Status Report: City of New York Air Quality Dataset Notes

So far, I have downloaded the data as a CSV file from Data.gov and inspected it using OpenRefine. Below are explanations and tables outlining the data schema and issues I found. The next order of business would be for either me or Tianqi to thoroughly review the EPA dataset in a similar manner to determine the best way to integrate the two sets. Analysis and visualization will follow.

General Findings

- Number of rows: 18,862
- Each Unique ID is numeric and, in fact, unique
- Each Indicator ID is numeric
- Each Geo Join ID is numeric
- Each Data Value is numeric
- The Time Period column is vague, but Tianqi provided a solution
- Every cell in the Message column is blank, so it can be removed

Indicator IDs and Their Meanings

This table records the Names, Measures, and units (Measure Info column) represented by each Indicator ID. The highlighted cells represent issues in the data.

- The entry “Âµg/m3” arises from an encoding error with the mu character, and can be changed to “mcg/m3”
- The highlighted “per 100,000” measure does not specify whether it refers to adults or children, as the other “per 100,000” measures do, so it may need to be assumed that both populations are represented
- The final two highlights show a grammatical mistake that can be corrected to “department”

Indicator ID	Name	Measure	Measure Info
365	Fine particles (PM 2.5)	Mean	mcg/m3
375	Nitrogen dioxide (NO2)	Mean	ppb
386	Ozone (O3)	Mean	ppb
639	Deaths due to PM2.5	Estimated annual rate (age 30+)	per 100,000 adults
640	Boiler Emissions- Total SO2 Emissions	Number per km2	number

641	Boiler Emissions-Total PM2.5 Emissions	Number per km2	number
642	Boiler Emissions-Total NOx Emissions	Number per km2	number
643	Annual vehicle miles traveled	Million miles	per square mile
644	Annual vehicle miles traveled (cars)	Million miles	per square mile
645	Annual vehicle miles traveled (trucks)	Million miles	per square mile
646	Outdoor Air Toxics - Benzene	Annual average concentration	Âµg/m3
647	Outdoor Air Toxics - Formaldehyde	Annual average concentration	Âµg/m3
648	Asthma emergency department visits due to PM2.5	Estimated annual rate (under age 18)	per 100,000 children
650	Respiratory hospitalizations due to PM2.5 (age 20+)	Estimated annual rate	per 100,000 adults
651	Cardiovascular hospitalizations due to PM2.5 (age 40+)	Estimated annual rate	per 100,000 adults
652	Cardiac and respiratory deaths due to Ozone	Estimated annual rate	per 100,000
653	Asthma emergency departments visits due to Ozone	Estimated annual rate (under age 18)	per 100,000 children
655	Asthma hospitalizations due to Ozone	Estimated annual rate (under age 18)	per 100,000 children
657	Asthma emergency department visits due to PM2.5	Estimated annual rate (age 18+)	per 100,000 adults

659	Asthma emergency departments visits due to Ozone	Estimated annual rate (age 18+)	per 100,000 adults
661	Asthma hospitalizations due to Ozone	Estimated annual rate (age 18+)	per 100,000 adults

Geo Type Names

The Geo Type Name column represents the type of geographic area measured, including the whole of New York City (Citywide), its five boroughs, 59 community districts (CD), UHF34 and UHF42 districts, which are the two districting methods of the United Hospital Fund of New York. The numbers in the two UHF schemes represent how there are 34 and 42 districts in each, respectively.

Boroughs and Citywide Geo Join IDs

The Geo Join ID codes each Geo Type to a number, but these IDs are not always unique. The table below shows how the ID for “Citywide” and “Bronx” are the same. A possible solution is to change the Citywide Geo Join ID to 0, making it unique to the column. However, the purpose of the Geo Join ID is to map data to other datasets, so further inspection of our EPA set in this regard is required.

Geo Join ID	Geo Type Name	Geo Place Name
1	Borough	Bronx
1	Citywide	New York City
2	Borough	Brooklyn
3	Borough	Manhattan
4	Borough	Queens
5	Borough	Staten Island

Limitations of Some Geo Types

The table below shows which Indicators are recorded for each Geo Type. Notice how the only Geo Types that record data for every Indicator are Citywide, the boroughs, and the UHF42 districts.

Geo Type Name	Indicator IDs with Data
Borough	Each borough records all IDs

CD	365, 375, 386, 643, 644, 645, 646, 647
Citywide	All IDs
UHF34	365, 375, 386
UHF42	Each district records all IDs

UHF42 Geo Join IDs

Due to these limitations, I have only thoroughly inspected the UHF42 Geo Type for now. Still, at a cursory glance at the other Geo Types, several of the IDs from the two UHF types overlap. This is because UHF district IDs are based on location, so UHF42 district 101 is in the Bronx because the first number “1” corresponds to the Bronx’s Geo Join ID, and the “01” is derived from it being the upper-leftmost of the Bronx districts. The same applies to UHF32.

The table below links the UHF42 Geo Join IDs to their place names and highlights some errors based on the codebook and map shown under the table. I also searched Apple Maps to double-check the Geo Place Names that did not match the codebook to verify the correct name of the locations they meant to reference. Note that I have not yet ensured every district actually encompasses the locations it is named after.

- The “Pk” in “Fordham - Bronx Pk” should be changed to “Park”
- “High Bridge - Morrisania” likely refers to the neighborhood called Highbridge (with no space) in this district. Although a bridge literally called High Bridge exists in the neighborhood, the district is likely named after the neighborhood itself, and should probably be changed to reflect that. However, it would be more appropriate to conform to the UHF code rather than to more accurate semantics
- “Downtown - Heights - Slope” should be “Downtown - Heights - Park Slope”
    - Again, “Downtown Brooklyn - Crown Heights - Park Slope” would be a more accurate description of the neighborhoods, but I have yet to see a codebook with district 202 named that way
- 301 also covers Inwood, so it is supposed to be “Washington Heights - Inwood”
- Rockaways is a spelling error; it should be “Rockaway”

Geo Join ID	Geo Place Name
101	Kingsbridge - Riverdale
102	Northeast Bronx
103	Fordham - Bronx Pk
104	Pelham - Throgs Neck
105	Crotona -Tremont
106	High Bridge - Morrisania

107	Hunts Point - Mott Haven
201	Greenpoint
202	Downtown - Heights - Slope
203	Bedford Stuyvesant - Crown Heights
204	East New York
205	Sunset Park
206	Borough Park
207	East Flatbush - Flatbush
208	Canarsie - Flatlands
209	Bensonhurst - Bay Ridge
210	Coney Island - Sheepshead Bay
211	Williamsburg - Bushwick
301	Washington Heights
302	Central Harlem - Morningside Heights
303	East Harlem
304	Upper West Side
305	Upper East Side
306	Chelsea - Clinton
307	Gramercy Park - Murray Hill
308	Greenwich Village - SoHo
309	Union Square - Lower East Side
310	Lower Manhattan
401	Long Island City - Astoria
402	West Queens
403	Flushing - Clearview
404	Bayside - Little Neck
405	Ridgewood - Forest Hills

406	Fresh Meadows
407	Southwest Queens
408	Jamaica
409	Southeast Queens
410	Rockaways
501	Port Richmond
502	Stapleton - St. George
503	Willowbrook
504	South Beach - Tottenville

UHF CODES

UNITED HOSPITAL FUND CODES

UHF code	Neighborhood name	Zip codes
Bronx		
101	Kingsbridge - Riverdale	10463, 10471
102	Northeast Bronx	10466, 10469, 10470, 10475
103	Fordham - Bronx Park	10458, 10467, 10468
104	Pelham - Throgs Neck	10461, 10462, 10464, 10465, 10472, 10473
105	Crotona - Tremont	10453, 10457, 10460
106	High Bridge - Morrisania	10451, 10452, 10456
107	Hunts Point - Mott Haven	10454, 10455, 10459, 10474
Brooklyn		
201	Greenpoint	11211, 11222
202	Downtown - Heights - Park Slope	11201, 11205, 11215, 11217, 11231
203	Bedford Stuyvesant - Crown Heights	11213, 11212, 11216, 11233, 11238
204	East New York	11207, 11208
205	Sunset Park	11220, 11232
206	Borough Park	11204, 11218, 11219, 11230
207	East Flatbush - Flatbush	11203, 11210, 11225, 11226
208	Canarsie - Flatlands	11234, 11236, 11239
209	Bensonhurst - Bay Ridge	11209, 11214, 11228
210	Coney Island - Sheepshead Bay	11223, 11224, 11229, 11235
211	Williamsburg - Bushwick	11206, 11221, 11237
Manhattan		
301	Washington Heights - Inwood	10031, 10032, 10033, 10034, 10040
302	Central Harlem - Morningside Heights	10026, 10027, 10030, 10037, 10039
303	East Harlem	10029, 10035
304	Upper West Side	10023, 10024, 10025
305	Upper East Side	10021, 10028, 10044, 10128
306	Chelsea - Clinton	10001, 10011, 10018, 10019, 10020, 10036
307	Gramercy Park - Murray Hill	10010, 10016, 10017, 10022
308	Greenwich Village - SoHo	10012, 10013, 10014
309	Union Square - Lower East Side	10002, 10003, 10009
310	Lower Manhattan	10004, 10005, 10006, 10007, 10038, 10280
Queens		
401	Long Island City - Astoria	11101, 11102, 11103, 11104, 11105, 11106
402	West Queens	11368, 11369, 11370, 11372, 11373, 11377, 11378
403	Flushing - Clearview	11354, 11355, 11356, 11357, 11358, 11359, 11360
404	Bayside - Little Neck	11361, 11362, 11363, 11364
405	Ridgewood - Forest Hills	11374, 11375, 11379, 11385
406	Fresh Meadows	11365, 11366, 11367
407	Southwest Queens	11414, 11415, 11416, 11417, 11418, 11419, 11420, 11421
408	Jamaica	11412, 11423, 11432, 11433, 11434, 11435, 11436
409	Southeast Queens	11004, 11005, 11411, 11413, 11422, 11426, 11427, 11428, 11429
410	Rockaway	11691, 11692, 11693, 11694, 11695, 11697
Staten Island		
501	Port Richmond	10302, 10303, 10310
502	Stapleton - St. George	10301, 10304, 10305
503	Willowbrook	10314
504	South Beach - Tottenville	10306, 10307, 10308, 10309, 10312

UHF MAP

UNITED HOSPITAL FUND MAP by BOROUGH