# Mining the Relationship between Crime, Poverty, and Noise Complaints in NYC

**Franco La Bruna**
**Yuchen Zheng**

CSC440 – Final Project

*Fall 2016*

## 1. Motivation and background

Nowadays, noise has been one of the major pollution sources, which, affects both people's health and behavior. Unwanted pollution, according to J.M. Field in his paper "*Effect of personal and situational variables upon noise annoyance in residential areas",* can damage psychological health and cause hypertension, hearing loss, sleep disturbances, and other harmful effects. In addition, crime and poverty are two other significant social issues with large negative externalities that constitute threats to safety of citizens and stability of society. Despite a large number of existing studies on these issues separately, it is frustrating to find no such topics on making connections and conclusions between crime, poverty and noise have been examined.

Data mining techniques have been rising as tools to analyze, in this case, data resulting from human beings' activity, stored in large databases. They can be applied in order to detect significant patterns and rules underlying human behavior. Inspired by Professor Jiebo Luo from the University of Rochester, we would like to study this undefined topic and see whether we can reveal the potential relationship between enormous noise complaints data and crime rates & poverty in New York City area with implementations of various data mining techniques. Any findings or conclusions will also be beneficial to future reference and local officials in terms of prevention of crime and decision making.

## 2. Related Work

As discussed in the first part, it is one of our incentives to do this project since previous literature has seemingly done very limited or even no research on our topic. One paper that is most relevant to what we intend to do is "Diagnosing New York City's Noises with Ubiquitous Data" by Zheng, Y. et. al, in which they employed noise complaint data to build a model that recovers the noise situation throughout NYC to inform people and officials' decision making. Though what they have accomplished is disparate from what we intend to do, one sentence in the introduction of that paper is genuinely enlightening, "As each complaint about noises is associated with a location, a time stamp, and a fine-grained noise category, … , the data is actually a result of “human as a sensor” and “crowd sensing”, containing rich human intelligence that can help diagnose urban noises.” For the project, our group is also going to use the data as a result of “human as a sensor” and “crowd sensing” to figure out the relationship between crime rates, poverty and noise in New York City.

### 3. Data Acquisition

Noise complaint data can be readily acquired from New York City Open Data. The data provided there contains millions of noise complaints to the 311 non-emergency service hotline from 2010 until the present day, and each noise complaint also lists the date of the complaint, the source of the complaint, a description of the offending noise, and location information by various metrics such as borough, street name, ZIP code, and GPS coordinates. Crime data is more nebulous, and will likely have to be compiled from multiple sources. After looking through all available datasets on the Internet, we decided that we would be using crime data from NYPD’s website since it must be the most reliable and accurate data source as it is the law enforcement agency in NYC. For poverty data, we were able to find an online database called ZIPAtlas, which has the percentage of population below poverty line in New York State by zip code.

### 4. Data Preprocessing

Data preprocessing is always the most fundamental task for any data mining related work and usually takes the most time. In our project, we first converted everything into data frames using Pandas in Python, which would enable us to manipulate and observe data very easily. Then we did data preprocessing involving data reduction, data cleaning, data integration, data transformation and data discretization upon having collected data of noise complaints, crime rates and poverty in New York City.

**(1).** Data Reduction:

Nearly all the raw datasets we pulled needed to be reduced. The noise complaint data from NYC Open Data had more than 50 columns that represent different attributes of noise complaints. For our research purposes we eliminated most of them and kept only several (less than eight) in the end to avoid redundancy and colossal dimensionality. Also, we discarded all noise complaints from 2016 due to crime data for that year being unavailable on the NYPD's website as of December, 2016. Within the noise complaint data, we noticed some redundancy as well. For instance, in the column "Borough", we found that there was one borough called "unspecified" in the dataset. Since it only took up a very small portion of the entire dataset, we got rid of all rows with column "Borough" to be "unspecified" for convenience. Other examples are abound in other columns such as "Precinct", "Zip Code" etc.

For crime data, we modified titles, eliminated all irrelevant ones that were not within 2010-2015 and deleted word descriptions from the downloaded data files. Compared with the other two, the poverty dataset had no unnecessary attributes, although we still needed to delete all the rows involving non-NYC zip codes.

(2). Data Cleaning:

A number of missing values were observed in the remaining columns after we reduced the noise complaint data. We employed a few techniques to tackle this problem. For example, we used a built-in function Pandas.DataFrame.fillna to fill all the blank cells in data frames with some values, which, we choose it to be "NaN" in our case.

No data from noise, crime and poverty should be deemed as noise since each represents some incident that actually happened and second, even if there was any abnormality, it will get screened out by pattern mining.

Inconsistency was largely found in NYPD's information on different precincts and their corresponding zip codes, which was crucial for merging noise and crime data. There were more than 50 missing zip codes and for some zip codes, they could be seen in more than one precinct. Since no such list for precincts and zip codes in NYC was found, we had to look up and append zip code with its corresponding precinct manually to a list.

(3). Data Integration:

For this part, we used several groupby and join functions (left join, right join, inner join, left out join, etc.) in Pandas that are analogous to SQL joins to integrate datasets on some shared columns among all. For each noise complaint filed, all the crime and poverty data for the precinct or the zip code that the location of the complaint belongs to, were appended for the sake of pattern mining. Melt function from Pandas was flexibly used to convert the data frame between long and wide format when integrating different data sources for the purpose of reducing a considerable number of “NaN” entries so that it would save time cost and memory in our computers with a lower dimensionality.

(4). Data transformation and Discretization:

As a result of the huge size of data, we have millions of disparate numeric values for statistics from crime and poverty data that will very highly result in getting no results from pattern mining in the next step. Hence, data transformation and discretization become undoubtedly a must in our project. After taking pros and cons of many available metrics into our consideration, we came to the decision that we would use interquartile range (IQR) together with five-number summary. Two major advantages for using IQR and five-number summary are: 1. Outlier Identification. It is easy to obtain an initial estimation of outlier data by looking at all the points that fall out of the range of 1.5*IQR, which is, in our project, particularly useful since we are interested in any patterns that involve any extremely high or low level of noise, crime or poverty so that observations can be made and further research be done. 2. Skewness. By comparing the median to the quartile values can we know whether and how much the column data is skewed, allowing us to find, for example, this area has more skewed crime data in which most of the data points are greater than median while some particular points stay way below median within IQR, compared with another area. The way we applied IQR method to the dataset was to write a Python function called “binning” that discretizes data in each column into six bins, which are, “low outliers”, “lower than IQR”, “within lower IQR”, “within higher IQR”, “higher than IQR”, “high outliers”.

## 5. Data Correlation

| | Violation Offenses | Misdemeanors | Major Felonies | Non-major Felonies | Poverty Rate |
|---|---|---|---|---|---|
| Brooklyn | 0.11 | -0.02 | -0.05 | -0.15 | 0.09 |
| Queens | 0.2 | 0.15 | -0.11 | -0.2 | -0.29 |
| Manhattan | -0.36 | -0.38 | -0.36 | -0.28 | -0.16 |
| Staten Island | 0.25 | 0.68 | 0.39 | 0.69 | 0.44 |
| Bronx | -0.89 | -0.91 | -0.19 | 0.45 | 0.2 |

Table 1：Correlations between boroughs and crime & poverty rate

With the goal of getting a better idea of the relationship between noise complaints, crime and poverty, our group did calculation on find correlations, which for sure would provide us with some insight on what to expect from pattern mining, between the number of noise complaints in different boroughs in NYC and four kinds of crime as well as poverty rate. As can be observed from the table, all four categories of crime and poverty have very small correlations with noise complaints which are negligible in

Brooklyn and Queens. That indicates that despite the fact that noise, crime and poverty are commonly distributed throughout these two boroughs, the tendency of them promoting or demoting each other seems to be stochastic.

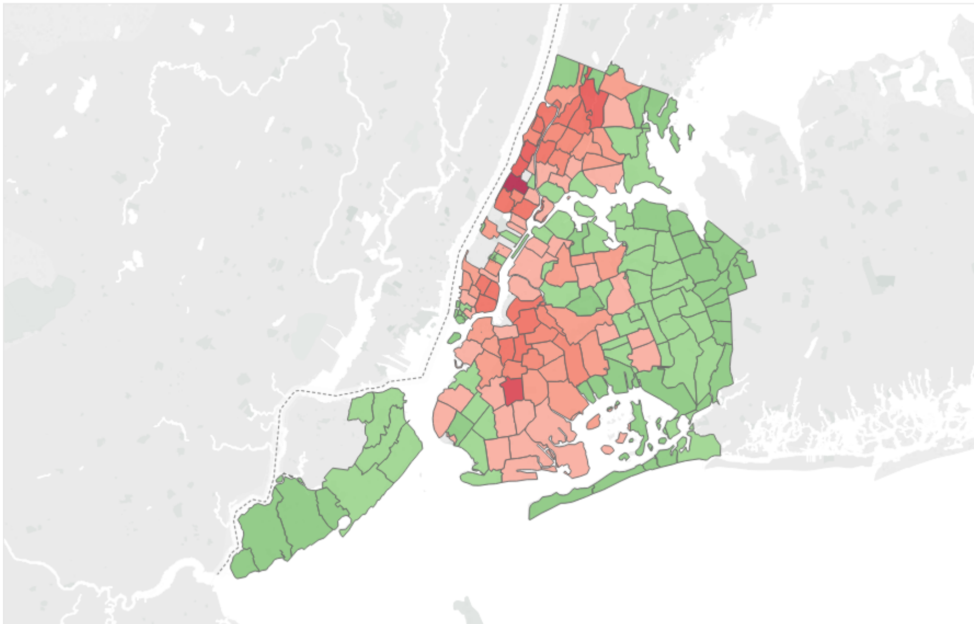
For Manhattan, all correlation coefficients are negative and larger in terms of values compared with those for Brooklyn and Queens, which implies some but limited linearity especially for violation offenses, misdemeanors and major felonies, whose correlations with noise are close to -0.4.

For Staten Island, notice that noise complaints has high positive correlations with misdemeanors and non-major felonies(0.68 and 0.69 respectively) while the other correlations being low yet, similar to Manhattan, still meaningful to some extent. The high positive correlations for misdemeanor and non-major felonies reveal a high probability that noise occur with misdemeanors and non-major felonies separately.

Lastly, for Bronx, what stands out is that noise complaints has very high negative correlations with violation offenses & misdemeanors (-0.89 and -0.91), which shows a strong repellence between noise and these two kinds of crime, although the causation behind this remains unknown.

## 6. Maps

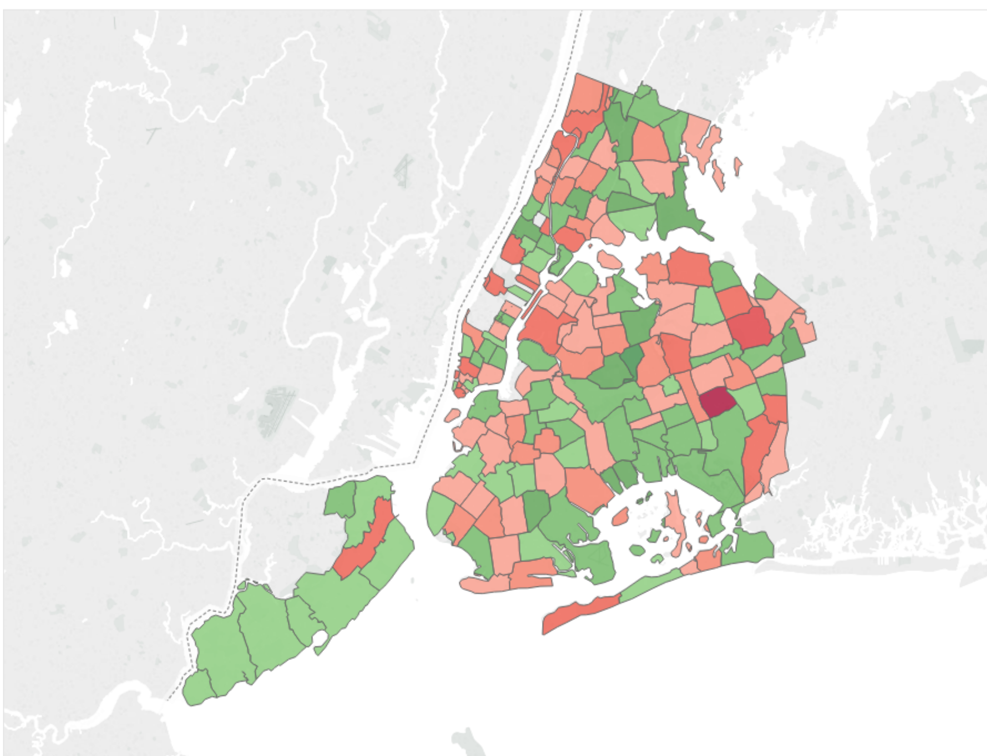
Map of Zip Codes and Number of Noise Complaints



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Noise Complaint. Details are shown for Zip Code.

Figure 1: Frequency map of noise complaints based on zip codes in NYC

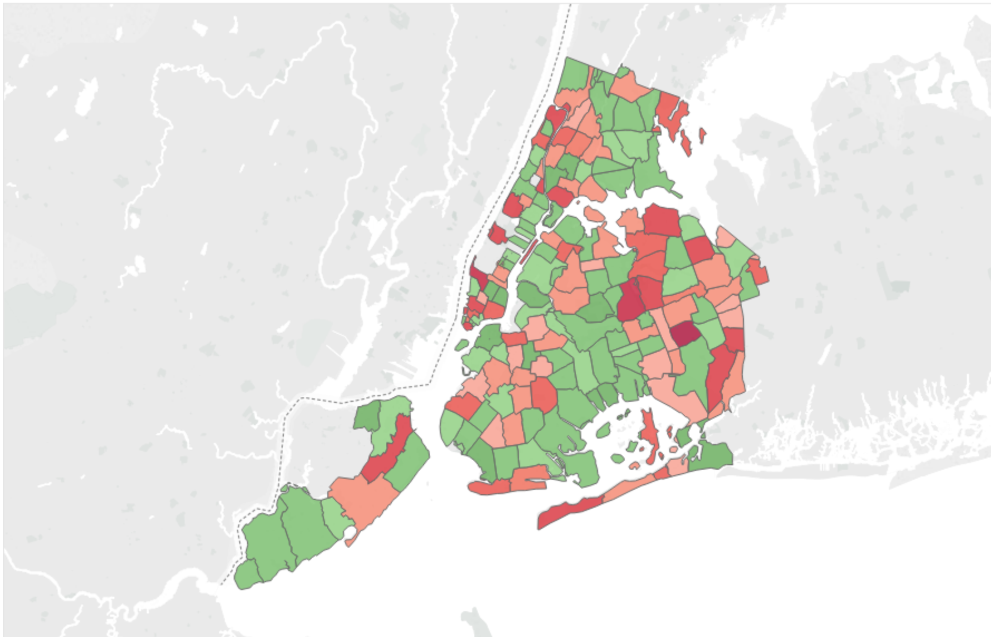
Map of Zip Codes and Major Crimes



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Major Crime. Details are shown for Zip Code.

Figure 2: Frequency map of major crimes based on zip codes in NYC

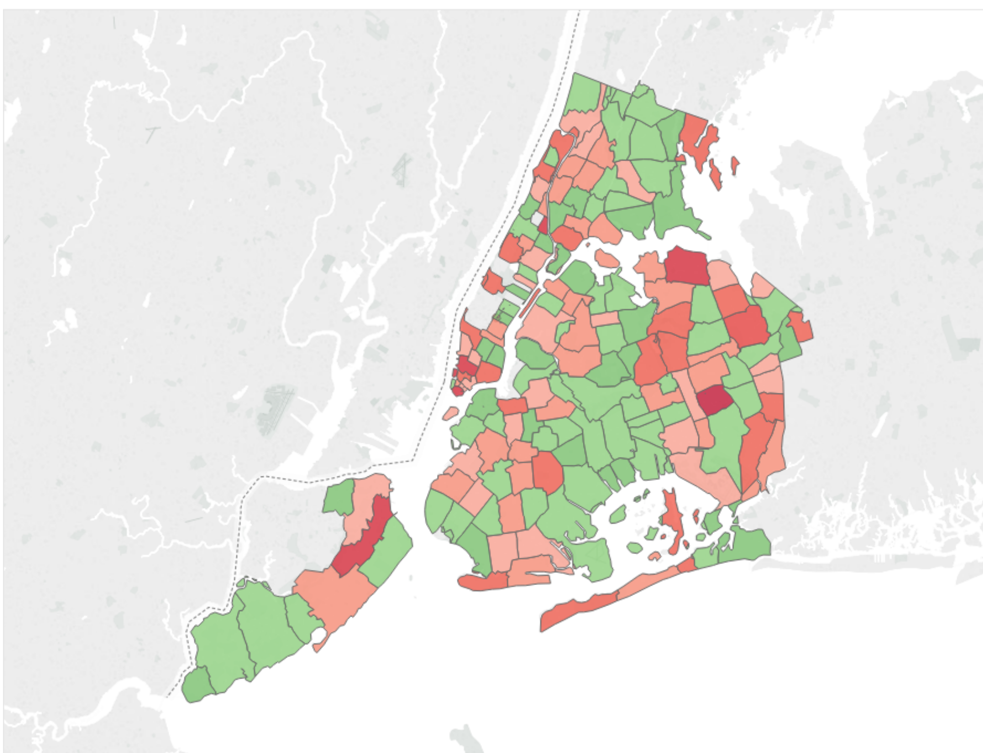
Map of Zip Codes and Nonmajor Crimes



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Nonmajor Crime. Details are shown for Zip Code.

Figure 3: Frequency map of non-major crimes based on zip codes in NYC

Map of Zip Codes and Misd Crimes



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Misd Crime. Details are shown for Zip Code.

Figure 4: Frequency map of misdemeanors based on zip codes in NYC

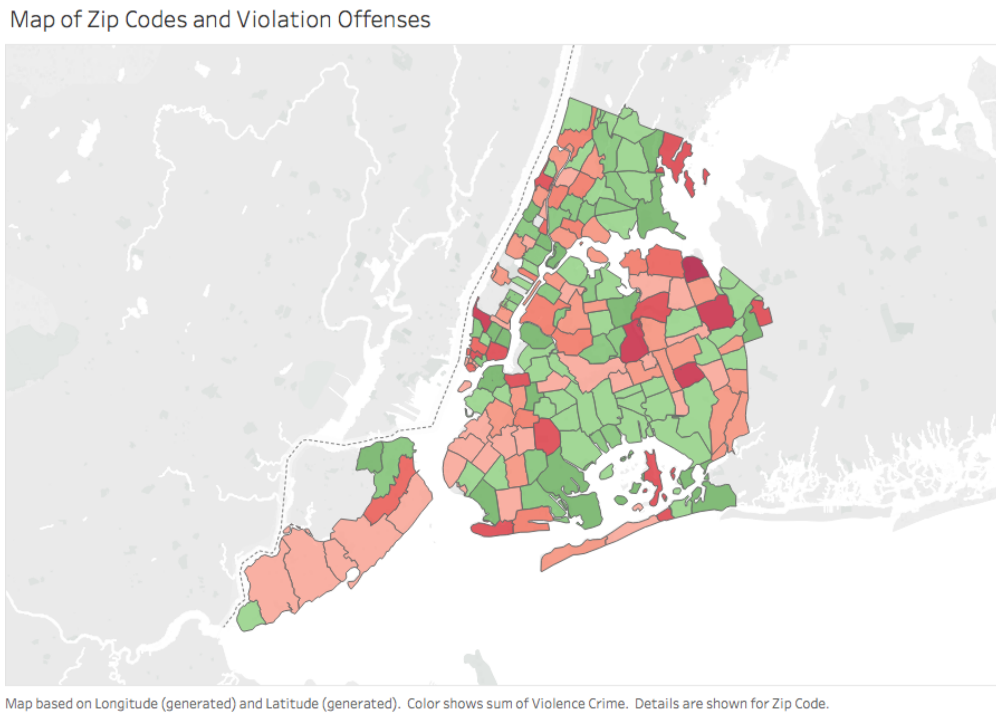



Figure 5: Frequency map of violation offenses based on zip codes in NYC

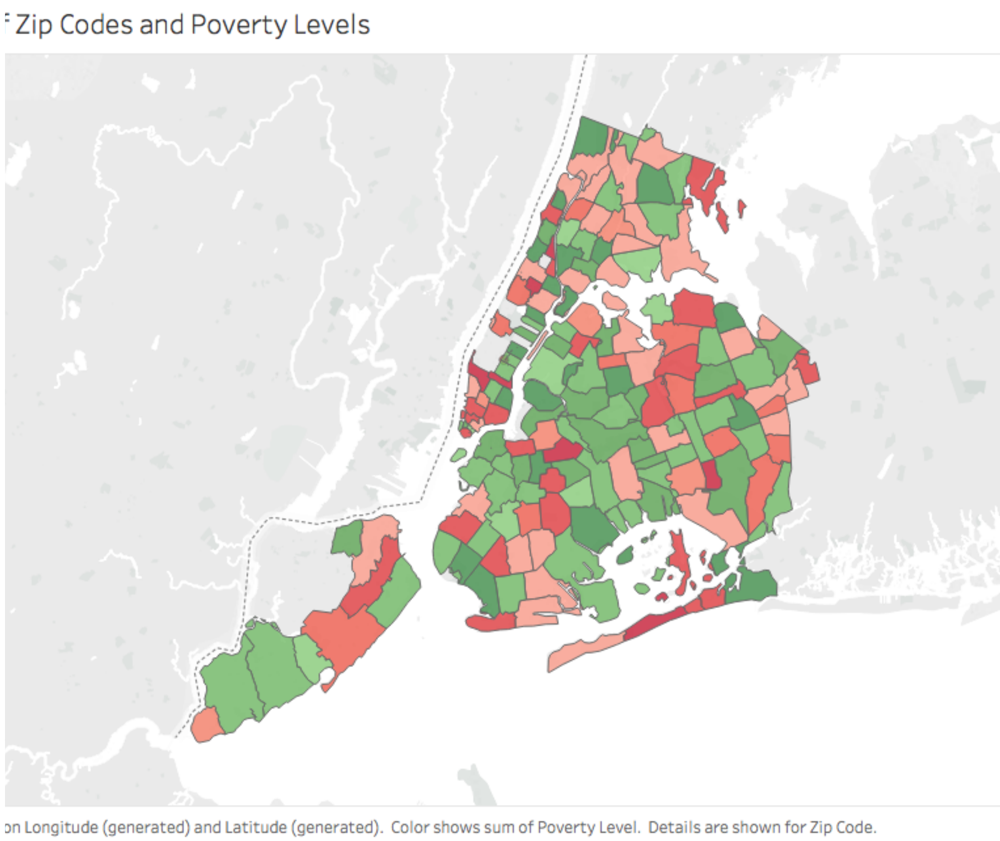



Figure 6: Frequency map of poverty levels based on zip codes in NYC

For the same purpose of calculation of correlations, we plotted filled maps in Tableau to foster us understand the results from pattern mining better, as we can see from Figure 1-6, which are the visualization for frequencies of different things we study based on zip codes in NYC, with all year's data combined.

In these maps, we employed median to divide colors, of which red indicates "higher than median" and green "lower than median". The darker the color for one area is, the more incidents took place. Since the data for calculating correlations and plotting maps are the same, we expect some coherent patterns on the maps here.

From Figure 1,3 and 5, we can clearly see that for most parts of Bronx, the results we had from getting correlations remain the same. When the areas on the noise map are red, it is highly likely that they are green on the other two maps for non-major felonies and violation offenses and vice versa. This confirms what we have found in the previous part.

Other relatively high correlations for Staten Island, on the other hand, seem less obvious but still noticeable: Given all areas on Staten Island are green in the noise map (Figure 1), figure 3 and 4, which are the maps for non-major felonies and misdemeanors, have less red areas compared with Figure 5 and 6. What is bizarre here is that Staten Island on the map for major felonies (Figure 2) has the least red areas, for which we would expect a higher positive correlation between noise complaints and major felonies. However, according to Table 1, we did not observe so. This requires that we dig deeper in the next part to find any potential explanations.

## 7. Pattern Mining

**Methods:**

To determine patterns between noise complaints, crime, and poverty, we stuck with the traditional data mining approach of finding how many noise complaints occur with certain markers of crime and poverty on the citywide level.

Beyond overall patterns, we sought to find temporal patterns and patterns by borough. In the former case, we used Pandas to separate the database by year, and removed the “year” column due to redundancy. In the latter case, we separated the database using the borough column and removed said column to prevent redundant patterns.
Our algorithm of choice is the FP-Growth algorithm, given its reputation of scalability over large itemsets versus Apriori. In particular, we chose the Python implementation by github user Evan Dempsey, since it allows the use of mutable elements like lists as inputs and also includes a function for generating association rules. We chose a support threshold of 10% and a minimum confidence of 70%; the former was initially 20%, but this resulted in some patterns, like the neighborhoods in the boroughs, not showing up at all.

1. **Results**

   *Note that all results are included with this submission. The results detailed here only list interesting patterns and association rules.*
   Citywide Patterns, all years

   First and foremost, we notice that Residential Noise Complaints make up about 57.4% of all noise complaints filed in NYC between 2010 and 2015. Furthermore, these noise complaints almost always occur with “Residential Building/House”, which indicates the complaint came from a residential building or house. Not a surprising result.

   What is surprising is the fact that while complaints about loud music or parties occur about 53.5% of the time, those complaints only come from residential callers about 38.3% of the time.

   We also notice that Manhattan makes up the single largest borough for noise complaints, with 17.9% of all noise complaints originating from there. Manhattan is followed by Brooklyn (17.7%), Bronx (14.5%), and Queens (17.7%), while Staten Island is not common enough to occur at 15%.

With crime, we note that noise complaints tend to occur in areas with lower than median crime. For example, the highest crime pattern we found was that noise complaints occur in areas with both lower than median misdemeanor and non-major felony crime rates at 21.4%, whereas the most common pattern with higher than median is violation offenses, at 16.4%.

Among crime patterns, we find that when noise complaints occur with lower than median misdemeanor, violation offense, non-major felony, or major felony rates, they also occur with other lower than median rates. For example, lower violation offenses occurred with lower major felony, non-major felony, and misdemeanor rates nearly 78% of the time, which is quite high.

The inverted case also holds: higher than median crime rates tend to occur with other higher than median crime indicators. However, these occurrences are much rarer, the strongest of which (higher major felony and violation offense) makes up only 14% of all noise complaints.

However, the patterns with poverty are not as clear citywide, though the patterns seem to edge towards higher poverty areas filing more complaints. For example, noise complaints happen that happen to be residential occur 29.7% of the time, whereas residential noise complaints occur in lower poverty areas make up about 27.7% of all noise complaints. While this difference is slight, it does suggest that noise complaints tend to originate in areas where poverty is higher.

The relationship between noise complaints, crime, and poverty is a fair bit clearer, however. We find that noise complaints that occur in areas of lower crime tend to occur in areas of less poverty, an association that is backed up by high confidence (>70%). Likewise, the converse is true in that noise complaints that occur in areas with higher crime also tend to occur in areas with higher poverty, also backed by high confidence.

We also notice that there it is hard to tell if the population density of a ZIP code is associated with noise complaints, given significant amount of overlap. We also notice a slight citywide trend towards areas of higher noise complaint density to file more complaints, but this should not be surprising. We will have to dig into the boroughs themselves to verify these patterns.

Regarding years, we did notice that 2013 made up the largest share of noise complaints at 16.0%, followed by 2012 at 14.2%, 2015 at 13.1%, 2011 at 12.5%, 2010 at 12.5%, and 2014 at 12.2%. This goes against the expectation that the six years provide approximately equal shares of the noise complaint total, and suggests there could be something interesting lying in wait if we mine patterns by year.

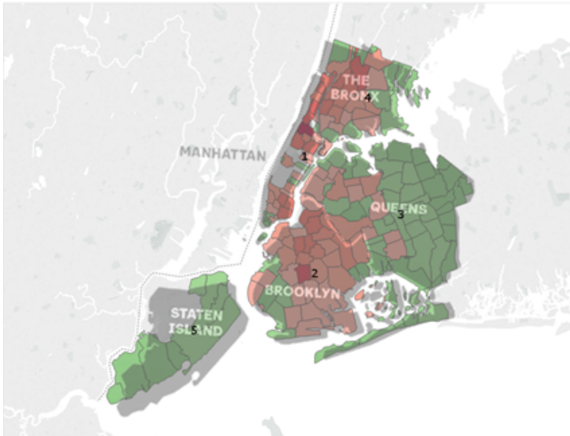
**Citywide Patterns, by Year**

In actuality, we find few to no major differences of note between the patterns of each year and compared to the citywide patterns for all years. This suggests that the distribution of noise complaints are independent of year, though we are unsure why 2013 had more noise complaints total than the other years. Perhaps the only interesting pattern to pop up is that for 2015, complaints about noise on the street or sidewalk show up about 18.4% of the time, compared to ~15% for the previous years and 16% for the city over all years. However, since 2015 is not the most common overall year of the years we studied, the overall patterns average to favor residential noise complaints.

**Citywide Patterns, by Borough**

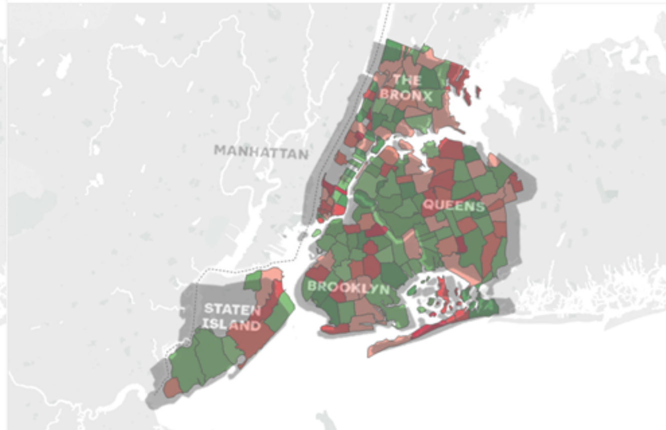
Since useful info could not be affirmatively derived from the citywide data about poverty, we decided to dig deeper and try to figure out why this is the case on a borough-by-borough basis.

Map of Zip Codes and the Number of Noise Complaints



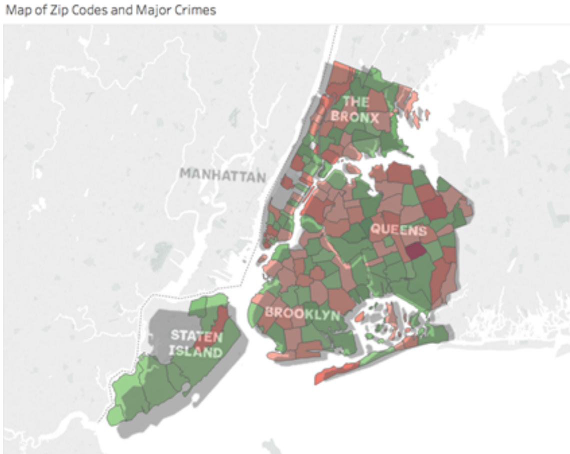

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Created Date. Details are shown for Zip Code.

Map of Zip Codes and Poverty Levels



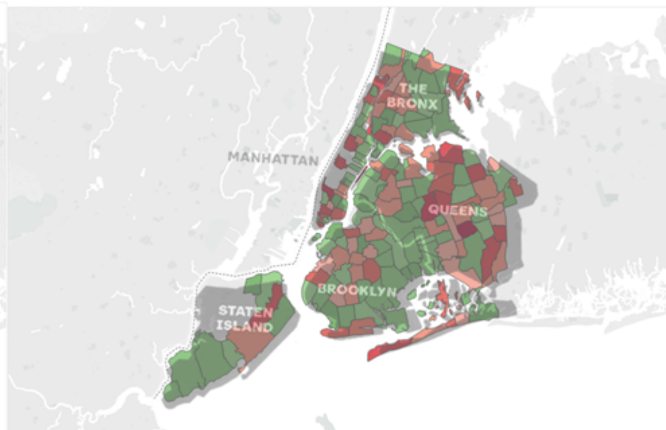

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Poverty Level. Details are shown for Zip Code.

Map of Zip Codes and Major Crimes



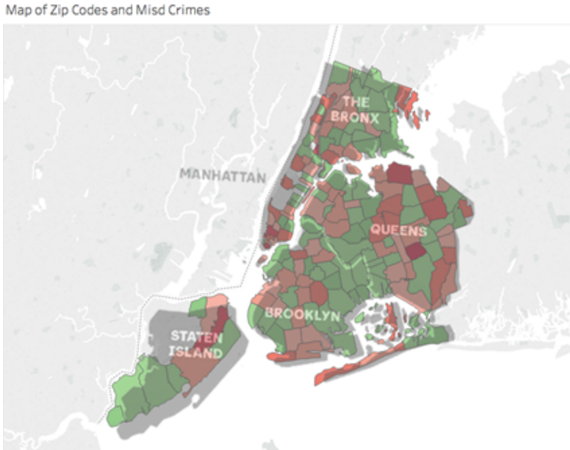

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Major Crime. Details are shown for Zip Code.

Map of Zip Codes and Nonmajor Crimes



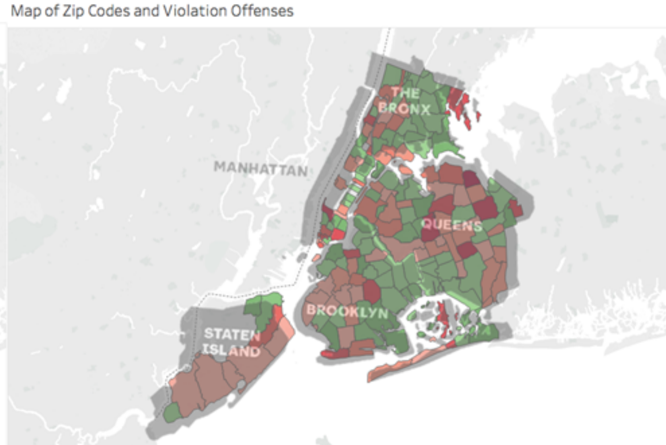

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Nonmajor Crime. Details are shown for Zip Code.

Map of Zip Codes and Misd Crimes




Map based on Longitude (generated) and Latitude (generated). Color shows sum of Misd Crime. Details are shown for Zip Code.

Map of Zip Codes and Violation Offenses




Map based on Longitude (generated) and Latitude (generated). Color shows sum of Violence Crime. Details are shown for Zip Code.

<u>Brooklyn:</u>

Immediately, we notice that 58.5% of all noise complaints in Brooklyn are residential in nature, versus 57.4% for the city as a whole, which suggests Brooklyn more closely resembles the distribution of the city than the other boroughs. Brooklyn also associates noise complaints with areas of greater noise complaint density and higher population density.

With regards to crime, we notice that in Brooklyn noise complaints tend to occur in areas with lower than median crime, with the notable exception of major felony rates, and possibly misdemeanor rates. The strongest pattern with major felony rates higher than median occurs 20% of the time, whereas the strongest pattern with lower than median major occurs at 15%. We found an interesting pattern with regards to misdemeanors: while noise complaints occur slightly more frequently in lower misdemeanor areas (24.5%), we found that when those noise complaints are residential noise complaints, the misdemeanor rates are higher than median (22.2%). Because of this, it is difficult to say affirmatively if noise complaints are associated with fewer misdemeanors overall, but it is suggested that areas with more misdemeanors tend to file more residential noise complaints. Moreover, higher misdemeanor rates and higher major felony rates frequently occur together 20.7% of the time, which further suggests that areas that file many residential noise complaints may also have higher misdemeanor and major felony rates.

In Brooklyn, we found that noise complaints are skewed towards lower poverty areas, with 57.7% of all noise complaints originating from such areas. This extends to residential noise complaints, which make up 58% of all noise complaints in Brooklyn; 31.1% of all noise complaints that happen to be residential originate from low poverty areas versus 27.3% from high poverty areas.

Unlike Manhattan below, no one neighborhood dominates Brooklyn, with Central Brooklyn at 15%, and Bushwick and Williamsburg, Northwest Brooklyn, and Flatbrush all showing up about 11-12% of the time. Given our threshold of 10%, the other neighborhoods are too infrequent to show up.

With regards to noise complaints, crime, and poverty, we find that lower poverty tends to occur with lower crime rates across the board, despite the relatively high occurrence of misdemeanors.

Manhattan:

We first notice that residential noise complaints make up only 41.6% of all noise complaints in Manhattan, which is much lower than the citywide figure of 57.4%. We postulate that this is because Manhattan is likely more commercial than other areas of New York City, so noise complaints are more likely to come from the street or sidewalk (15%) and commercial callers (13.7%) than the other boroughs, which are more residential. We also noticed that Manhattan had strong patterns towards lower population density and higher noise complaint density versus the city at large.

With regards to crime, nearly 60% of all noise complaints occur in areas of lower than median violation offense rates, and nearly 59% of all complaints occur in areas of lower than median major felony rates. We also find that lower than median crime rates tend to occur with other lower than median crime rates. For example, about 35.7% of all noise complaints were filed in areas that had both lower major felony and lower violation offense rates.

We find that in Manhattan, noise complaints tend to occur significantly more often with higher than median poverty levels than with lower than median poverty levels, at 58.2% versus 28.4%. Given that Manhattan makes up a large part of all noise complaints, this could partially explain why it is hard to tell whether lower or higher poverty rates are associated with noise complaints on a citywide level. Between poverty and crime rates, it is evident that in Manhattan noise complaints tend to occur in areas of lower crime and higher poverty, which puts it at odds with the other boroughs, such as Brooklyn and the Bronx. For example, the most common poverty-related pattern is that noise complaints occur in areas with higher than median poverty and lower than median major felony rates about 34.4% of the time. The next highest shows higher than median poverty occur with lower than median major felony rates 26.3% of the time.

Another thing of note is that out of the neighborhoods in Manhattan, Inwood and Washington Heights is by far the most common in terms of noise complaints, appearing as often as 21.0% of the time; the next most common neighborhood is Central Harlem, at a distant 15.9%. Inwood and Washington Heights also share many of the previous observations, having higher than median poverty yet lower crime overall. This may be skewed by the fact several ZIP codes in Manhattan may not be properly matched to a precinct. The commonality of Inwood and Washington Heights may also explain the tendency towards lower population density, despite Manhattan having a perception of being very urbanized, since that neighborhood has relatively lower-density housing than the rest of Manhattan.

Queens:

Off the bat, we notice that about 71% of all noise complaints in Queens come from ZIP codes that file fewer noise complaints on median than NYC as a whole. On top of that, 62% of all noise complaints are residential noise complaints, compared to 57.4% to the city as a whole. Queens associated noise complaints with lower population density and lower noise complaint density versus NYC at large at least 44.6% of the time.

With regards to crime, it is evident that Queens follows the common trend of lower crimes overall with noise complaints; with the significant exception of higher than median major felony and violation offense rates, which tend to occur far more frequently with noise complaints than lower than median cases; higher than median major felony rates occur with residential noise complaints 29.6% of the time, and higher violation offense rates 28.8% of the time. Unlike the other boroughs, higher than median major felony rates also tend to occur with other lower than median crime rates, such as higher major felony rates co-occurring with lower than median non-major felony rates 26.1% of the time. This also applies to the higher violation offense rates, which occur with lower misdemeanor rates 20% of the time. This is despite the fact that lower than median misdemeanor, non-major felony, and violation offense rates tend to occur with one another.

Queens definitely shows a strong pattern of lower than median poverty ZIP codes filing noise complaints more so than more impoverished ZIP codes, with the latter occurring 38.3% versus 32.7% of the time.

Going from the crime observations from before, lower poverty codes correspond to higher major felony and violation offense rates, along with lower non-major felony and misdemeanor rates - a pattern backed up by a strong association of 76%.

West Queens and Southwest Queens are the most common neighborhoods for noise complaints, making up 17.3% and 16.3% of all noise complaints, respectively. Jamaica and Northwest Queens show up less frequently at about 12% each. We find that Southwest Queens has strong patterns towards higher major felony and violation offense rates, with its strongest pattern involving both of them - suggesting that every ZIP code in this neighborhood has a higher than median major felony and misdemeanor rate, along with lower crime statistics elsewhere. West Queens has no particularly strong crime patterns until it gets one with lower than median violation offenses at 11%. This suggests that the strong patterns in Southwest Queens are pulling the data in Queens towards that strange result. However, none of the neighborhoods had any interesting patterns with regards to poverty, which suggests that the neighborhoods in Queens are fairly diverse financially.

<u>Bronx:</u>

We can immediately notice that 75% of all noise complaints filed in the Bronx are residential noise complaints, which is much higher than the citywide figure of 57.4%. The Bronx also associates noise complaints with areas of higher population density at least 70% of the time, and those areas tend to file more noise complaints per ZIP code than NYC on median.

The Bronx significantly deviates from the other boroughs in that noise complaints tend to occur in areas of higher than median crime. For example, noise complaints originate in areas with higher than median misdemeanors about 64% of the time, which is exceptionally high. Like the other boroughs, high crime tends to occur with high crime in the Bronx, with high felony rates occurring with high nonmajor felony rates (36.6%), for example. This is backed by a high confidence of nearly 100%.

Noise complaints in the Bronx are definitely associated with higher poverty levels, as residential noise complaints occur with the latter 44.4% of all noise complaints, versus 30.9% for all noise complaints.
Since the Bronx associates noise complaints with higher crime and higher poverty rates, it would be logical to assume that crime and poverty are directly associated with each other, which is true. For example, higher misdemeanor rates are associated with higher poverty rates for 40.2% of all noise complaints; nonmajor felonies are associated with both higher misdemeanors and higher poverty rates 33.6% of the time; while higher violation offense and higher major felony rates occur with higher poverty levels 27.3% of the time, and all of these patterns are backed by high confidence. These are exceptionally strong patterns by the standards of the boroughs, and show that the

Among the neighborhoods of the Bronx, Bronx Park and Fordham had the strongest showing at 24.2%. Central Bronx comes next at 17.5%, followed by High Bridge and Morrisania at 17.2%. The latter is interesting because unlike the former two, its noise complaints come predominantly from ZIP codes with lower than median poverty rates. Its noise complaints also occur with lower major felony rates, in contrast to the other neighborhoods and Bronx as a whole, though misdemeanor, violation offense, and nonmajor felony rates are still higher than median. The former two, while more common, share common observations with the Bronx as a whole, which is not surprising given they are the largest sources of complaints in Bronx.

Staten Island:

Residential noise complaints make up 57.6% of all noise complaints on Staten Island, versus 57.4% for New York City as a whole. Like Queens, Staten Island associates noise complaints with lower than median population density 78.8% of the time. Another interesting pattern is that every noise complaint occurs in an area that files fewer noise complaints than the median of NYC, but this may be explained away by Staten Island having the smallest population of any NYC borough.

Staten Island sees most of its noise complaints occur in areas of lower than median crime, and like the city as a whole, lower than median crime rates tend to occur with other lower than median crime rates. For example, 59.5% of all noise complaints come from areas with lower than median major felony rates, and lower major and non-major felony rates occur with each other 36.8% of the time.

Surprisingly, 56.5% of all noise complaints on Staten Island come from ZIP codes that are poorer than the median, which is a higher proportion than the other boroughs.

Stapleton and St. George are by far the most common source neighborhoods of noise complaints, making up 49.4% of all complaints. South Shore comes next, making up 32.6% of all complaints. We observed both strong patterns shared by Staten Island as a whole for lower crime and higher poverty. Port Richmond brings up the rear at 14.7%, and is slightly interesting in that it has its noise complaints come from areas with lower poverty; since it still has the same patterns of lower crime, it is possible this is the case because Port Richmond just happens to be wealthier than the rest of Staten Island, but is not large enough to offset the former two neighborhoods. Mid-Island is not frequent enough to appear, but presumably is below 10%.

## 8. Results Discussion

Overall, the data suggests that most noise complaints originate from ZIP codes of New York City with lower overall crime rates. It is difficult to say definitively that higher poverty areas file more noise complaints than lower poverty areas, and deeper pattern mining by borough suggests that the poverty relationship varies among boroughs. Brooklyn and Staten Island have reasonably strong relationships between noise complaints and lower poverty levels, whereas Manhattan, the Bronx, and Queens have stronger relationships between noise and higher poverty rates. Trying to explain why this may be case would require a deeper analysis of the demographics of the boroughs, most likely taking into account the housing situation. For example, is it possible that areas where more people live in project high-rise apartments like Manhattan and the Bronx are poorer than areas where more people live in dense urban housing like Brooklyn, or suburban housing like Queens and Staten Island? This might explain why noise complaints in Brooklyn come from wealthier, high density areas versus poorer, higher density areas like in the Bronx.

Even then, we feel it is fair to say that if you are in an area of New York City that gets a lot of noise complaints, you can rest assured that area is safer, even if it is likely to be poorer depending on which borough you are in.

## 9. Future Work and Areas of Improvement

For crime data, we had the issue of the precincts not corresponding to specific ZIP codes geographically. This required mapping the precincts to the ZIP codes they approximately correspond to. Perhaps in the future, we could try figuring out a model that approximates crime levels to each ZIP code depending on the surrounding precincts, but due to time constraints this was not looked into with serious detail.

Pattern mining has the limitation of not showing us how areas that have a low occurrence of noise complaints are distributed. We cannot say for certain that areas with fewer noise complaints have higher crime, or less poverty; we can only say that areas with more noise complaints tend to have less crime and slightly greater poverty, dependent of borough. We could probably remedy this by compiling a list of areas with fewer noise complaints and trying to find patterns or correlations, but given the fact we only have 176 ZIP codes total in NYC, our dataset would be quite small and possibly biased towards certain attributes. Controlling for population may help, but we will not be certain of this unless we actually attempt this operation.

Also, there is deficiency of running all algorithms for pattern mining on the city level that we have learnt due to different volumes of data for different boroughs, which may lead to skewness or distortion of results. Although running them on each borough looks like one solution, it is not plausible because we for sure will lose insight and grasp on the macro level. In future research, we are going to improve, for example, FP-Growth by designing some mechanisms or models specifically for this real-world scenario that takes more meaningful parameters so that it will take all boroughs equally on the city level.

## Sources


1. J.M. Field, *Effect of personal and situational variables upon noise annoyance in residential areas*, Journal of the Acoustical Society of America, 93: 2753-2763 (1993)
2. Zheng, Y. et. al. "Diagnosing New York City's Noises with Ubiquitous Data." Ubicomp 2014. Published 17 September, 2014. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Diagnosing20NYC27s20noise20using20ubiquitious20data-yu.pdf.
3. NYC Information Technology and Telecommunications, and NYC Mayor's Office of Data Analytics. "NYC Open Data." NYC Open Data. The City of New York, n.d. Web. 01 Nov. 2016. https://data.cityofnewyork.us/.
4. NYPD Historical Crime Statistics Database. New York City Police Department. Web. 15 Dec. 2016. http://www.nyc.gov/html/nypd/html/analysis_and_planning/historical_nyc_crime_data.shtml
5. ZIPAtlas: Percentage of Population below Poverty Line in New York State by ZIP Code. ZIPAtlas. Web. 15 Dec. 2016. http://zipatlas.com/us/ny/zip-code-comparison/population-below-poverty-level.htm
6. Dempsey, E. "PyFPGrowth". Github. https://github.com/evandempsey/fp-growth
7. Python Packages: PANDAS, Numpy, Tableau