Early Warning Data Project

**Goal:**

*Support the development of a usable database containing opensource property information.*

**Major problems (for Lucy):**

* BBLs being thrown out in DoB set b/c trash text
* **Some HPD violations have approved date missing?**

**To-Do:**

* ~~transform dates~~
* ~~change HPD Violations date to HPD approved~~
* ~~error check HPD violations with both opendata and department pages (ex: 3044540060)~~
* ~~put total # hpd complaints on lookup page~~
* ~~add code keys to raw data sheet (for error checking)~~
* zero problem in excel sheet
* ~~fill-in cells for HPD viols~~
* ~~documentation of data sets~~
* apartment # error?
* ~~sorting by date: can I do in excel sheet, or need in output?~~
* DoB complaints are only since 2014 – do the same for other fields??
* ~~De-dup DoB complaints by ID~~
* ~~TABLES WITH DATES: can do tiered sorting in pandas:~~ DF.sort\_values([col1, col3]) will return DF sorted by column 1, and internally by col 3
* manual data:
  + ~~Furman center core data~~
  + ~~tax bills – link from Lucy email, possibly talk to John Krauss~~
* implement ACRIS and HPD:

ACRIS

-Update ACRIS data to call up every most recent mortgage or deed filed for each property (doc type will only say "mortgage" or "deed" with no extra words)

-For ACRIS data display a zero if the doc amount is zero instead of it being blank

-reformat date on summary page for ACRIS data as m/d/yyyy like on property report

~~HPD~~

~~Complaints criterion: change to~~**~~all~~**~~complaints, open and closed, ever. The count should also be that number.~~

* ~~error check Lucy’s problems (see below)~~
* error catching and resilience measures
* final prep and install
  + ~~Excel links~~
  + tie everything together -- .sh file
  + set up github
  + Archiving – check Lucy emails
  + get everything on DuBois’ computer

**Key Links:**

* [Pandas docs](http://pandas.pydata.org/pandas-docs/stable/dsintro.html)
* Sources:
  + HPD [complaints](https://data.cityofnewyork.us/Housing-Development/Housing-Maintenance-Code-Complaints/uwyv-629c/data) and [problems](https://data.cityofnewyork.us/Housing-Development/Complaint-Problems/a2nx-4u46/data)
  + HPD [violations](https://data.cityofnewyork.us/Housing-Development/Open-HPD-Violations/csn4-vhvf/data)
  + DOB [Com](https://data.cityofnewyork.us/Housing-Development/DOB-Complaints-Received/eabe-havv)plaints
  + DOB Violations
  + [DOB ECB](https://data.cityofnewyork.us/Housing-Development/DOB-ECB-Violations/6bgk-3dad/data) Violations
* possible [source code](https://gist.github.com/clhenrick/d4c7aea465071363097d/revisions) for getting BBL from street address
* street address [webpage](http://a030-goat.nyc.gov/goat/Default.aspx) using URL search
* [SODA docspodshello](https://dev.socrata.com/consumers/getting-started.html)street address [webpage](http://a030-goat.nyc.gov/goat/Default.aspx) using URL search
* **ACRIS:** [RP Master](https://nycopendata.socrata.com/City-Government/ACRIS-Real-Property-Master/bnx9-e6tj/data), [RP Legals](https://data.cityofnewyork.us/City-Government/ACRIS-Real-Property-Legals/8h5j-fqxa/data), [RP Parties](https://data.cityofnewyork.us/City-Government/ACRIS-Real-Property-Parties/636b-3b5g/data)

**Implementation Notes:**

DOB Complaints:

* no BBL or zip code; only address and community board, BIN

DOB Violations:

* only BBL
* issue date is TEXT
* even restricting to active violations in Brooklyn, it’s >200,000
* perhaps if small-ish number of streets?

DoB ECB Violations:

* BBL and zip code available; very short list, likely will have none

DoB Permits Filed (Job applications)

* community board and BBL available
* need sum per BBL and change in proposed units

DoB Permit Issuance

* BBL, community board, and zip code available

**Program Plan:**

* one program for all #-per-BBL fetches
* customized function for each datasource – getting a small enough batch
* merge BBLs (or get BBLs)
* for simple counts, Pandas method to return number of instances in a column?
* for Permits filed, iterate through and increment BBL count

**ACRIS Fix**

In Python:

* call real property legals as in current program
* eliminate all document entries with non-ENY BBLs
* request *every document ID* from Real Property Master
* eliminate all non-DEED and non-MORTGAGE entries
* join with parties, etc

In Postgres:

* upload csv files
* join legals to master by Doc ID
* Combine BBL fields (use python??)
* eliminate all non-ENY BBLs and all non-DEED and non-MORTGAGE
* order by BBL
* purge all but most recent for each BBL (python?)

**Error-check HPD Violations**

TO DO:

DOB

There's still the filed unit change number that seems funky. Could you send me documentation/remind me what our parameters are for DOB stuff? I just can't make heads or tails of it.

Example discrepancy properties: [3039740001](tel:(303)%20974-0001), 3100 Atlantic Ave. DOB numbers very different. Possible explanation\*

*Ton of buildings, same lot -* [*http://a810-bisweb.nyc.gov/bisweb/PropertyBrowseByBBLServlet?requestid=0&allborough=3&allblock=3974&alllot=1*](http://a810-bisweb.nyc.gov/bisweb/PropertyBrowseByBBLServlet?requestid=0&allborough=3&allblock=3974&alllot=1)

Also to document/investigate: \*it's possible that you'll only get one building up when you click on the DOB link, and the numbers split between all properties on a lot (e.g. 3100 Atlantic Ave)

[3043120034](tel:(304)%20312-0034) 796 Cleveland St has 5 complaints and 4 issues in EWS but looks like way more on property page.

I’m using respondent\_zip to get ENY – but respondent for some of these violations is listed at address in Williamsburg.

Workaround for now: download whole CSV (1.15M rows), upload into Pandas

why does the 'filed for change in number of units' have a number value rather than a yes/no (or 0/1) value? e.g. 778 livonia ave 3040880019 has the number 15.

HPD complaints BBL [3037920038](tel:(303)%20792-0038)

Complaint History on HPD website doesn’t show if open or closed – the complaints that were missing from EWS may have been already closed

* they’re in the newly-implemented complaints system
* random check, careful matching
* record which BBLs to check
* first check raw data against both OpenData and HPD website; then check workbook

3044540060:

* 12 entries in sheet match 12 in opendata
* apparent duplicate in raw data 🡪 Opendata has same violation on different dates and different violation ID, but website deduplicated it
* 252 DoB complaints in raw data, but 221 on DoB website.

3040880036

* HPD complaints and violations match up
* DoB complaints: opendata has duplicate entries from complaints being updated, but it doesn’t have complaints older than ~ 2010. **Solution: de-dup by complaint ID**

3038400031

* DoB jobs: website count is “15”, but these are from pre 2010 jobs and duplicate entries of updates on same job. Worksheet is accurate

**Lucy’s Notes:**

Here are the other data sources:

HPD Violations: all open violations for each property in dataset. For reference when I did this manually, got ~26,000

<https://data.cityofnewyork.us/Housing-Development/Open-HPD-Violations/csn4-vhvf/data>

DOB Data: I just need the count of each of these per property (except #4, see below), because a link to the DOB page for the property quickly gives you all the info you need. See the attached doc for reference. We could do it ourselves, but is there any reason to? But the fields I need a count for are:

1. DOB Complaints <https://nycopendata.socrata.com/Housing-Development/DOB-Complaints-Received/eabe-havv/data>

2. DOB Violations <https://data.cityofnewyork.us/Housing-Development/DOB-Violations/3h2n-5cm9/data>

3. DOB ECB Violations <https://nycopendata.socrata.com/Housing-Development/DOB-ECB-Violations/8vf4-ne9x/data>

4. DOB Permits Filed <https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2/data>

5. Filed for change in Units - this refers to two columns in DOB Permits Filed. This field should have a positive or negative integer based on the difference between fields "Existing Dwelling Units" and "Proposed Dwelling Units" (positive for proposed increase, negative for proposed decrease)

6. DOB Permits Issued <https://nycopendata.socrata.com/Housing-Development/DOB-Permit-Issuance/ipu4-2q9a/data>

**Other types of data to integrate:**

DOF Documents Filed: I think up to you if you want to get the data from ACRIS or from the OpenData API where you have to link tables together. We want a report of all the documents filed for a property (I think ever, it shouldn't be too unwieldy) and most importantly, a flag if there's any new document filed for a property. The information that would be relevant in a dashboard would be the document type, recorded date, document date, and amount. It would be great if there were a link that searched automatically in [ACRIS](https://a836-acris.nyc.gov/DS/DocumentSearch/BBL), which is not something I've seen anyone do so far. But at the very least, Dubois wants to know when a new document is filed so he can pull it up in ACRIS and look at the actual document images.

Rolling Sales Data: this is released monthly in pdf and Excel format [here](http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page). It probably always has the same file format, and Dubois is interested in the time period as it's published, which is the previous 12 months updated each month (e.g. in December 2016 you get December 2015 to November 2016). We would want all sales for each of our properties.

Evictions: This data is available through OpenData (<https://data.cityofnewyork.us/City-Government/Evictions/6z8x-wfk4/data>), but it doesn't have a BBL field. Only an address. I can manually do a spatial join in GIS, but I don't think that solves our automation problem. I just found this [BBL lookup](http://webapps.nyc.gov:8084/CICS/fin1/find001I), I wonder if there's a way to query it automatically? And how reliable it would be, based on the format of addresses as they're entered in OpenData.

Manual data: We've also got bad acting banks to flag, sales listings, and differences between legal and observed units in a building, all which would probably be entered manually by someone in the office. I think this is a conversation to hold off on, assuming you have enough to do with the other stuff!

**Overview of Script Outputs:**

**Heatseek Housing Data Tool**

* sources: 311 complaints, HPD complaints, HPD registrations, DoB Permits, DoB Violations, Pluto, Rent Stabilization, Alternative Enforcement Program
* *not included:* DoB Complaints, DoB ECB Violations, HPD Violations, HPD Problems, DoF, manual data
* this git repo is meant to get central SQL database, create reports for activists, and publish online