Improving the Response to NYC "Social Distancing Violation" Reports

Greg Berry

April 09, 2020

1. Introduction

1.1 Background

As the Covid 19 pandemic sweeps across the United States many states, counties and cities have implemented "social distancing" (SD) laws relating to a restriction on the number and distance at which citizens can congregate. New York City (NYC) has facilitated enforcement of their local distancing ordinance by allowing citizens to report suspected violations by phone, and on-line using a "311" service, which is analogous to the "911" emergency response system but for lower urgency issues. The NYC Police Department (PD) assigns each report to one or more officers for review, and investigation at the location of the complaint. Many of these reports turn out to be non-actionable (i.e. the investigation does not result in a successful resolution of the issue or issuance of a violation to suspects.)

1.2 Problem

Responding to non-actionable reports requires a police department to task officers to non-productive work that could be assigned to other emergency related tasks. This project attempts to lower the number of unactionable social distancing reports by looking at associated date from two approaches:

A. Determine if there are any correlations between 311 SD-Reports (311SDR) and the types of social venues within the proximity (i.e. 160 meters, or about 2 city blocks) of the 311SDR.

B. Detect and define distinct clusters within the 311SDR that might indicate locations that could be included in more routine Police Department patrols to deter social distancing violators and thus reduce the necessity of citizen reports and one-off dispatches of officers to handle those calls.

C. Review 311SDRs to consider factors that might be contributing to the probability of a report being actionable.

1.3 Interest

This type of information should be of interest to police department leadership as a factor in their pursuit of effectively assigning officers to types of neighborhoods to patrol, as well as in their efforts to predict which reports would likely result in effective action, so as to allow them to prioritize their response to reports in times when police staffing resources are limited. The leaders of the communities (e.g. community boards) where social distancing reports were made would also likely be interested in understanding how their community compared to others in terms of social distancing related incidents.

2. Data Acquisition and Cleaning

2.1 Data Sources

311SDRs can be found in the NYC Open dataset found here: (https://opendata.cityofnewyork.us/calls). Data related to the social venues in the immediate region around the location of the social distancing reports can be found using the Foursquare API found here: (https://foursquare.com/).

2.2 Data Cleaning

NYC Open Data for 311 was retrieved using a RESTful call to NYC 311 Data API with the with a descriptor filter on "Social Distancing". This brought back a JSON response of about 1,000 rows of information, across 30 categories, detailing individual 311 calls. The JSON file was converted to a pandas DataFrame. Rows with no associated latitude or longitude were removed.

Foursquare data was retrieved using RESTful calls to their "venues/explore" API to capture most (i.e. I limited the number of venue records per 311DR to 100) venues (e.g. restaurants, and stores) within 160 meters of the latitude and longitude for each 311SDR. This resulted in a return of about 7932 rows, which was too many calls for the API call limit of my "sandbox" account on Foursquare and I had to register for a free "personal" account to increase my API call limit to an acceptable level. An additional RESTful call was made to the Foursquare "venues/categories" API to capture the text descriptions about each venue and a reference to the categories parent entry in the hierarchy. The venue category hierarchy was composed of 944 rows. Foursquare venue data was also initially in JSON format and I stored each row of foursquare venue data as an additional column on the 311SDR DataFrame which contains the whole of the Foursquare venue data for the 311SDR's latitude/longitude.

2.3 Feature Selection

After data cleaning the 311SDRs there were approximately 985 rows of 311SDRs, each with 14 columns of data, left in the DataFrame. After reviewing the available columns for the 311SDR data I determined there were multiple columns (e.g. cross-street1, ZIP, x_coordinate_state_plane, etc.) that were not needed for this project which were removed. Several columns (i.e. open data channel (i.e. online report or phone-in report), date/time, and community board) were converted to a floating-point representation to make it easier to do attribute correlations at a later phase.

The Foursquare venue category for a given foursquare venue was not initially retrievable. From the initial raw category data I defined a python dictionary to allow retrieval of the venue name for a given venue code (e.g. CA44234234 might equal "Samantha's Diner"), and to return the highest-level venue category for a given code (e.g. "Samantha's Diner" is a member of the "Food" hierarchy).

There are many thousands of different individual venue descriptions in NYC. To make the Foursquare venue data more readily computable for the purposes of this project I created different data buckets to contain counts of venues by top-level category for each 311SDR.

3. Exploratory Data Analysis

3.1 Calculation of target variable

The 311SDR data did not have a specific feature that summarized whether a call ended with an actionable or non-actionable resolution, however it did contain a field with one of a set of text strings describing the result of the call. I created an “actionable call” feature that mapped each 311SDR result text string to one of three possible floating-point values, each summarizing a different grouping of outcomes:

- 0.0 : Non-Actionable : This SDR311 report with the police department officer involved not being able to take a final action on the call for one of multiple reasons (e.g. “Not enough information was given in the report to locate the suspect(s)” or “The suspect(s) left the location before the police officer arrived”,etc)
- 0.5 : Partially-Actionable : This SDR311 report required more information / investigation by the police, and was still in progress for being resolved
- 1.0 : Actionable : This SDR311 report resulted in the police officer(s) successfully dispersing and/or ticketing the suspects.

3.2 Relationship between actionable calls and nearby Foursquare venues

It was a meta-goal of mine that this project would serve as the capstone project for an “IBM Data Science Professional Certification” on Coursera. One of the stipulations of the course’s capstone assignment was that Foursquare data be used as a part of a study. My original hypothesis was that the types and quantities of venues near the location of a 311SDR would have a correlation to whether the 311SDR would ultimately be actionable. The correlation analysis on the attributes of the 311SDR calls and Foursquare venue counts by category determined there was little to no correlation present.

	actionable_call
actionable_call	1.000000
online_call	-0.021291
part_of_day	-0.034183
community_value	0.084142
cat_counts_food	-0.025704
cat_counts_outdoor	0.050261
cat_counts_shop	-0.015046
cat_counts_nightlife	-0.018900
cat_counts_travel	0.000642
cat_counts_arts	0.053788
cat_counts_residence	0.037989
cat_counts_professional	-0.050298
cat_counts_event	NaN
cat_counts_college	-0.035259
TOTAL_count	0.017273

Figure 1 – Correlation Analysis of actionable calls to counts of venues in proximity to the call

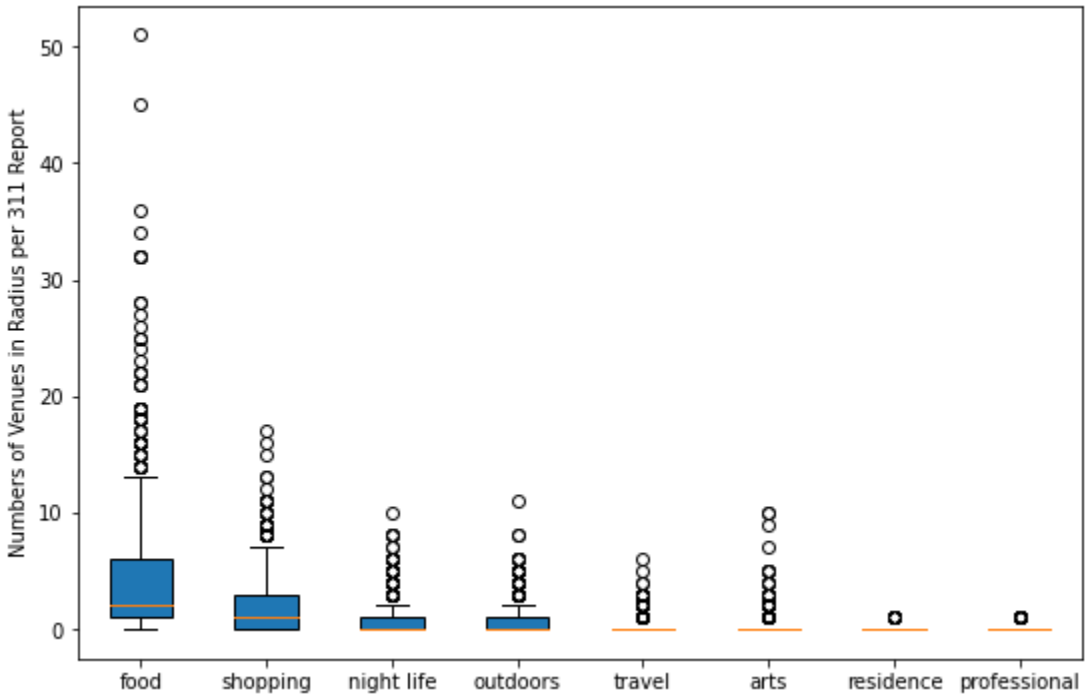

Figure 2 –Venue Category Counts Per 311SDR

I made multiple attempts to adjust the radius and limit of the Foursquare venues search calls, which resulted in significant changes to the counts for the venue categories and re-ran the correlation function but with little change in levels of correlation. This has led to two possibilities. The first, and likeliest, is that there is no correlation between the types of venues around a social distancing incident and the outcome of a police response to the incident, or secondly that finding such a correlation might be achieved by looking at venue counts at a lower level of the hierarchy of aggregation (e.g. Samantha's diner would be rolled up to the category of "American Food" instead of the higher level category of "Food") and/or combinations of categories. Unfortunately, such repeated re-grouping and re-analysis of the associated Foursquare data is outside the scope of effort that has been allocated to this project.

3. Clustering

There are five main types of clustering algorithms that could potentially be used to analyze the clustering characteristics of the 311SDR data including: K-Means, Man-Shift, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation – Maximization (EM) Clustering using Gaussian Mixture Models (GMM) and Agglomerative Hierarchical clustering. Because the 311SDR and Foursquare Venue data is primarily geography based (i.e. using latitude and longitude) and irregularly distributed I chose to use a DBSCAN clustering approach to this analysis.

3.1 Geographical Visualization

Even though there is no evident correlation between Foursquare venue locations and 311SDR locations it is worth looking at visualizations of these locations on a map as shown in the figure below.

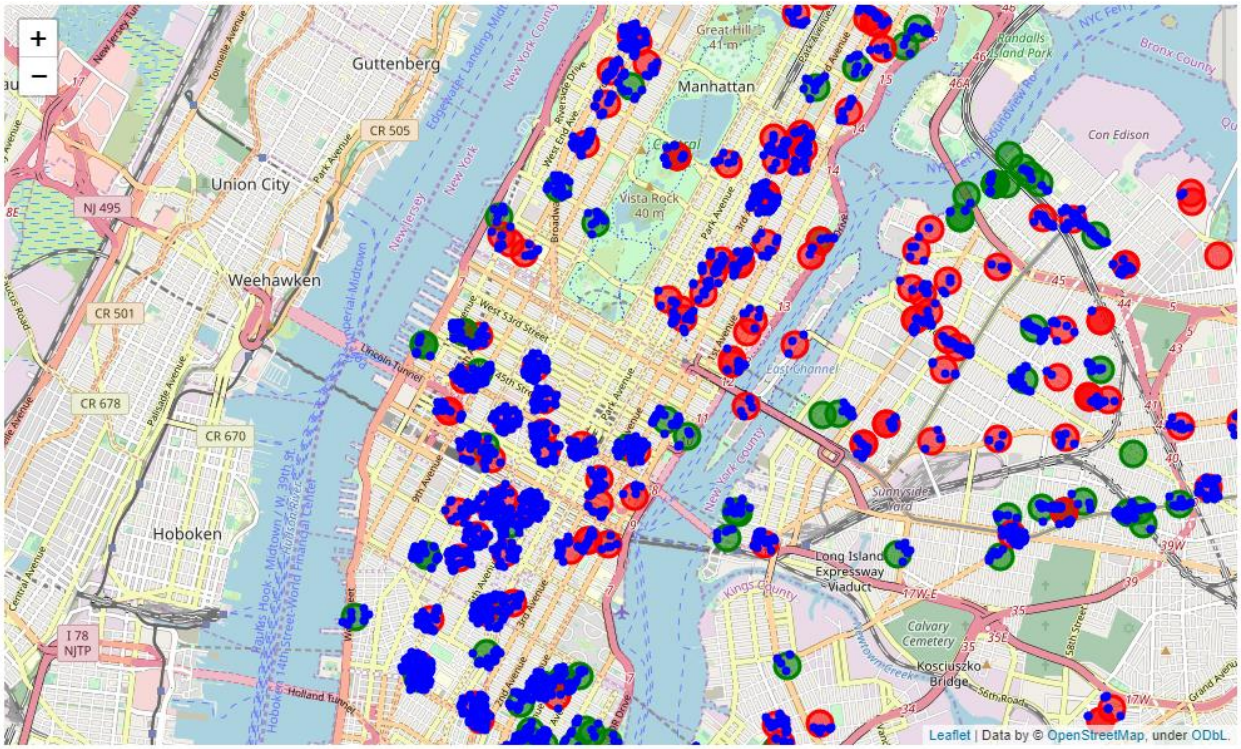

Figure 3 – Display map (Social Distancing Reports with Social Venues)

Blue dots are Foursquare Venues, Green Circles are Actionable 311SDR Calls, Red Circles are Unactionable 311SDR Calls

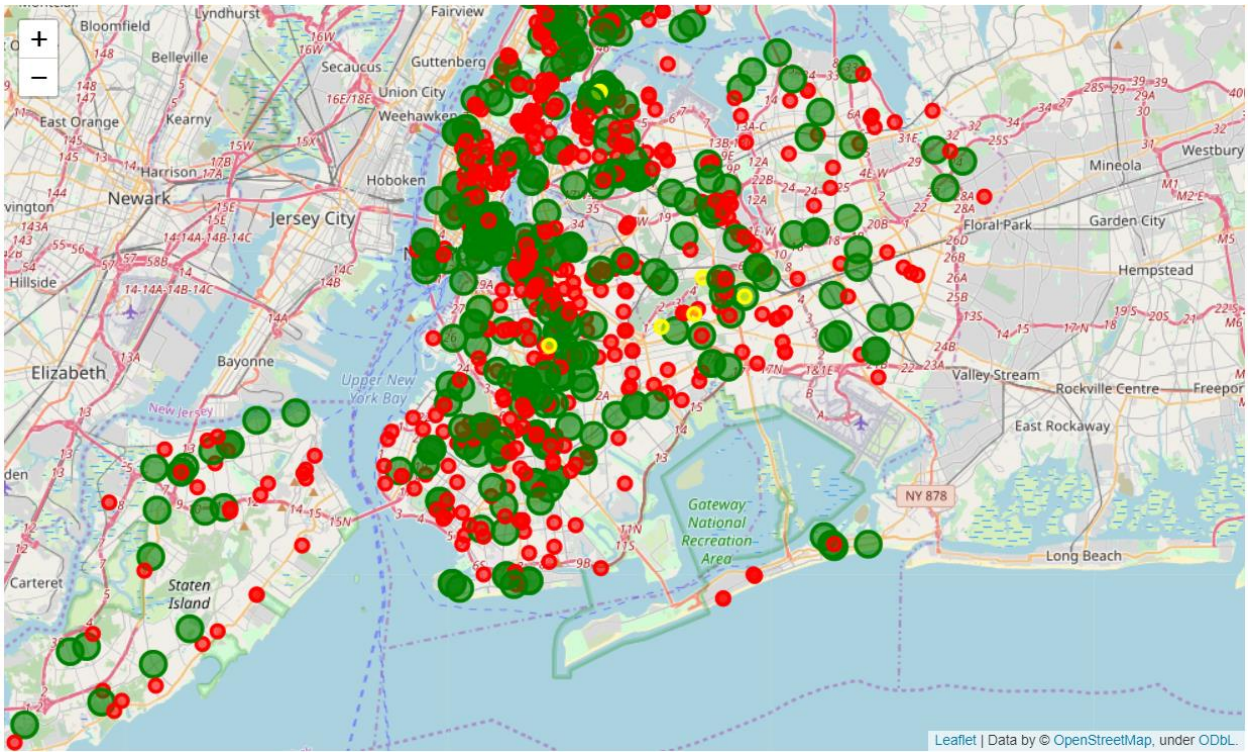

Figure 4 – Display map (Social Distancing Reports with Social Venues)

Green Circles are Actionable 311SDRs, Red Circles are Unactionable 311SDRsl, Yellow Circles are in-progress 311SDRs3.2 DBSCAN Visualization

The configurable variables affecting a DBSCAN clustering analysis are the epsilon, allowed distance between data points (i.e. latitude/longitude for this data set), and sample size that regulates how many points are required to be near each other for a cluster to be reported.

I ran several visualizations for this project to consider which might best reflect true clustering of data. When looking at the visualizations for the DBSCANs it is important to compare them with the geographical visualization that shows the data in context to a map and the colors/types of the data points to see if the clusters are best reflecting an “eyeball” assessment of the data. Three of these visualizations are provided in the following figures.

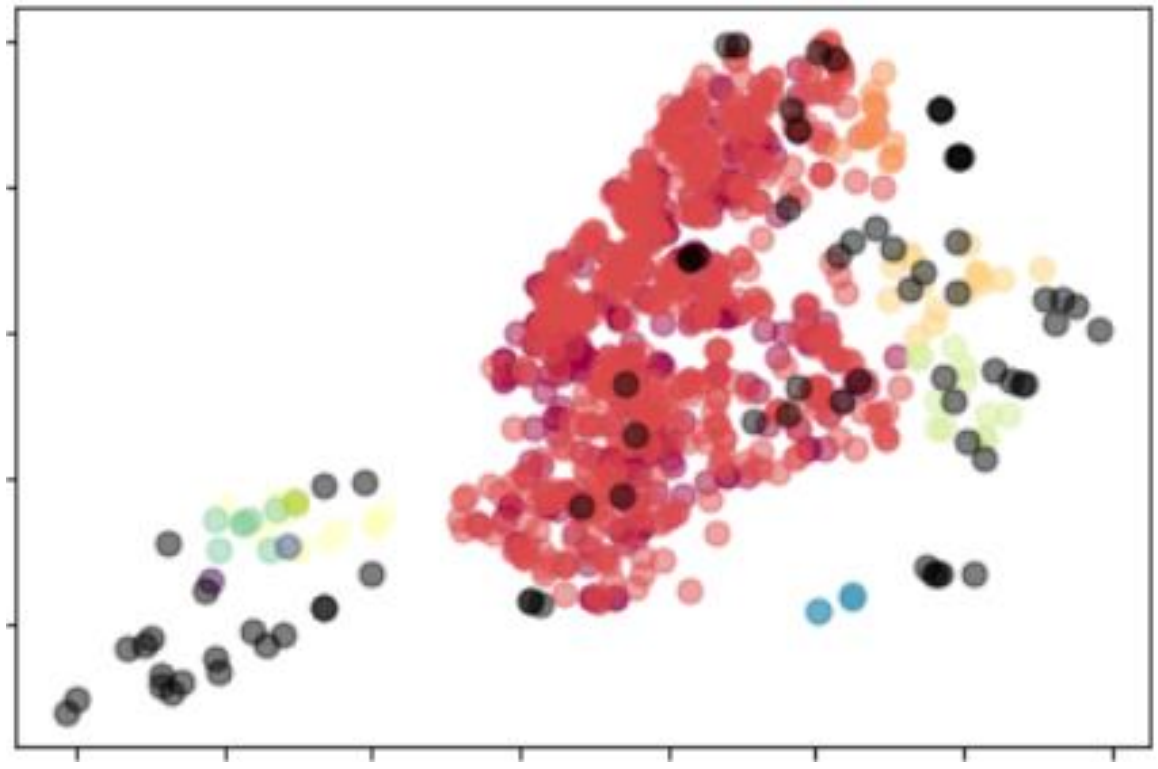

Figure 5 – DBSCAN (epsilon=.3, samples=5)

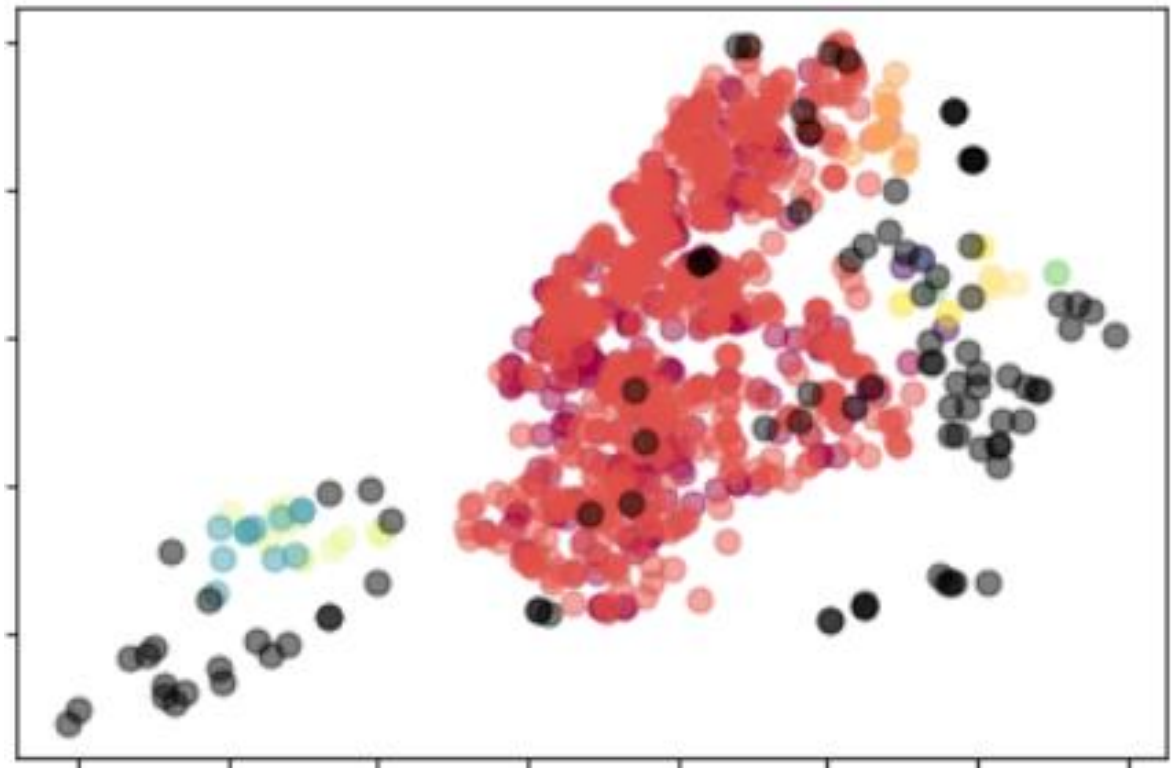

Figure 6 – DBSCAN (epsilon=.3, samples=6)

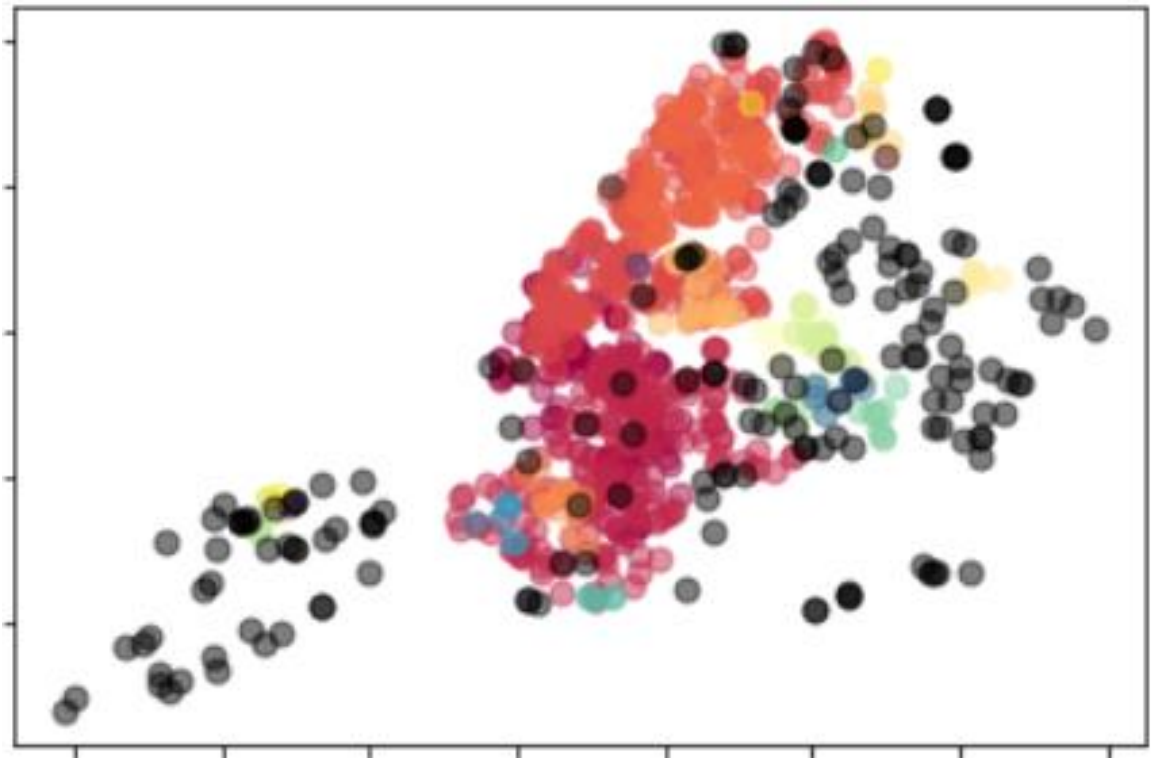

Figure 7 – DBSCAN (epsilon=.2, samples=5)

4. Conclusions

This study disproves my first hypothesis of any correlations between 311 SD-Reports (311SDR) and the types of social venues within the proximity (i.e. 160 meters, or about 2 city blocks) of the 311SDR. However, the study does show that there are significant clusterings of both actionable and non-actionable reports based on geographic location. These clustering bear closer investigation as they suggest there may be explainable reasons for why some areas of NYC have a much higher/lower actionable rate of 311SDRs. The NYC Open Data for 311 Calls data has “community_board” mapping, as showin in Figure 8 below, is an example of visual data that can potentially help interpret the borders of some of the clusters.

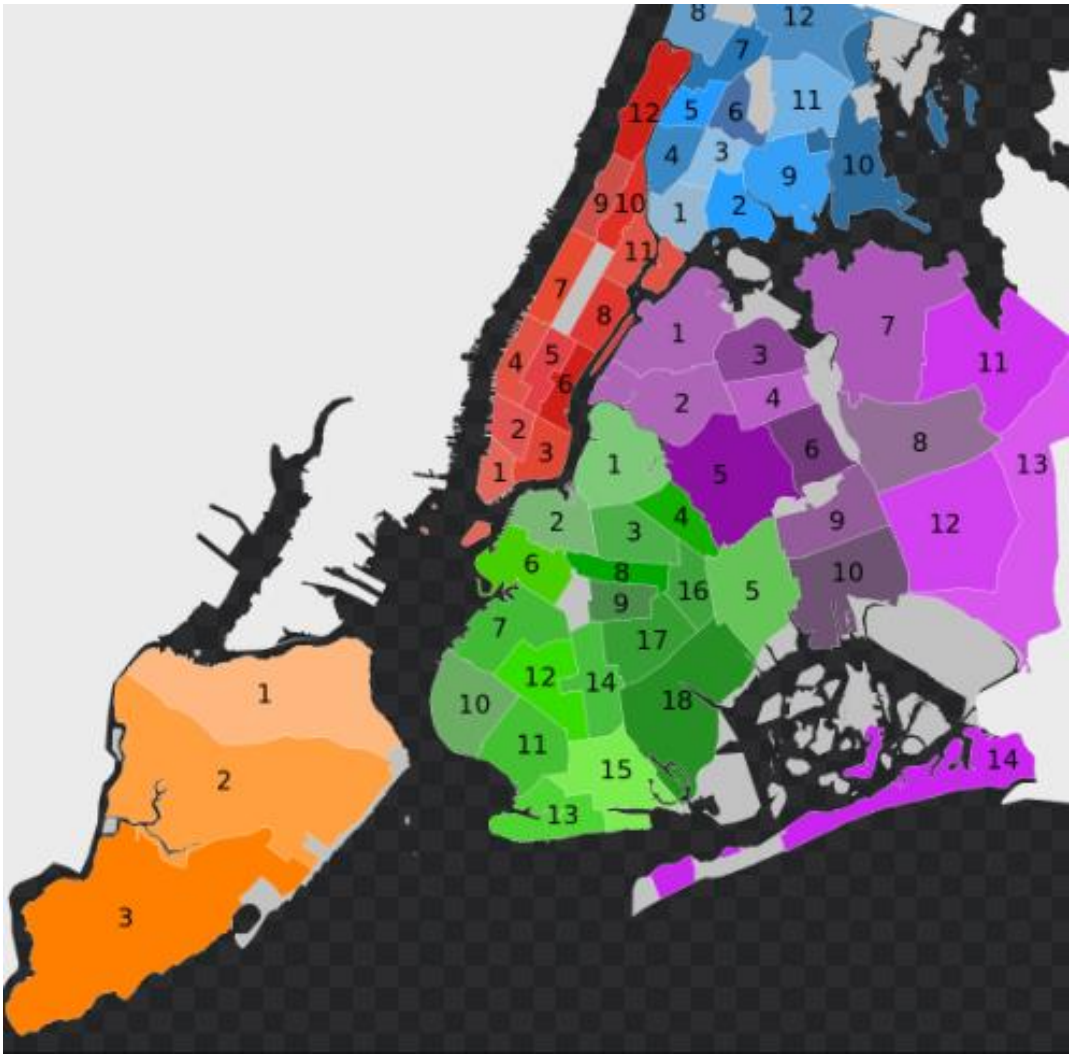

Figure 8 (From Wikipedia) – NYC by Community Board

5. Future Directions

As mentioned in the conclusion, there are definite clusters in the locations of actionable/non-actionable 311SDRs. I think that with additional data sources for such things as locations of police departments (to get a value for distance between 311SDR issue and police manpower), population density in NYC by latitude/longitude, wealth/income in NYC by latitude/longitude, many of the clusters could be explained and used to make recommendations for improved police department and neighborhood actions to social distancing issues. Hopefully in a few weeks the need to respond to 311SDRs will be moot, but the need to look at other subsets of 311 calls might make this study, and any improvements to its data sources, a valuable resource to future NYC planners and analysts.