Project Update

Ben Jakubowski
Haonan Zhou

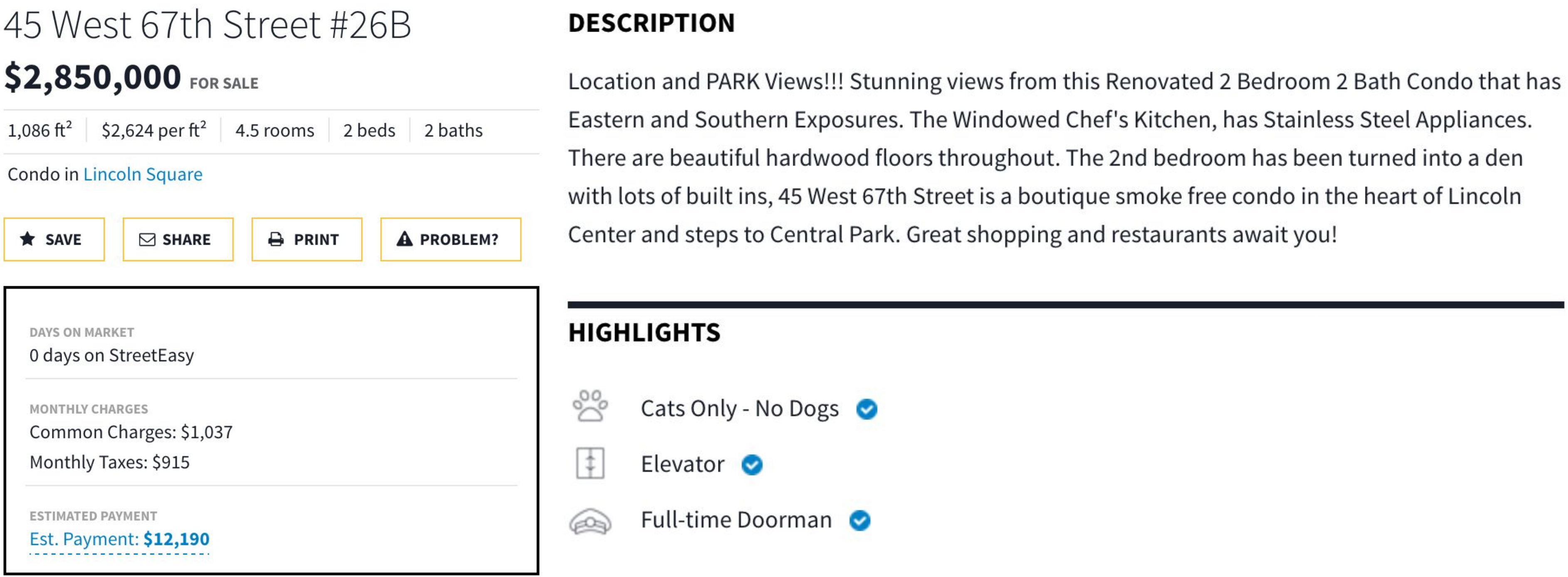
Update Outline

- Scraping StreetEasy Data
- Getting and cleaning NYC civic data:
    - Pluto
    - NYC Department of Finance Annualized Sales Data
- Merging datasets
- Questions

StreetEasy Data

- Each property unit has its own webpage that contains current or last listing price (and last closing price in some cases) along with a wealth of features
- Wrote python script to scrape this data using BeautifulSoup package
- Challenge is to find property listings in a systematic manner
- Iterated through current sales page for New York City and scraped data for all 12626 current listings
- Identifying current listings is straightforward:
  - http://streeteasy.com/for-sale/nyc?page=1
  - …
  - http://streeteasy.com/for-sale/nyc?page=974

StreetEasy Data - Features



45 West 67th Street #26B

$2,850,000 FOR SALE

1,086 ft^2 | $2,624 per ft^2 | 4.5 rooms | 2 beds | 2 baths

Condo in Lincoln Square

DAYS ON MARKET
0 days on StreetEasy

MONTHLY CHARGES
Common Charges: $1,037
Monthly Taxes: $915

ESTIMATED PAYMENT
Est. Payment: **$12,190**

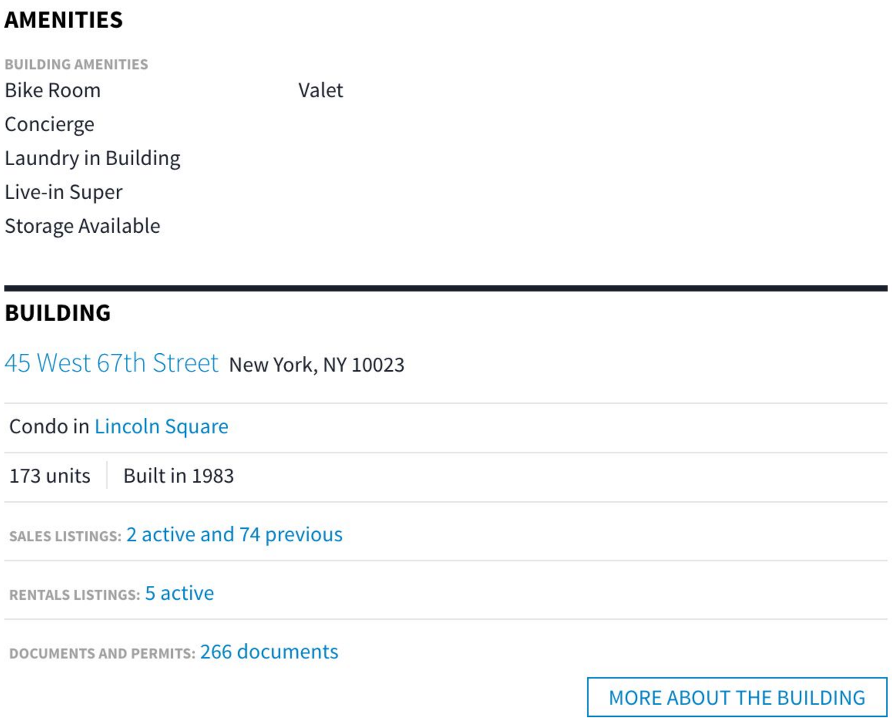
DESCRIPTION

Location and PARK Views!!! Stunning views from this Renovated 2 Bedroom 2 Bath Condo that has Eastern and Southern Exposures. The Windowed Chef's Kitchen, has Stainless Steel Appliances. There are beautiful hardwood floors throughout. The 2nd bedroom has been turned into a den with lots of built ins, 45 West 67th Street is a boutique smoke free condo in the heart of Lincoln Center and steps to Central Park. Great shopping and restaurants await you!

HIGHLIGHTS

- Cats Only - No Dogs ✓
- Elevator ✓
- Full-time Doorman ✓

StreetEasy Data - Features



AMENITIES

BUILDING AMENITIES

Bike Room
Concierge
Laundry in Building
Live-in Super
Storage Available
Valet

BUILDING

45 West 67th Street New York, NY 10023

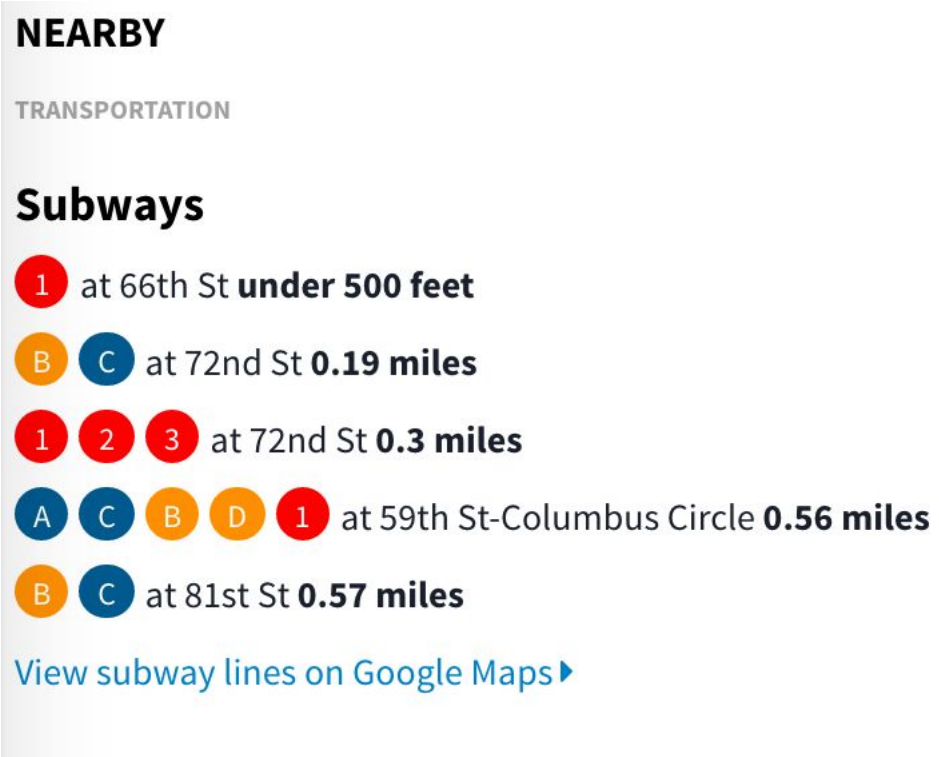
Condo in Lincoln Square

173 units | Built in 1983

SALES LISTINGS: 2 active and 74 previous

RENTALS LISTINGS: 5 active

DOCUMENTS AND PERMITS: 266 documents



NEARBY

TRANSPORTATION

Subways

1 at 66th St **under 500 feet**

B C at 72nd St **0.19 miles**

1 2 3 at 72nd St **0.3 miles**

A C B D 1 at 59th St-Columbus Circle **0.56 miles**

B C at 81st St **0.57 miles**

StreetEasy Data - Summary of Current Listings

Large range of prices:

- Mean: $2.84M
- Median: $1.29M
- Min: $45394
- Max: $96M

Missing data:

- 4187 listings (~30%) are missing square footage data
- Almost all listings (>98%) have number of bedrooms/bathrooms data

StreetEasy Data - Summary of Current Listings

Many different types of listings:

Condo	5158
Co-op	4235
Multi-family	1316
House	1054
Townhouse	572
Condop	156
Building	81
Apartment	29
Other	24

We will consider limiting the scope of our project to particular property types.

Streeteasy Data - Next Steps

- Scrape data for historical listings to expand size of dataset
- Restrict analysis to a subset of property types
- Use buildings pages to find links to past sales and listings
  - E.g. http://streeteasy.com/building/the-armory#tab_building_detail=2
- Straightforward to do since current data scraping script to can be used with very minor modifications for past sales and listings
- Challenge is constructing URL:
  - E.g. http://streeteasy.com/building/the-armory/7j vs. http://streeteasy.com/sale/1214429
- Clean and merge with PLUTO data

Civic Data Introduction

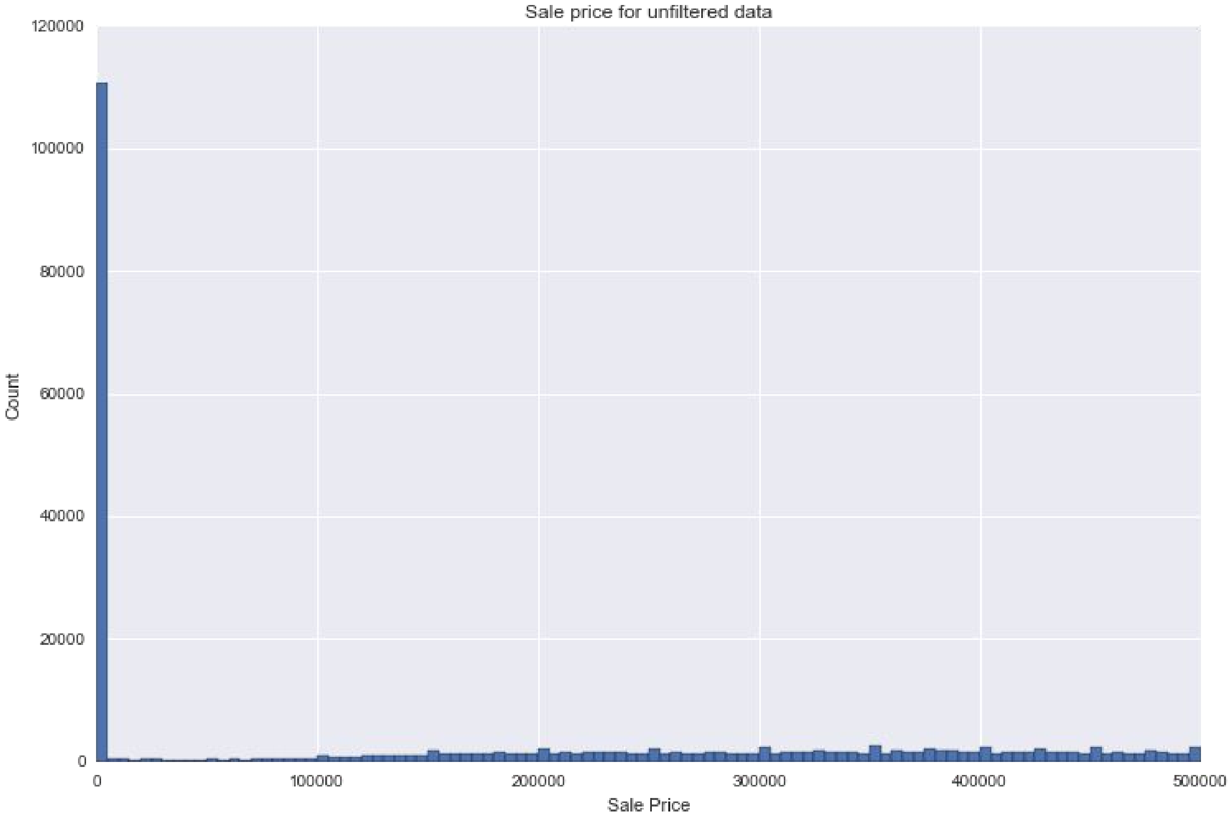
PLUTO:

- Primary Land Use Tax Lot Output (PLUTO) data.
- Developed by the New York City Department of City Planning's Information Technology Division (ITD)/Database and Application Development Section.
- It contains extensive land use and geographic data at the tax lot level.

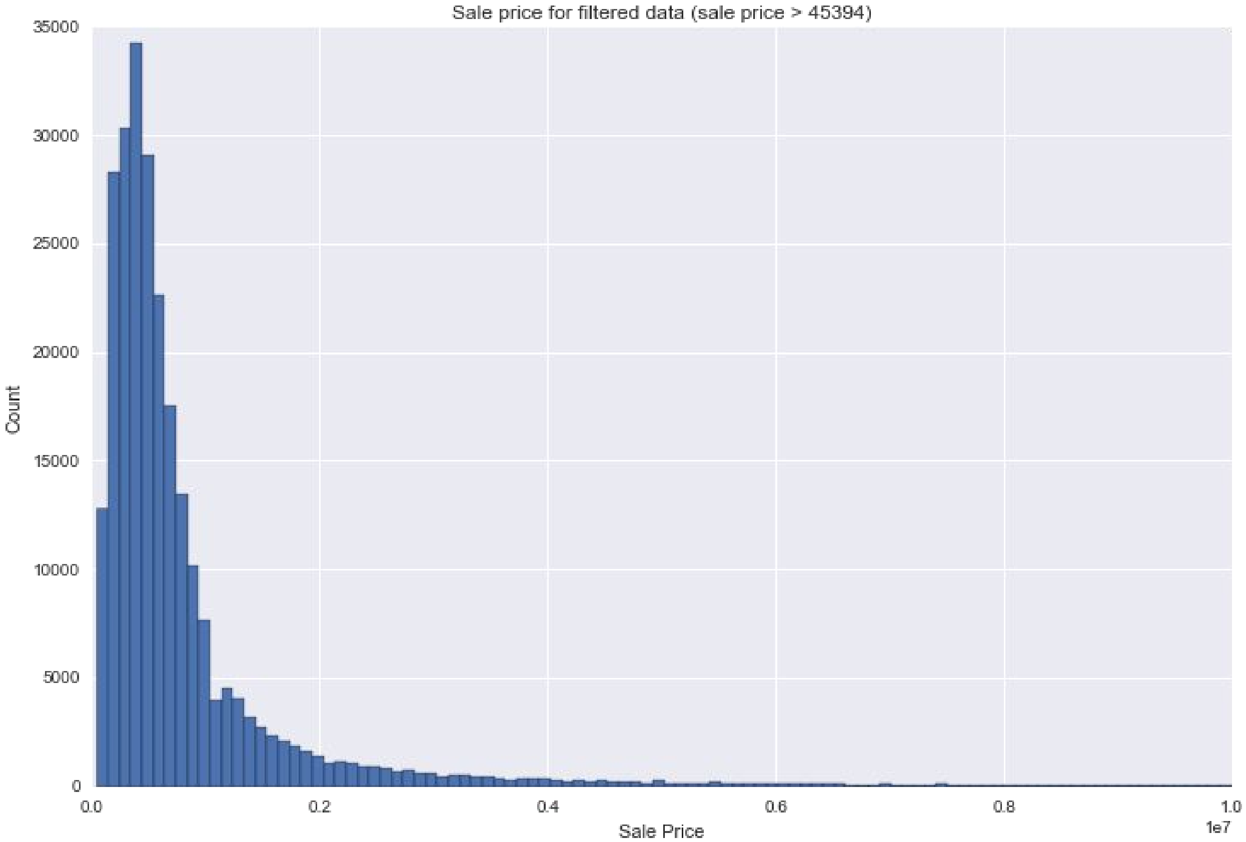
Annualized Sales Update:

- Yearly sales information of properties sold in New York City.
- Maintained by NYC Department of Finance.
- These files also have information such as neighborhood, building type, square footage and other data.

Sample Problems with Raw Civic Data



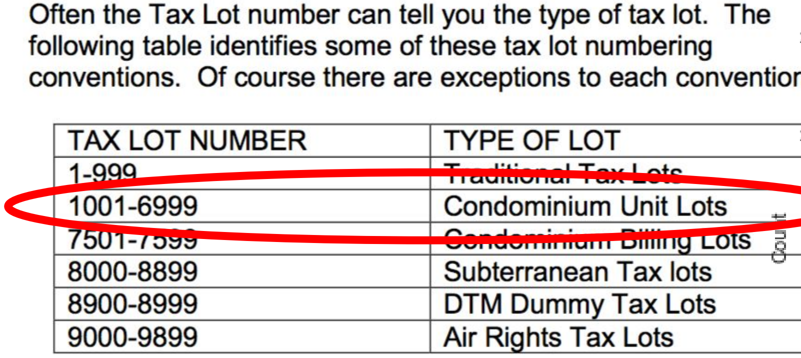
Sample Problems with Raw Civic Data



Merging Civic Datasets

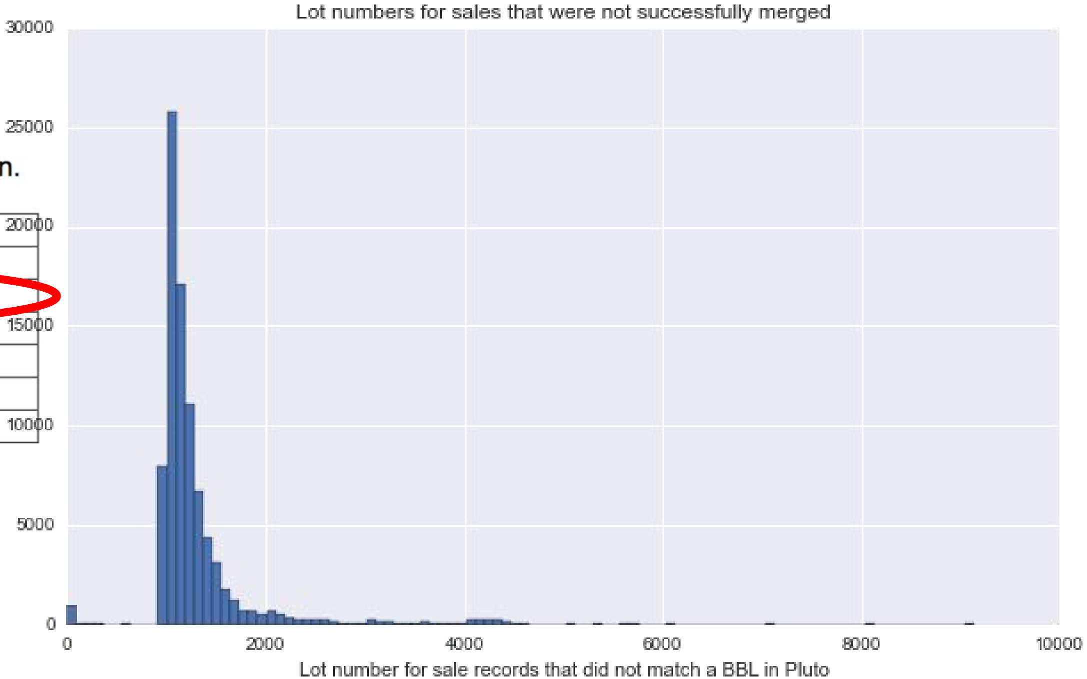
- Merge by BBL (Borough, Block, Lot number- a unique tax lot identifier).
- Merge issues:
  - Number of Sales for Tax Class 1 and 2 (residential), 2011-2015:
    - 369636
  - Number of Records where Sale.BBL matched PLUTO.BBL:
    - 281148

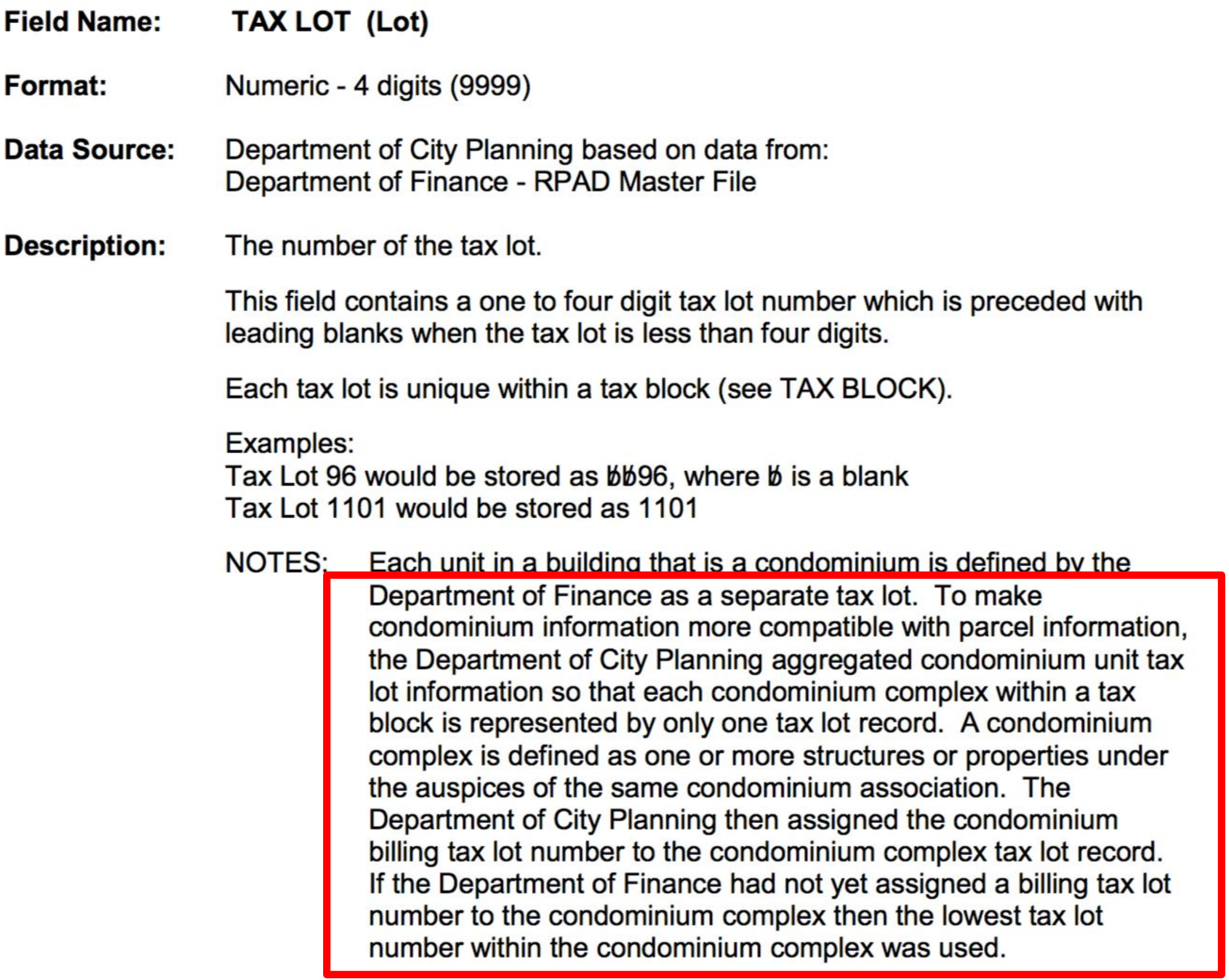
Merging Civic Datasets



Often the Tax Lot number can tell you the type of tax lot. The following table identifies some of these tax lot numbering conventions. Of course there are exceptions to each convention.

TAX LOT NUMBER	TYPE OF LOT
1-999	Traditional Tax Lots
1001-6999	Condominium Unit Lots
7501-7599	Condominium Billing Lots
8000-8899	Subterranean Tax lots
8900-8999	DTM Dummy Tax Lots
9000-9899	Air Rights Tax Lots





Field Name: **TAX LOT (Lot)**

Format: Numeric - 4 digits (9999)

Data Source: Department of City Planning based on data from:
Department of Finance - RPAD Master File

Description: The number of the tax lot.

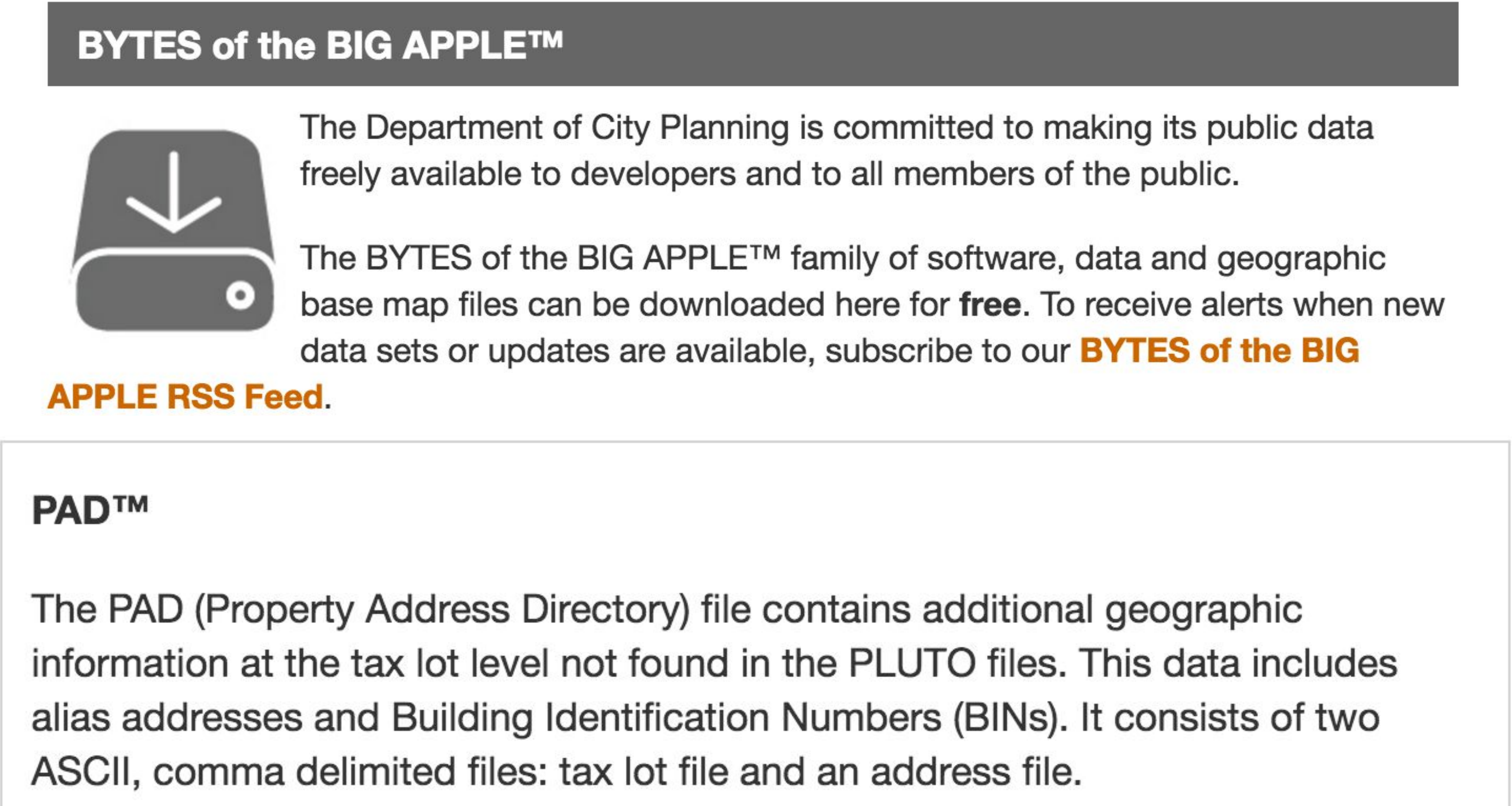
This field contains a one to four digit tax lot number which is preceded with leading blanks when the tax lot is less than four digits.

Each tax lot is unique within a tax block (see TAX BLOCK).

Examples:
Tax Lot 96 would be stored as ƀƀ96, where ƀ is a blank
Tax Lot 1101 would be stored as 1101

NOTES: Each unit in a building that is a condominium is defined by the Department of Finance as a separate tax lot. To make condominium information more compatible with parcel information, the Department of City Planning aggregated condominium unit tax lot information so that each condominium complex within a tax block is represented by only one tax lot record. A condominium complex is defined as one or more structures or properties under the auspices of the same condominium association. The Department of City Planning then assigned the condominium billing tax lot number to the condominium complex tax lot record. If the Department of Finance had not yet assigned a billing tax lot number to the condominium complex then the lowest tax lot number within the condominium complex was used.

Attempted Solution



BYTES of the BIG APPLE™

The Department of City Planning is committed to making its public data freely available to developers and to all members of the public.

The BYTES of the BIG APPLE™ family of software, data and geographic base map files can be downloaded here for **free**. To receive alerts when new data sets or updates are available, subscribe to our **BYTES of the BIG APPLE RSS Feed**.

PAD™

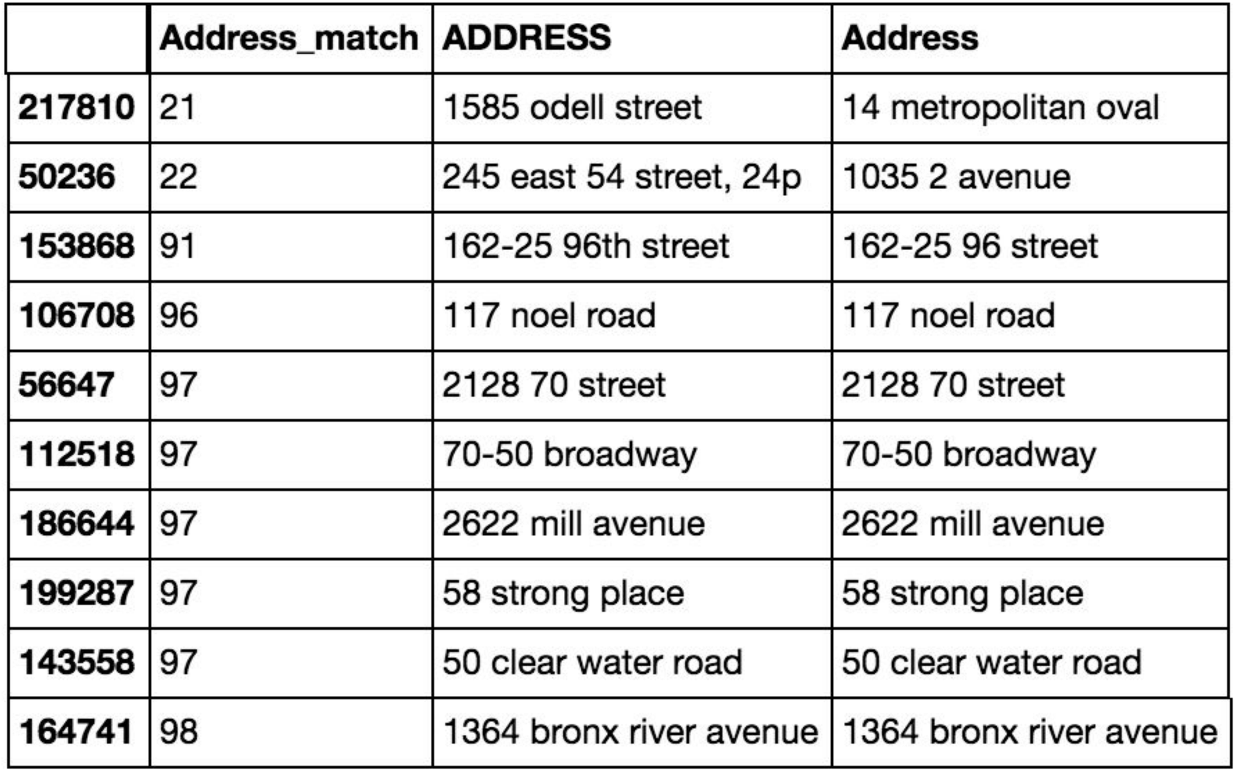
The PAD (Property Address Directory) file contains additional geographic information at the tax lot level not found in the PLUTO files. This data includes alias addresses and Building Identification Numbers (BINs). It consists of two ASCII, comma delimited files: tax lot file and an address file.

Attempted Solution

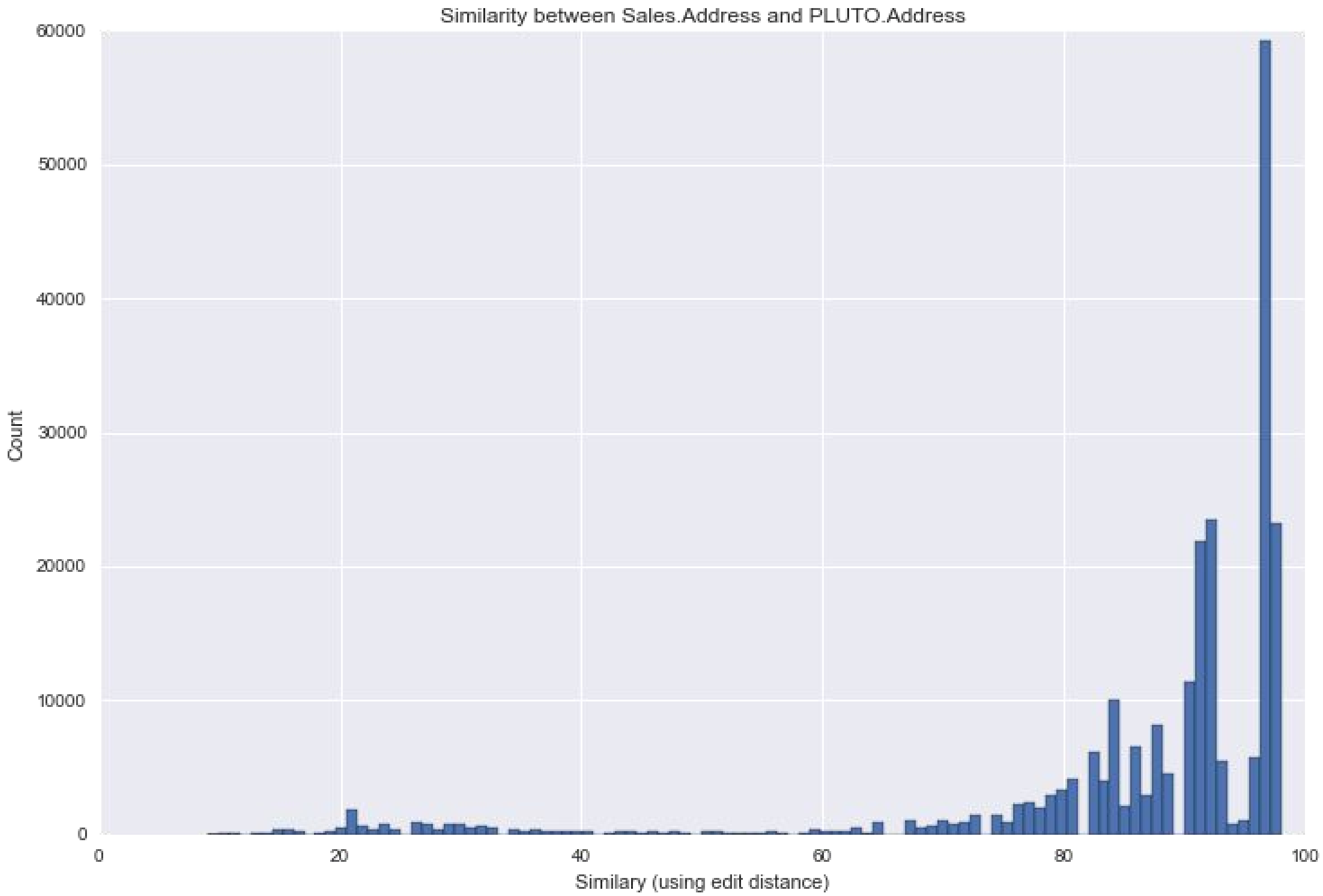
- In PLUTO, Condos are apparently listed by either (i) their condo billing BBL, or (ii) the lowest BBL in the range of BBLs contained in the condo.
- Using the condo billing BBL and BBL range provided by PAD, an intermediate merge allowed us to ultimately map 99.4% of sales records to PLUTO records.
- This introduced new problems, since some of the produced BBL matches between Sale and PLUTO records don't correspond to matching physical properties.

Issue with Sales/Pluto merge

This is apparent when comparing Sales.ADDRESS, PLUTO.Address, and Address_match, the similarity of the two strings (based on edit distance).

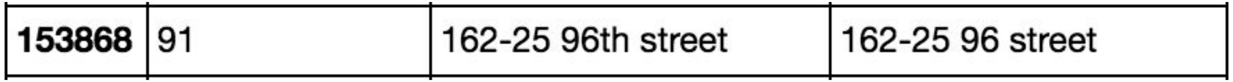


	Address_match	ADDRESS	Address
217810	21	1585 odell street	14 metropolitan oval
50236	22	245 east 54 street, 24p	1035 2 avenue
153868	91	162-25 96th street	162-25 96 street
106708	96	117 noel road	117 noel road
56647	97	2128 70 street	2128 70 street
112518	97	70-50 broadway	70-50 broadway
186644	97	2622 mill avenue	2622 mill avenue
199287	97	58 strong place	58 strong place
143558	97	50 clear water road	50 clear water road
164741	98	1364 bronx river avenue	1364 bronx river avenue


Similarity between Sales.Address and PLUTO.Address
Count
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100
Similary (using edit distance)

Civic Data - Next Steps

- Parse and format the Sales.ADDRESS and PLUTO.address features to address obvious differences, such as



153868	91	162-25 96th street	162-25 96 street

- Define a threshold edit distance at which we accept the match; alternatively, ideally parsing the addresses and ensuring consistent formatting will allow us to identify matching addresses perfectly.
- Use address as a field to potentially generate URLs for properties for StreetEasy Webscraping.

Questions?