Uncovering the Realities of NYC: A Data-Driven Analysis of Complaints, Evictions, and Storefront Vacancies in the Big Apple

Author: Nikola Baci

## Introduction

This paper provides an analysis of three data sets from the NYC Open Data program to explore the relationships between building complaints, evictions, and storefront vacancy rates from 2019-2021 in the New York City (NYC) area. This analysis can be used by city officials and real estate professionals to make informed decisions about policies and investments to address building complaints, evictions, and storefront vacancies in the NYC area.

In the report we aggregate the data based on zip code and year and performs linear regression analysis to understand the trends and patterns. The paper begins by introducing the data sets and the specific columns that will be used. It also outlines the assumptions made and the limitations of the data. The paper then presents the findings from the analysis, including histograms, heatmaps, and tables, to provide insights into the patterns and trends of building complaints, evictions, and storefront vacancies in the NYC area. Finally we build a linear regression model to pin down the relationship between number of complaints with number of evictions and vacancies.

## Data

In this analysis I will be using three data sets provided by the NYC Open Data program.

The first data set is provided by Department of Building where we can find all the complaints entered in the system. These complaints where filed with 311 and there are 3.3 million complaints as of March 13, 2023. Additionally, the data set has 15 columns, but we will be focusing only on the column that holds the year and zip code data. You can view or download the data set [here](https://data.cityofnewyork.us/Housing-Development/DOB-Complaints-Received/eabe-havv).

The second data set is provided by the New York City Marshals where we find information on pending, scheduled, and executed evictions in the five boroughs from 2017 to the present. As of now, the data set has 72K rows and 20 columns, but as with the DOB data set we will only be using the column that hold the year and zip code data. You can view of download the data [here](https://data.cityofnewyork.us/City-Government/Evictions/6z8x-wfk4).

The third data set provided by an online portal which allows owners to provide information about ground and second floor storefronts on their properties complying with Local Law 157. To date the data set has 214K rows and 24 columns. You can view of download the data [here](https://data.cityofnewyork.us/City-Government/Storefronts-Reported-Vacant-or-Not/92iy-9c3n). This data set however only track records as of December 31 of each year. Meaning owners report if the store was vacant or not as f the end of the year. Keep in mind a store could have been vacant for only a few days before reporting or a long time vacant store could get occupied a few days before reporting.

All three data sets will be aggregated to show the total sum of complaints, evictions, and vacant vs non vacant storefronts respectively. The aggregated information will be based on the year and zip code. The year will be limited between 2019 and 2021 (3 years) while the zip codes will be limited on the NYC (5 boroughs) area.

The aggregated data will be put together in on data frame and linear regression analysis will be performed. To enable uniform and reliable merging of the tables I will be using a zip code data set that I have obtained for the NY state.

The data sets are in a good condition with very few missing values and proper documentation to help understand the columns. It is worth noting that the aggregated data is not exhaustive and so when we merge them we notice missing values. This happens when in year y and zip code x there have been complaints filed with the DOB but there were no evictions (for the same y-x combination).

Here we also make the assumption that is a (year-zip) combination does not appear on the data set, then no complaints or evictions then the count is zero.

## Anlaysis

### DOB Complaint Analysis

Below we create an histogram chart to show the distribution of complaints received from 2019-2021.

As you can see, the levels of complaints for each month range from 10,000 to 30,000+ per month in all of NYC. It is quite noticeable by the humps of the chart that the number of complaints tends to go up during the summer months.

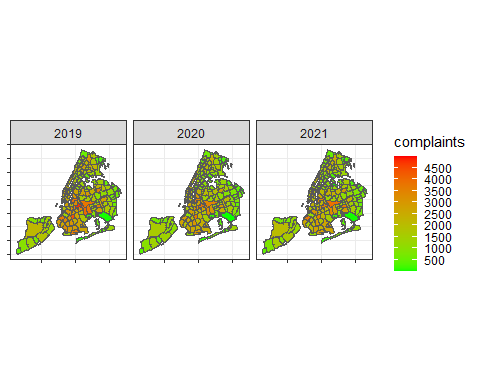
Although it might seem like 2020 was the year with the far fewer complaints, crunching the numbers shows that in fact there is not big difference as the table below suggests.

| Year | Count |
| --- | --- |
| 2019 | 309,490 |
| 2020 | 257,502 |
| 2021 | 257,046 |

Thur further analysis, we find that over all three years there are three neighborhood that make the cut for the top 5 most complaints and consequently the top 5 overall:

| zip | borough | neighborhood |
| --- | --- | --- |
| 11221 | Brooklyn | Bushwick |
| 11226 | Brooklyn | Flatbush |
| 11385 | Queens | Ridgewood |

Below we take an eagle’s eye view of the whole NYC area using a heatmap where each zip code is colored depending on the total number of complaints for the given year.



From the maps we notice that Brooklyn and west-Queens stand out compared to other boroughs with a medium to high number of complaints. The reasons why are unknown at this time and further investigation is necessary.

### Evictions Analysis

Let’s break eviction data set down and understand its content. In the NYC area we are dealing with 27,455 evictions across all 5 boroughs. The tables below paint a quick picture about those evictions.

* Table 1: the court ruled in ejectment of 18 or 0.08% of evictions cases.
* Table 2: 90% of eviction cases are residential

Further analysis shows that out of the 17 ejectments only 1 was commercial, the rest was residential.

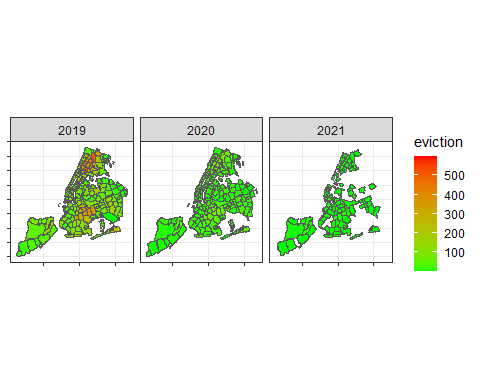
Table 1 Table 2

|Action |Count |Type |Count | |————|——| |————|——| |Ejectment | 18| |Commercial | 2145| |No Ejectment|22,387| |Residential | 20260|

| Type | Count |
| --- | --- |
| Commercial | 2145 |
| Residential | 20260 |

The map below, similar to the complaints map, is showing a geographical distribution of the evictions separated by year. It appears that Bronx and Brooklyn account for the majority of the evictions in 2019, however this number goes down in ’20-’21. This could the result of the Covid-19 pandemic at time when NYC took [strict measures to protect tenants](https://hcr.ny.gov/covid-19-eviction-protections-tenants).

The zip codes that have no evictions are removed from the map.



### Storefront Analysis

The table below shows a quick summery of the total count of reports, the respective numbers for vacant and not vacant storefronts as well as a calculated ration between those two numbers.

| Year | Count | Vacant | Non Vacant | Ratio Vacant:Non Vacant |
| --- | --- | --- | --- | --- |
| 2019 | 75250 | 6940 | 68310 | 1:10 |
| 2020 | 74469 | 8428 | 66041 | 1:8 |
| 2021 | 63409 | 6598 | 56811 | 1:9 |

We notice that in 2021 we had about 10,000 reports fewer than the previous years. And as some of you might expect, there were more stores empty during the pandemic bringing the ration down to 1:8 (for every 8 stores non vacant, you find a vacant store).

Additionally,

| Business | Count |
| --- | --- |
| RETAIL | 60881 |
| OTHER | 33744 |
| FOOD SERVICES | 41226 |
| MISCELLANEOUS OTHER SERVICE | 15174 |
| HEALTH CARE or SOCIAL ASSISTANCE | 8427 |

### Linear Regression Model

In this analysis we will investigate if there is a relationship between complaints received in a zip code with the percent of vacant stores and eviction.

To do this, we use Ordinary Least Square (OLS) linear regression model that will fit the line in such a way as to minimize the total difference of each point from the line hence the name OLS.

Because of dimensionality we cannot show a graphical illustration of the model. The model shows that for each additional eviction a zip code area in NYC goes up by 5 complaints and for each additional vacant store the zip code has there are 11 more complaints in the neighborhood.

Both those findings are statistically significant since the p-value or the likelihood that we would see such a result is negligible.

It is important to note that just because there is a concurrence we have limited information to imply causation. This means vacant storefronts does not make people file more complaints with the DOB, rather it happens that when the number of vacant storefronts is high so is the number of complaints.

##   
## Call:  
## lm(formula = df$complaints ~ df$eviction + df$vacant, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1923.1 -659.8 -170.7 501.6 2794.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 661.7819 40.8124 16.21 <2e-16 \*\*\*  
## df$eviction 5.1067 0.4338 11.77 <2e-16 \*\*\*  
## df$vacant 11.7105 0.7644 15.32 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 803.8 on 681 degrees of freedom  
## Multiple R-squared: 0.3824, Adjusted R-squared: 0.3806   
## F-statistic: 210.8 on 2 and 681 DF, p-value: < 2.2e-16

### Conclusion

In conclusion, this analysis of building complaints, evictions, and storefront vacancies in the NYC area has provided valuable insights into the patterns and trends from 2019-2021. The analysis shows that building complaints tend to increase during the summer months, and that certain neighborhoods in Brooklyn and Queens have higher numbers of complaints. The analysis also shows that most of the eviction cases are residential, and the majority of the evictions occur in the Bronx and Brooklyn. Finally, the analysis reveals that storefront vacancies are concentrated in certain neighborhoods, and that the vacancy rate has remained more or less the same.