raw_106.csv*
- *data/processed/mls/share/combined_processed.csv*: merged raw MLS data
- *data/processed/mls/share/spells_processed.csv*: processed MLS data by listing spells *data/processed/mls/share/ corelogic_mls_merged.csv*: processed MLS data merged with CoreLogic data for main analysis*data/processed/mls/share/*

*2.5 Attestations*

Data provided in the replication package will be preserved indefinitely.

Additionally, per the JPE replication policy, as we are unable to publish this restricted use data in the JPE dataverse, the authors commit to preserving the data and code for a period of no less than five years following the publican of the paper, and will provide reasonable assistance to requests for clarification and replication.

## 3. Software requirements

The analysis code was run on a x86_64-pc-linux-gnu platform, with R version 4.3.0 (2023-04-21). The replication code was additionally verified on a Windows machine running R version R version 4.2.1 (2022-06-23 ucrt).

The following R packages, available publicly, are used in the analysis:

- tibble
- data.table
- doParallel
- fixest
- ggplot2
- Hmisc
- lfe
- nleqslv
- scales
- stargazer
- stats
- stringr
- survival
- zoo

The expected run-time of the replication package with the simulated data is as follows:

- Reduced form analysis: < 10 minutes.
- Structural analysis: 2-3 hours.

## 4. How to run the package

Each of the code directories (*data_prep*, *reduced_form*, and *structural_model*) contain R files that generate all the analysis. Each directory also contains a single file "*_create_data.r*", "*_run_reduced_form_analysis.r*", and "*_run_structural_analysis.r*", which, when sourced, will create all of the tables and figures. To execute the code in R, make sure all dependencies are installed, and run the following commands:

1. source('_create_data.r')
2. source('_run_reduced_form_analysis.r')
3. source('_run_structural_analysis.r')

For tables, the code generates .html formatted output, which we further format before placing into the manuscript (e.g., not displaying coefficients for non-central controls, or describing the sample or fixed effects in the regression).

**5. Creation of specific tables and figures.**

The following contains a schematic of which function generates each table or figure in the manuscript:

***Reduced Form***

Table 1: Summary Statistics

- Panel A: *reduced_form/1_Summary_Stats.r*
- Panel B: *reduced_form/8_Listings_data_analysis.r*

Table 2: Transaction Behavior

- Panel A: *reduced_form/8_Listings_data_analysis.r*
- Panel B: *reduced_form/2_Pricing_Behavior.r*

Table 3: Pricing Determinants and Limits

- Panel A: *reduced_form/3_Pricing_Errors.r*
- Panel B: *reduced_form/3_Pricing_Errors.r*
- Getting liquidity MLS coefficients: *reduced_form/8_Listings_data_analysis.r*

Table 4: Gross Returns and Ease-of-Pricing

- Panel A: CoreLogic: *reduced_form/3_Pricing_Errors.r*

Figure 1: Summary Statistics

- Panel A: *reduced_form/1_Summary_Stats.r*
- Panel B: *reduced_form/1_Summary_Stats.r*
- Panel C*: reduced_form/1_Summary_Stats.r*
- Panel D: *reduced_form/1_Summary_Stats.r*

- Panel E: *reduced_form/3_Pricing_Errors.r*
- Panel F: *reduced_form/3_Pricing_Errors.r*

Reduced Form Appendix Tables

- Table 3: Matching Tie-Out: *reduced_form/8_Listings_data_analysis.r*
    - *Please Note: Panel A3.B, generated with simulated data, omits some rows because the sampled data does not include all cases (e.g., iBuyer buys that match within a day). The code will generate the full table when run with the full data.*
- Table 4: Listing Robustness: *reduced_form/8_Listings_data_analysis.r*
- Table 5: Coefficients from pricing errors: *reduced_form/3_Pricing_Errors.r*
- Table 6: Gross Returns*: reduced_form/5_pnl_analysis.r*
- Table 9: Model Validation
    - A.9.1: *reduced_form/6_Regional_Analysis.r*
    - A.9.2: *reduced_form/6_Regional_Analysis.r*
    - A.9.3: *reduced_form/7_Mobility_analysis.r*
    - A.9.4: *reduced_form/7_Mobility_analysis.r*
- Table 14: Speed of Transaction Listing: *reduced_form/9_Days_to_Pending.r*

***Structural***

Table 5: Calibration and Fit

- Panel A: Moments: This table is created by *structural/1_structural_model.r*; numbers in the table are derived as below.:
    - List price: *reduced_form/4_Extra_Calibration_Moments.r*
    - HH time on market: *reduced_form/8_Listings_data_analysis.r*
    - XS Price Variation: *reduced_form/4_Extra_Calibration_Moments.r*
    - TS Price Variation: *reduced_form/4_Extra_Calibration_Moments.r*
    - Impatient Price Delta: *reduced_form/8_Listings_data_analysis.r*
    - Impatient TOM: *reduced_form/8_Listings_data_analysis.r*
    - iBuyer Share: *reduced_form/4_Extra_Calibration_Moments.r*
    - iBuyer TOM: *reduced_form/8_Listings_data_analysis.r*
    - iBuyer Spread: *Table 2 Panel B*
    - iBuyer Buy Discount: *Table 2 Panel B*
    - Share Unprofitable: *reduced_form/4_Extra_Calibration_Moments.r*
    - dS / dLiq: *reduced_form/4_Extra_Calibration_Moments.r*
    - dS / dErr: *reduced_form/4_Extra_Calibration_Moments.r*
    - P(Cut Prices): *reduced_form/8_Listings_data_analysis.r*
- Panel B: External Calibrations: *(Citations inline)*

- Panel C: Calibrated Parameters: *structural_model/1_structural_model.r*

Table 6: Counterfactual

- Panel A: House Prices: *structural_model/1_structural_model.r*
- Panel B: Welfare: *structural_model/1_structural_model.r*
- Panel C: Shares/Time on Market: *structural_model/1_structural_model.r*

Figure 2: Counterfactual decomposition

- Model Code: *structural_model/1_structural_model.r*

Figure 3—5: Factors in Intermediation

- Model Code: *structural_model/1_structural_model.r*

Structural Appendix Tables

- Table 8*: structural_model/2_impatience_robustness.r*
- Table 10: *structural_model/1_structural_model.r*
- Table 11: *structural_model/1_structural_model.r*
- Table 12: *structural_model/1_structural_model.r*
- *Table 13: structural_model/1_structural_model.r*