NYC Opportunity EDS – Take-home Assessment for Data Curator/Scientist

Paolo Rivas
Email: privas.legua@gmail.com
#: 347 968-5109

Task 1: Director Stats
The Executive Director of the Mayor's Office to Protect Tenants is interested in providing stats on Housing Court activity in New York City for a presentation for a Deputy Mayor. The Housing Data Coalition makes housing-related datasets easily available to the public. See their GitHub page here (instructions to access the dataset can be found here). Using the OCA Housing Court Records dataset, answer the following summary statistics.

Q1: How have many total cases for residential housing there been in each NYC borough?

A1: If we consider only the County Civil Courts of each borough and the Harlem Community Justice Center as part of the borough of Manhattan, the total number of cases in NYC are the following: Bronx: 375,667, Brooklyn: 290,108, Manhattan: 218,419, Queens: 161,632 and Staten Island: 23,123.

Q2: Which month of the year sees the highest number of housing court cases?

A2: If we assume the filed date as a marker for the beginning of a court case, January is the month with the most housing court cases with 107,133 initiated in that month, followed by February with 102,755 and March with 99,562.

Q3: What are the top 10 NYC ZIP codes with residential court cases before March 2020?

A3: The following are the top 10 NY zip codes with more residential cases: 1) 10456, 2) 10453, 3) 10467, 4) 10457, 5) 10452, 6) 11212, 7) 10458, 8) 10468, 9) 11226, 10) 11207.

Q4: Are there any neighborhoods (based on ZIP codes) that show relatively more cases after March 2020 compared to pre-pandemic?

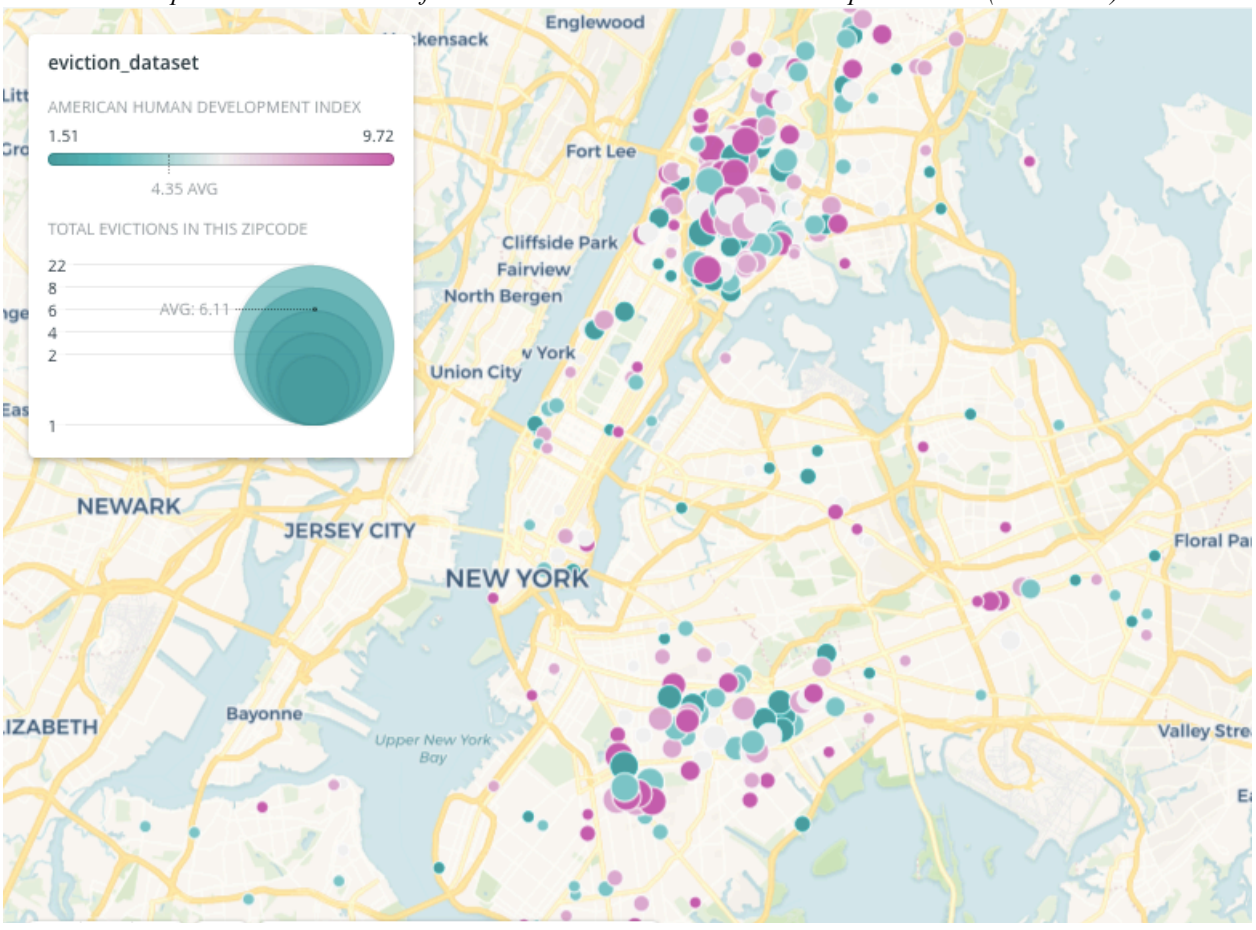
A4: Excluding the exception of postal code 1191-5702 (belonging to Far Rockaway), which based on monthly mean had one more case than the pre pandemic average of observations, all cases reported in the database either dropped or stayed relatively the same because of the pandemic. The majority of them, once again, on average dropped significantly because of the virus.

Task 2: Tenant Outreach
Develop a set of recommendations for the Mayor's Office to Protect Tenants to improve outreach efforts to tenants who might be at risk of appearing in court or eviction. You can start with the OCA dataset you use for Task 1, eviction data from NYC Open Data, NYC PLUTO data, census data, or other data sets. You are not required to use all suggested data sets, but also, we encourage you to use any methodologies or additional datasets that would help you answer this question.

Memorandum

To: Mayor's Office for Economic Opportunity
Subject: Evictions and Human Development Index are not statistically associated for NYC zip code.
In despite of what it seems to be an immediate visual association (see graph 1.1), evictions are not statistically correlated to Life expectancy, Education, and Earnings; or at least are trivial as part of the components of the American Human Development Index. One possible explanation on why these two variables tested insignificant might be the lack of a larger sample to test more rigorously this hypothesis or the dispersity that can be found between different units of measurements and zip codes. This finding is crucial to demystify some of the mental association most New Yorkers have regarding the abuse of authority and structural socio-economical dynamics of the city that might not be reflected in the data.

Graph 1.1: Total Number of Evictions vs. American Human Development Index (2017-2019)



Task 2: Tenant Outreach (technical analysis)

Disclaimer & data sources: Due to the lack of time to perform a more complex task, for this part of this exercise I am going to do a simple but informative correlational test followed by ab effective visualization exercise. The goal will be to join the OCA Evictions datasets and the 2020 American Human development index (AHDI) from OurHome.nyc. AHDI is a composite index (0-10) of USALEEP Life expectancy, Education, and Earnings. For more information see http://measureofamerica.org/human-development/. From the original dataset I have already selected and downloaded only zip codes from NYC and ADHDI values. If the original dataset is needed, it can be fully downloaded over here:
https://ssrc.formstack.com/forms/index.php

Problem statement: According to data, evictions are more likely to be executed in less affluent areas where systemic and structural social inequality conditions are predominant.

Hypothesis: The occurrence of evictions in specific zip code is negatively correlated with that particular area Human development index. Therefore, the higher a particular zip code is in the AHDI, the less likely is to be evicted.

Statistical Technique: Coefficient of relation: Person's R

Outcome of test: no Association between variables R: -0.017, P-value: 0.75

Hypothesis of negative testing: Problems with the representativity of the sample used might have severely limited the outcome/ Dispersity might have also been higher than expected between values in zip code neighbors

Caveat/Next Steps: Try a different grouping technique beyond zip codes and test again. One alternative is to create macro-zip codes areas with k-means and look for correlations with separate socio-demographic variables (not exclusively AHDI) and see if any of those variables are highly correlated with evictions.

Task 3: Data Documentation Exercise
These are open-ended questions and intended to help understand your approach to data documentation

Q1: What do you think are ideal sources for data documentation?

Ideally, Data dictionaries, standardized codebooks (in harmony with ICPSR guidelines) or any well documented "readme" style metadata (Cornell's Readme template is well known for this). At least, a concise document that has a title of dataset, creation date, Investigator names, keywords, purpose of study, research questions/hypotheses (in case it's an academical setting). Regarding the specific of the documentation, I would ideally expect to find also file formats, content, size, relationship among files; as well, Data source, provenance, copyright permissions. Data identifiers (DOI, URI) and Information on confidentiality, access & use conditions are also crucial.

Q2: How can data documentation affect your analysis of the data?

If I do not have access to good documentation, I can make several mistakes not only in my analysis (for instances, misreading the content of multiple variables) but also in the uses, rights and authorizations for my data practices. In addition, without good documentation valuable time is wasted in identifying basic elements of any first assessment like formats, keywords, data cycles code, syntax or software uses.

Q3. Why is it important to have thorough data documentation?

Quantification is representation. It is fundamental to keep an eye on what are we tracking and storing; for how long and how we can access to this information; where are those repositories. It is particularly important when we are talking about public information: data and algorithm accountability are practices that safeguard public trust and data governance policies are crucial for any form of e-government and e-democracy.