Linking New York City housing sales to deeds and recorded mortgages

Pablo Loschi
Independent researcher, Berlin, Germany
loschi.pablo@gmail.com | ORCID 0009-0004-9455-4713

Abstract

New York City's sales and land-registry files contain complementary information but share no transaction identifier. Condominium sales are particularly difficult to link because the sales parcel often identifies the building, while deeds and mortgages identify a unit. This dataset connects the sources through documented rules for parcel, amount and date. It retains 729,047 residential sales dated 2016–2025. Among positive-price house and condominium sales in the four boroughs covered by the land registry, 97.2% of house sales and 96.0% of condominium sales are linked to a deed. Recorded-mortgage indicators use three windows around the sale. Deed-level flags describe possible company buyers, shared surnames, shared mailing addresses and lender acquisitions. Remaining unlinked condominium sales are concentrated at higher prices, so recorded-mortgage shares describe a selected subset. An offline example reports those shares alongside their linkage denominators. The deposit includes source extracts, code and documentation. Coverage and computational agreement are reported separately from accuracy, which has not been measured against verified instruments.
Keywords: housing; record linkage; administrative data; mortgages; condominiums; New York City

Related dataset

Loschi, P. (2026a). NYC residential sales linked to ACRIS deeds and mortgages, 2016–2025 (Version 1.2.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.23557585
This note documents dataset version 1.2.0. The version DOI and the deposited checksums identify the release used.

Introduction

A recorded sale price becomes more useful when it can be connected to the document transferring ownership and to any mortgage filed around the sale. These connections support studies of financing, transaction taxes and the construction of house-price indices. New York City publishes the relevant records, but researchers must build the connections themselves, and different linking rules can produce different samples and financing labels from the same public sources.
I built this dataset to provide a documented join for houses and condominium units sold in 2016–2025. It also retains co-operative apartment sales and Staten Island sales, marked as outside the deed linkage, and it supports my companion studies of tax thresholds, foreclosure consideration and the cash–mortgage price gap (Loschi, 2026b, 2026c, 2026d). Here I describe the files and their limits, with an example that a user can run without accessing the live registry.
The linking problem is relevant beyond New York, because administrative sources often record the same transaction using different property identifiers or dates. A reusable join needs to disclose how those differences are resolved, which records remain unmatched and which conclusions depend on the accepted matches.

The linking problem

The Department of Finance sales file reports a property's location, building class, sale date and recorded price (NYC Department of Finance, 2026d). ACRIS, the Automated City Register Information System, publishes separate tables for document headers, parcels and parties (NYC Department of Finance, 2026a, 2026b, 2026c). It covers Manhattan, the Bronx, Brooklyn and Queens, while Staten Island documents are recorded separately by the Richmond County Clerk.
The sales file has no ACRIS document identifier, and even where the sources share a borough-block-lot parcel identifier, dates and amounts can differ. Condominium sales often carry the building's billing lot, whereas the deed and mortgage use the unit lot. Several units on a block can sell for the same amount near the same date. A deed may also convey a unit together with parking or storage, and bulk apartment purchases can appear on

several sales rows at the full transaction price. Treating every candidate as a match would overstate what can be attributed to an individual sale.

Methods

I retrieved the source extracts through NYC Open Data on 26 September 2026, paging Sales, ACRIS Master and ACRIS Legals in fixed key order and checking retrieved counts against the source counts at retrieval. The deposit preserves these extracts and their query manifests. I used ACRIS Parties records to compute deed flags. The corrected release retains the earlier party extract and adds records for 8,190 newly linked deeds retrieved in October 2026; the names and addresses are not deposited.

I attempt linkage for positive-price house and condominium sales in the four ACRIS boroughs. A deed qualifies when its amount is within 0.5% of the recorded sale price and its document date is within 45 days of the sale. Recording date replaces document date when the latter is missing, and house deeds must name the sale's parcel. Condominiums are searched within the borough and block among unit lots 1001–6999, the range documented by the Department of City Planning (2022, pp. 2–3). A unique single-parcel candidate supplies the unit lot, and if several units qualify, the sales file's separate unit-lot field can select one. Where no single-parcel deed qualifies, a deed can match if it names that unit and at most two additional parcels and matches no other sale. Bulk purchases that fail this condition remain unlinked. A link requires a single candidate parcel, which is what the value `unique` in the `deed_match` field records. If several deeds qualify on that parcel, I keep the lowest document identifier; this selects one deed without showing which is correct. It occurs for 946 of the 305,175 retained links (698 houses and 248 condominiums). The `link_rule` field records the rule used.

The three condominium rules link 99,086, 12,715 and 9,462 sales respectively. Another 4,046 sales fail the bulk-purchase condition and are counted as unmatched. Users can restrict the sample to `single_parcel` links to retain that rule, although expanding the candidate range changes which units qualify. The rules use explicit parcel restrictions rather than the probabilistic matching framework of Fellegi and Sunter (1969) and Christen (2012).

Recorded-mortgage indicators search from 15 days before the sale to 30, 90 or 180 days after it. The search uses the house's own parcel or the condominium unit identified by the deed, and for condominiums matched through a multi-parcel deed, mortgages naming the unit with other parcels also count. The files give the number of mortgage matches, the largest recorded amount in the base window and its document identifier, with a separate flag for consolidation, extension and modification agreements.

Company, surname, mailing-address and lender-event flags are computed from deed parties using the released classification rules. The company flag identifies legal naming forms, which may differ from beneficial ownership or investment purpose, and a shared surname or address marks a possible connection between parties that I have not verified as a family relationship.

Data and coverage

The sales table has 729,047 rows: 388,791 house sales, 169,887 condominium sales and 170,369 co-operative sales. It keeps all recorded prices, including 213,457 zero-price entries, so users can choose an appropriate screen; zero-price entries should be kept apart from observed market prices.

The links table has one row per sales-table row. Users join them on `sale_id`, then join linked deeds to the 304,334-row deed-flags table on `deed_document_id`. A deed can appear on more than one sales row: 722 deeds link to 1,563 sales. Sales entries and distinct deeds therefore have different denominators.

Linking is attempted for 189,279 house sales and 126,251 condominium sales. Deed links are retained for 183,912 houses (97.2%) and 121,263 condominiums (96.0%). Table 1 reports the linkage and recorded-mortgage denominators. Houses have financing indicators even without a deed match because their financing parcel is known. Unlinked condominiums have no financing classification, because the deed match supplies their unit parcel, so blank financing fields must stay blank and not be converted to zero.

Table 1. Linkage and base-window mortgage counts

Property group	Eligible sales	Deed link	Deed coverage	Financing measured	Base-window mortgage match
Houses	189,279	183,912	97.2%	189,279	135,741
Condominiums	126,251	121,263	96.0%	121,263	67,950

Mortgage counts describe matching instruments; purchase purpose is unverified.

Condominium coverage is 94.7% in Manhattan, 98.8% in the Bronx, 97.7% in Brooklyn and 96.1% in Queens. It is 99.0% for recorded prices from $500,000 to below $1 million and 71.6% at $6 million or more. Linked condominium sales have a median recorded price of $978,000, compared with $4 million for unlinked sales. These differences describe selection into the linked sample; the release supplies no weights for recovering all condominium sales from that subset.

The deed table includes 67,065 company-buyer flags, 11,184 shared-surname flags and 45,354 shared-address flags, all screens whose error rates have not been measured. Party names are excluded from the derived tables, but the release is not anonymous: property addresses and public document identifiers can lead users back to identifiable public records.

A runnable example

The package's `examples/recorded_mortgage_example.py` reads its `build/sales.csv.gz` and `build/links.csv.gz` tables and produces annual condominium linkage and recorded-mortgage counts. It needs Python 3.10 or later, uses the standard library and runs offline:

```
python examples/recorded_mortgage_example.py build --output annual_condo_coverage.csv
```

In 2021, 16,736 of 17,734 eligible condominium sales link to a deed (94.4%). Of those linked sales, 11,106 have a base-window mortgage match (66.4%); financing is unmeasured for the other 998 eligible sales. In 2025, 11,976 of 12,497 eligible sales link (95.8%), and 5,657 linked sales have a mortgage match (47.2%). These percentages describe recorded mortgage matches among linked sales; they do not estimate the financing rate of all condominium purchases.

The example rejects duplicate join keys, missing eligible sales and financing assignments to unlinked condominiums, and its output keeps no-match and unmeasured counts separate. Users can adapt the same structure to report linkage denominators beside mortgage shares for another period or borough.

Reliability and reuse

Coverage, agreement and accuracy answer different questions (Bailey et al., 2020): coverage is the proportion of eligible records linked, agreement checks whether related code applied to overlapping sources returns the same labels, and accuracy requires evidence that a linked document belongs to the sale and that a mortgage financed the purchase.

Among the 99,086 condominium sales linked by the single-parcel rule, the recovered unit agrees with the sales file's separate unit-lot field for 98,661 (99.6%). This is agreement between two fields, and does not measure accuracy against verified documents. The other rules use the unit-lot field to select candidates, so its agreement cannot validate them independently. Comparisons with archived companion-study panels likewise measure computational agreement.

A positive mortgage indicator establishes a qualifying recorded instrument within a date window, though purchase purpose remains unverified, and a zero establishes no match under that rule; late-recorded loans, loans on other collateral and missed links can remain. Multi-parcel deed amounts apply to all conveyed parcels and cannot automatically be read as the apartment's price alone. Co-operative share financing is outside this real-property mortgage search.

The sales extract contains 336 rows whose BBL encodes a different block from the separate block column. Of these, 136 eligible sales remain unlinked. The release follows the BBL and does not substitute the other block automatically; candidate deeds under that alternative identifier require source checks.

The deposit contains a codebook, retrieval manifests and software instructions. Sales and linkage tables can be rebuilt offline from the archived extracts. Deed flags are a fixed snapshot: reproducing them requires party

records and a frequent-surname list that are not deposited. Comparison outputs also require the separately archived companion-study inputs, and a fresh retrieval from NYC Open Data may change results because records are revised or added. The `sale_id` key is release-specific; use the documented source or deed identifiers for comparisons across releases.

Derived tables, code and documentation are released under CC BY 4.0; source records retain New York City's open-data terms. I recruited or contacted no participants. Users should respect the source terms and the identifiability of the underlying public records.

Concluding remarks

The dataset makes an explicit connection between two public sources that users would otherwise need to link independently. I see its value in reusable rules, visible unmatched cases and outputs that preserve the distinction between a mortgage match and purchase finance. The coverage comparisons and runnable example help users decide whether the linked subset fits their question. Independent checks of accepted and missed links remain the next step toward measuring accuracy.

Declarations

The study received no external funding. The author founded HolaCasaClara, a comparator of Argentine mortgage loans with no connection to this dataset or New York housing records, and reports no other competing interests. Claude and ChatGPT assisted with coding, checking and writing. The author reviewed the work and is responsible for its content.

References

Bailey, M. J., Cole, C., Henderson, M., & Massey, C. (2020). How well do automated linking methods perform? Lessons from US historical data. Journal of Economic Literature, 58(4), 997–1044. https://doi.org/10.1257/jel.20191526

Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer. https://doi.org/10.1007/978-3-642-31164-2

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Loschi, P. (2026a). NYC residential sales linked to ACRIS deeds and mortgages, 2016–2025 (Version 1.2.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.23557585

Loschi, P. (2026b). The shape of the tax: Round prices and bunching at New York City's 2019 transfer-tax notches [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23414901

Loschi, P. (2026c). When a deed is not a market sale: Foreclosure transfers, statutory consideration, and repeat-sales house-price measurement [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23414919

Loschi, P. (2026d). Who gets the house? Related-party transfers and the cash–mortgage price gap in New York City [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23414584

NYC Department of City Planning. (2022). PLUTO data dictionary (May 2022, pp. 2–3). https://www.nyc.gov/assets/planning/download/pdf/data-maps/open-data/pluto_datadictionary.pdf

NYC Department of Finance. (2026a). ACRIS—Real property legals [Data set]. NYC Open Data. https://data.cityofnewyork.us/d/8h5j-fqxa

NYC Department of Finance. (2026b). ACRIS—Real property master [Data set]. NYC Open Data. https://data.cityofnewyork.us/d/bnx9-e6tj

NYC Department of Finance. (2026c). ACRIS—Real property parties [Data set]. NYC Open Data. https://data.cityofnewyork.us/d/636b-3b5g

NYC Department of Finance. (2026d). Citywide annualized calendar sales update [Data set]. NYC Open Data. https://data.cityofnewyork.us/d/w2pb-icbu