DRAFT

DATA ENGINEERING INTERNSHIP

Jingyi Huang and Andrew Brieff

NYC PLANNING

Jingyi Huang

Enterprise Data Management; Data Engineering

- **Hometown**: Beijing, China
- **Current Residence**: Pittsburgh, Pennsylvania
- **Studying**: Public Policy and Data Analytics at Carnegie Mellon University
- **Passionate about:** Data-driven urban policies, transportation justice, and anthropogeography
- **Loves**: Indie Rock, Photography, Documentary Film

Andrew Brieff

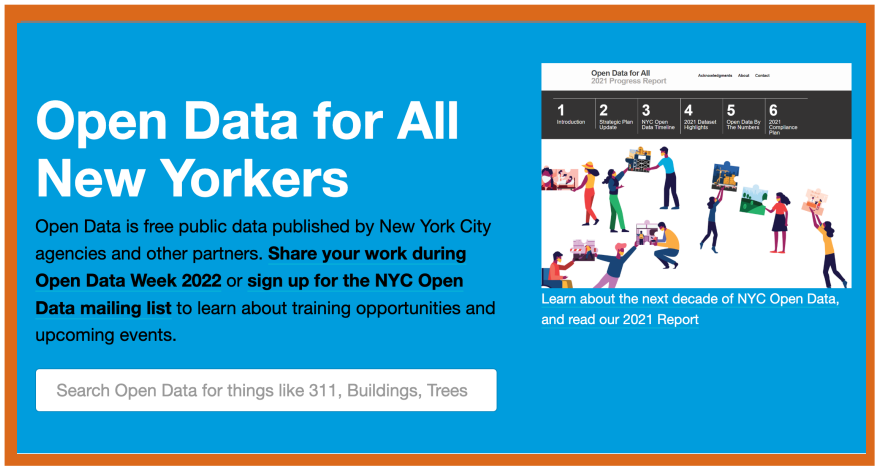
Enterprise Data Management; Data Engineering

- **Hometown**: Huntington, NY
- **Current Residence**: Somerville, MA
- **Studying**: Urban and Environmental Planning at Tufts University
- **Passionate about:** Civic Tech, Sustainable Transportation and Climate Change
- **Loves**: Hiking, Kayaking, Exploring, Live Music

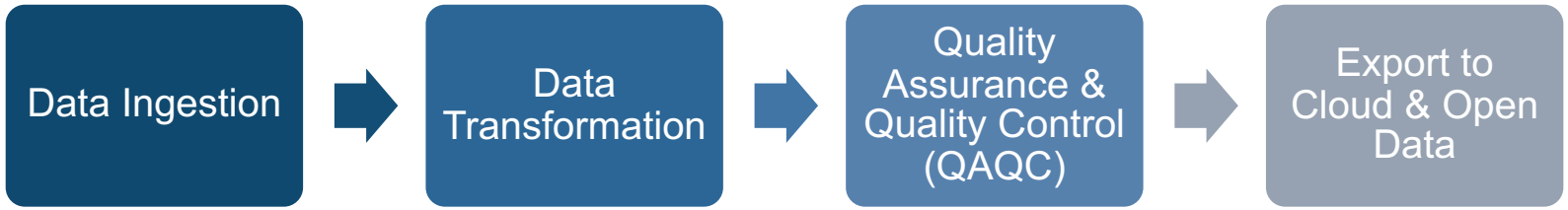
Data Engineering

What we do:

- Manage DCP's core data products
- Reimagine and improve data pipelines
- Improve accuracy of underlying data
- Use modern open-source technologies (PostGIS, Python, GitHub Actions)



Source: NYC Open Data

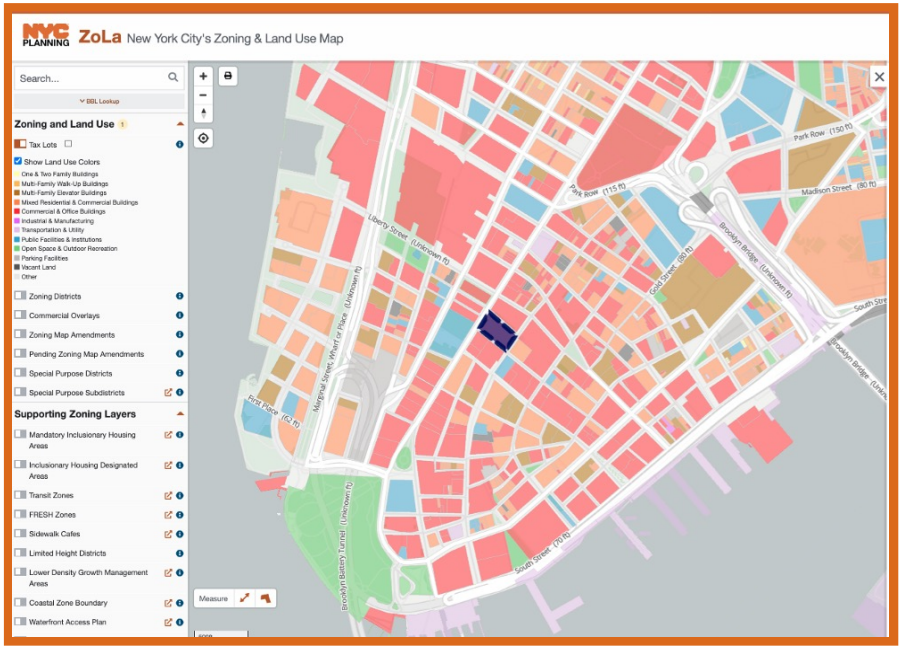


What Data Products did we work on?

PLUTO

What is PLUTO?

- Primary Land Use Tax Lot Output
- Combines Data from:
  - Department of City Planning (DCP),
  - Department of Finance (DOF)
  - Department of Citywide Administrative Services (DCAS)
  - Landmarks Preservation Commission (LPC)
- One record per tax lot (except condominiums)
- Contains extensive tax lot and building characteristics (>70 fields)
- Includes geographic/political/administrative boundaries

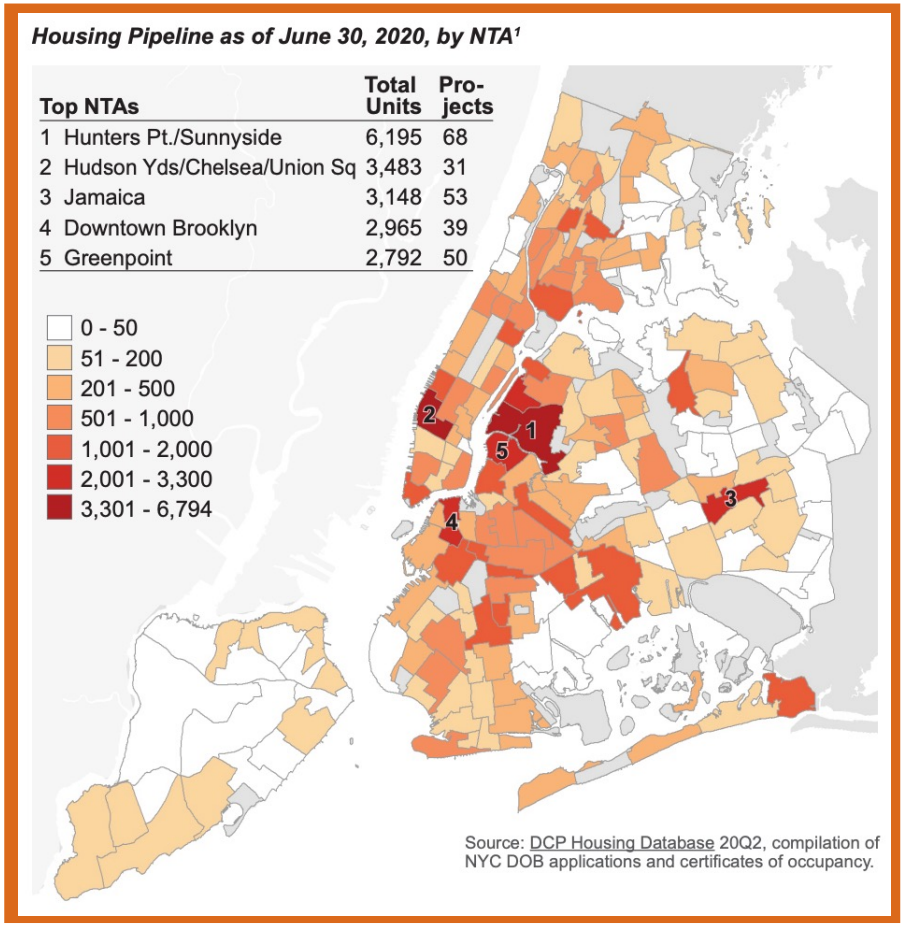


ZoLa: NYC's Zoning and Land Use Map, with data from PLUTO

Housing Database

What is the Housing Database?

- Based on DOB job and permit information
- Tracks housing production across the city
- We add geocoding
- Can be aggregated at various political, geographic and administrative boundaries
- Underlying data is self-reported, can lead to inaccuracies



Source: 2020 Mid-year Housing Production Snapshot

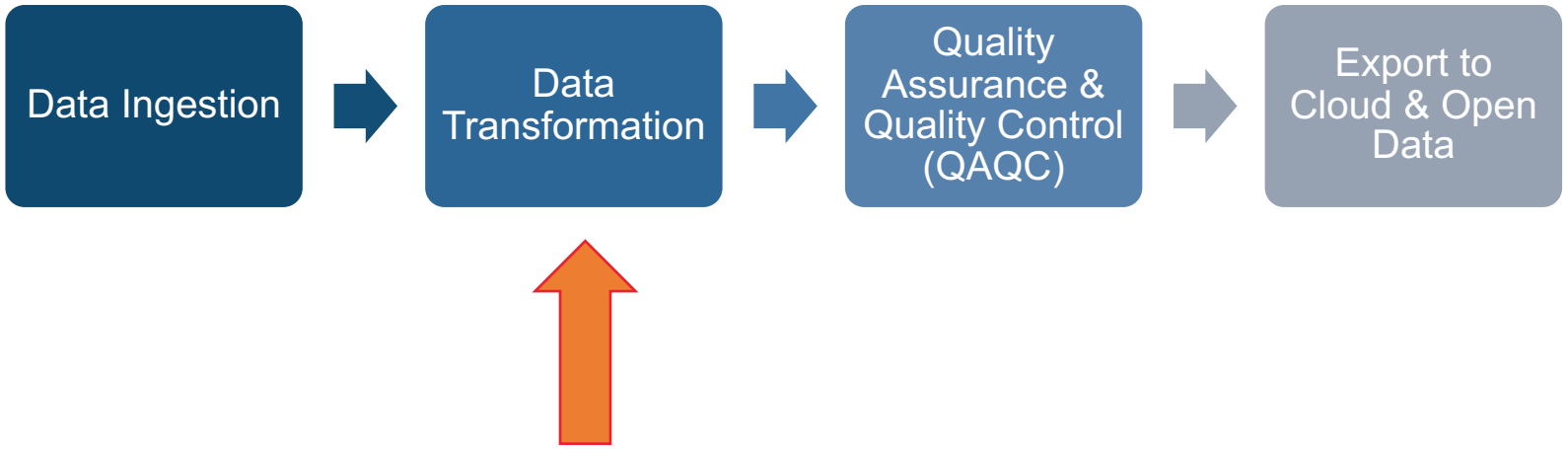
City Owned and Leased Properties Database

What is the City Owned and Leased Properties Database (COLP)?

- Input data is the Integrated Property Information System (IPIS), a real estate database maintained by DCAS.
- Contains each property's use, owning/leasing agency, location, and tenant agreements
- Existing QAQC checks mainly are designed to identify invalid data in IPIS
- Some are designed to help GRU research potential new addresses

Gracie Mansion, Source: Jim.Henderson, CC0, via Wikimedia Commons

Data Pipeline: Transformation



Data Transformation - PLUTO

What does Data Transformation look like for PLUTO?

- Standardizing precision of fields like Lot Depth, Building Frontage
- Standardizing Address Formatting
- Populating missing data (for example, using COLP to populate owner type)

Data Transformation - PLUTO

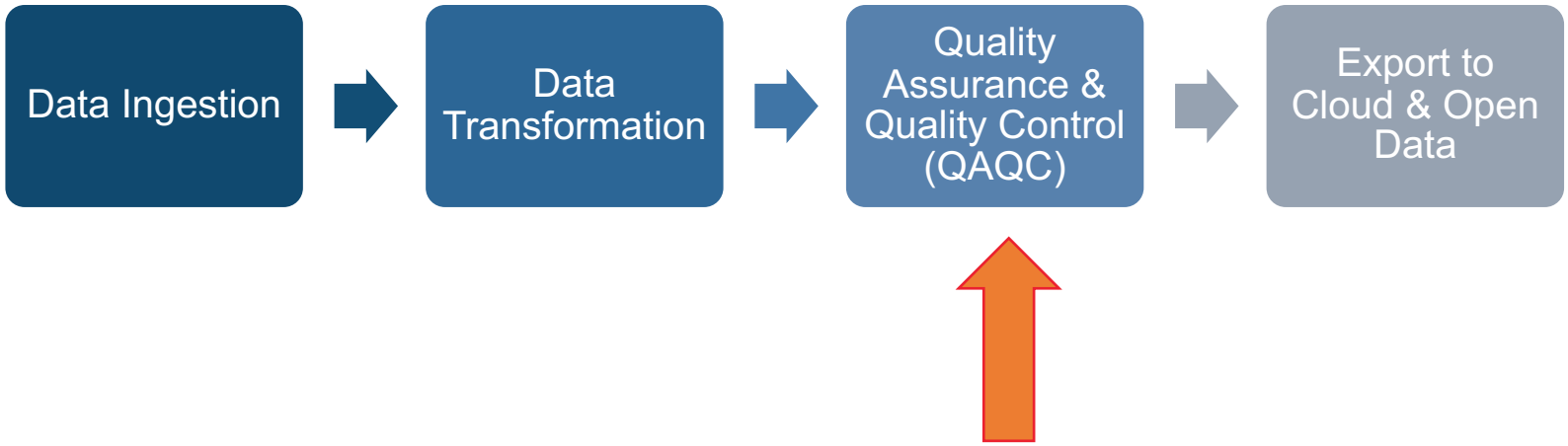
Enhancements we implemented:

- Adding missing geographic information by calculating lot centroids
- Removing invalid data for number of floors
  - .5 floors -> NULL floors
- Removing invalid leading characters from owner names
  - ",.NYC DEPARTMENT OF EDUCATION" -> "NYC DEPARTMENT OF EDUCATION"

```
UPDATE pluto a
SET ownername = trim(regexp_replace(a.ownername, '^([.,;><\-_!?`%]+)(.*)', '\2')),
    dcpedited= 't'
WHERE a.ownername ~ '^([.,;><\-_!?`%]+).*';
```

- Improving ability to test new builds in GitHub Actions

Data Pipeline: Quality Assurance & Quality Control



Quality Assurance & Quality Control (QAQC)

What does QAQC look like?

- Need for automatic processes that can detect anomalies
- Ability to track changes over time
- Protects against introducing mistakes
- Produce app-specific QAQC reports for:
  - Summary Statistics
  - Verion-to-version Comparisons
  - Outlier identification
  - Geospatial consistency
- Worked to build and enhance these reports
- Visualized reports on QAQC web application

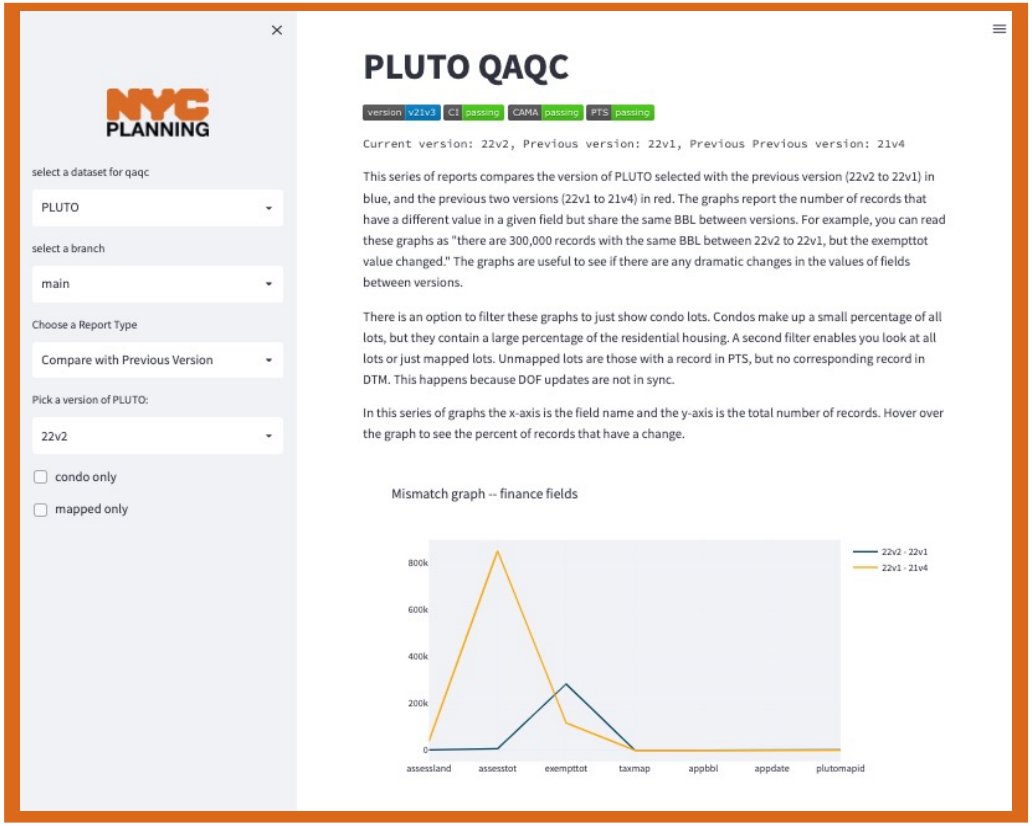
Pluto QAQC Report: Residential Area per Unit

bbl	resarea	unitsres	res_unit_ratio
3008927502	11036	64	172.4375
3010057502	37880	131	289.1603053435
3010347501	12951	81	159.8888888889
3011297502	64851	297	218.3535353535
3011560070	49500	889	55.6805399325
3011747502	6950	64	108.59375
3014520066	6913	77	89.7792207792
3014520070	6913	77	89.7792207792
3014520071	6913	77	89.7792207792
3014520072	6913	77	89.7792207792
3014520073	6913	77	89.7792207792
3014520074	6913	77	89.7792207792
3014520076	6915	77	89.8051948052

QAQC Application

The QAQC Web Application:

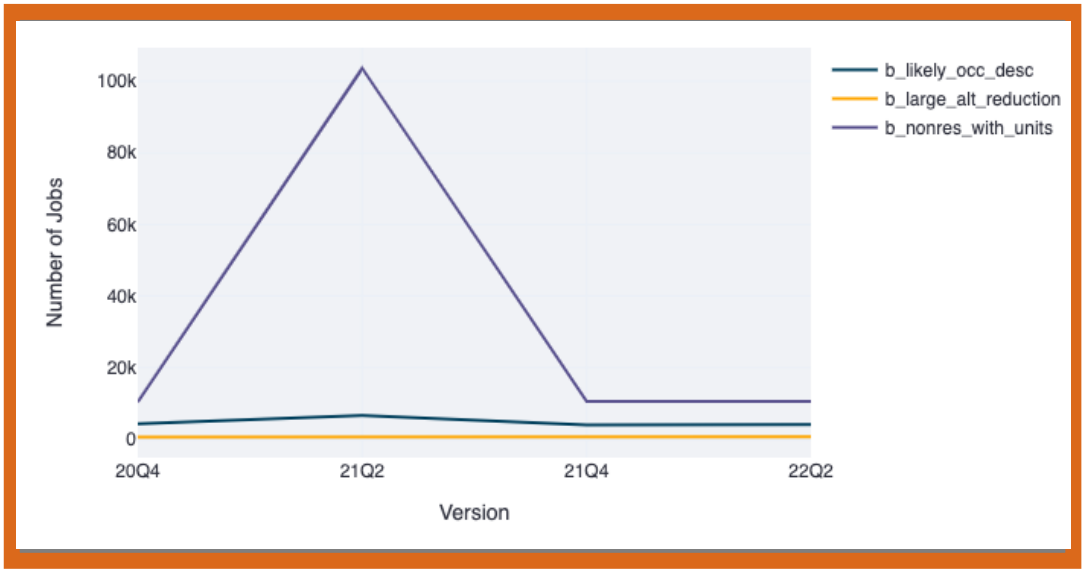
- Built using Python and Streamlit
- Ingests QAQC Reports
- Slices data on the fly
- Data Visualizations for:
  - PLUTO
  - Facilities Database
  - Housing Database
  - Capital Projects Database
  - City Owned and Leased Properties Database

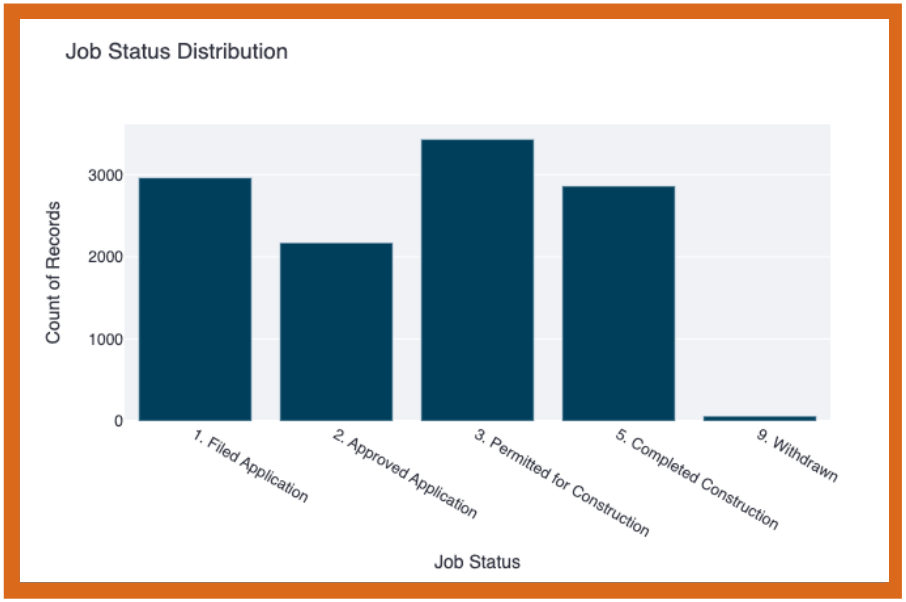


QAQC - Housing Database

Reports we Built:

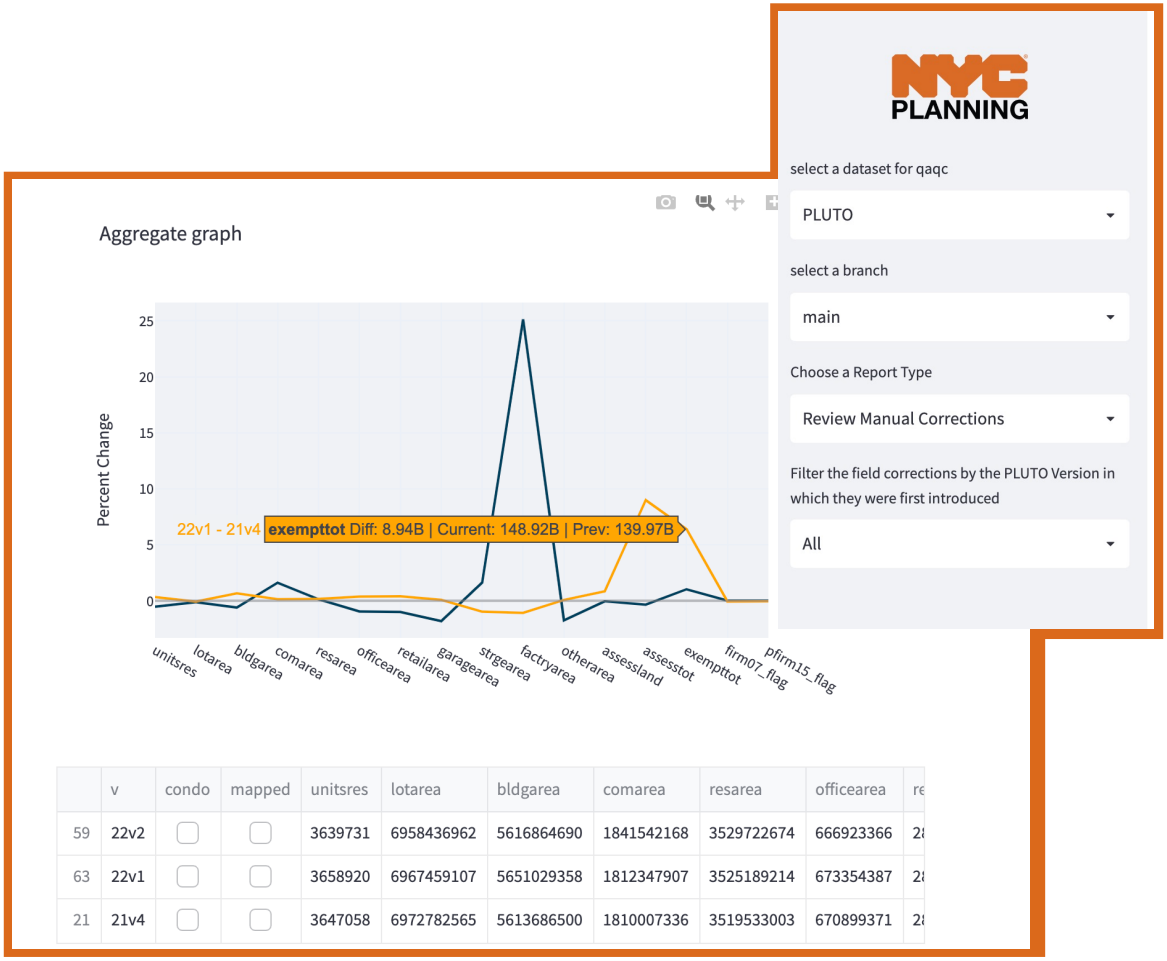
- Summary Statistics of Job Types, Job Statuses, Completed Quarter
- Version-to-Version Comparisons of the number of jobs failing a series of quality checks
- Providing tables of jobs for each failed quality check
- Allows for quick eyeballing of key fields, and spotting potential mistakes in the data





QAQC - PLUTO

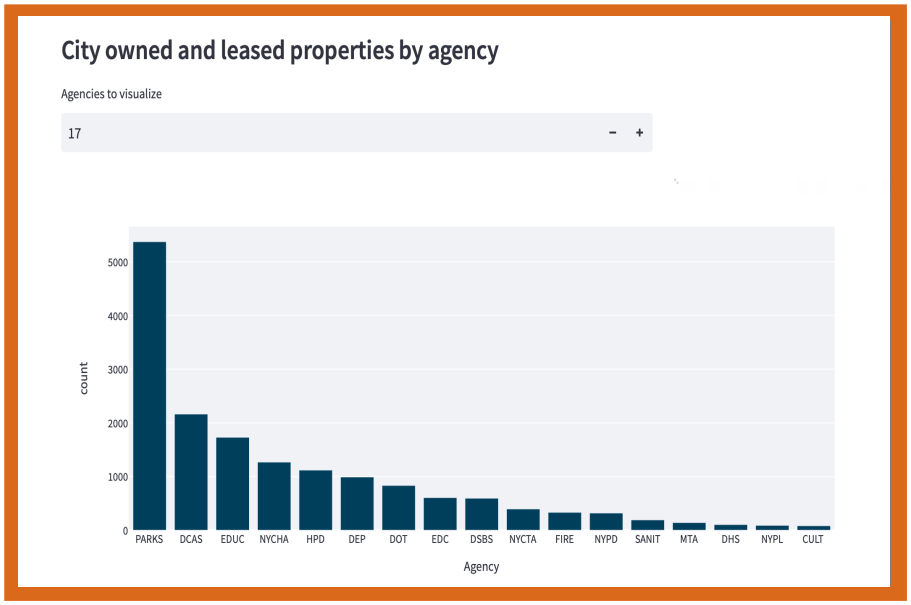
- **Enhancements that we've made:**
  - Improve the functionality and user interface
  - Improve the existing graphs by adding clear tooltips, descriptions, axis, etc.
  - Add manual Correction Report for programmatic changes and research and user reported changes
  - Outlier Analysis



	v	condo	mapped	unitsres	lotarea	bldgarea	comarea	resarea	officearea	re
59	22v2	☐	☐	3639731	6958436962	5616864690	1841542168	3529722674	666923366	2
63	22v1	☐	☐	3658920	6967459107	5651029358	1812347907	3525189214	673354387	2
21	21v4	☐	☐	3647058	6972782565	5613686500	1810007336	3519533003	670899371	2

QAQC - City Owned and Leased Properties

- **Reports We Built:**
  - **Display Useful Graphs** (number of records by agency/ use type)
  - **Version-to-version comparison** of changes in the number of records per use type
  - **Outlier Report:** display the mismatch between IPIS Community and PLUTO
  - **Geospatial Check**: check whether all properties are within NYC borough boundaries and inconsistency in geographies
  - **Manual Corrections Check:** display graphs and data frames of Manual Corrections Applied and Not Applied by field



Final Thoughts - Jingyi

- I'm thrilled to see how data products that I have used before as an individual researcher are being taken care of by DCP.
- By working on different data products and switching from one to another, I managed to learn tons of new skills and knowledge.
- I really love how the work environment helps us build collaborations and teamwork, and be adaptable, resilient, and productive.

Final Thoughts - Andrew

- The Data Engineering team wears a ton of different hats.
- Despite the huge range of data products we are responsible for, the team often provides attention to ensuring the accuracy of the smallest details of analysis.
- The development process is more organic and bottom-up than I would have expected throughout DCP.

Thank You

Thank you to:

- Amanda, Lynn, and everyone at DCP for creating this opportunity and supporting us
- Sasha, Te and Max for your patience and feedback and the opportunity to learn from your work
- Coding it Forward for creating this internship program