Analyzing the Correlation between NYPD Complaints Location and Nearby Venues

Daniel Ramos

September 29, 2019

1. Introduction

1.1 Background

The city of New York is one of the most popular cities and one iconic cultural and financial center in the United States. It is one of the densest city in the world with a population of around 20 millions of habitants.

Bigger cities have bigger problems, security is one of them. The New York Police Department (NYPD) have around 95,883 complaints reported in 2018.

1.2 Problem

The main purpose of this investigation is to determine what are the most dangerous zones in the city and what venues are most popular around those points. For practical purposes only the borough of Manhattan is being used for observation.

1.3 Interest

The security of the citizens must be the highest priority to any security law enforcement organization and its governments.

2. Data acquisition and cleaning

2.1 Data sources

The main dataset used in this project is made available publicly by the New York Police Department.

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data

In this dataset are all the complaints reported to the Police Department in the current year and only in the borough of Manhattan. This filtered dataset is made available in the GIT repository of this project in the following link.

https://raw.githubusercontent.com/devrrad/Coursera_Capstone/master/NYPD_Complaint_Data_2019_Manhattan.csv

The API service provided by Foursquare is also used to get info about the venues around.

2.2 Data cleaning

The time component of each complaint is separated into two string columns (CMPLNT_FR_DT and CMPLNT_FR_TM) as the first step after the download of the dataset is the merge of these two columns in one Datetime column for easier management of the dates.

2.3 Feature selection

The data contains the geospatial information of each complaint, date, hour, suspect, and other fields used internally by the department. For this purpose only the location, date and hour of each complaint will be used.

3. Exploratory Data Analysis

3.1 Relationship between the time of day and location of the complaints

First is the creation of 3 time windows to analyze the variations produced between the occurrence of the events

Time Windows 1: 00:00 to 07:59
Time Windows 2: 08:00 to 15:59
Time Windows 3: 16:00 to 23:59

After the dataset is filtered and the points are plotted as heat maps we can observe differences in the intensity and location of the hotspots.

Figure 1. Complaints generated during time window 1

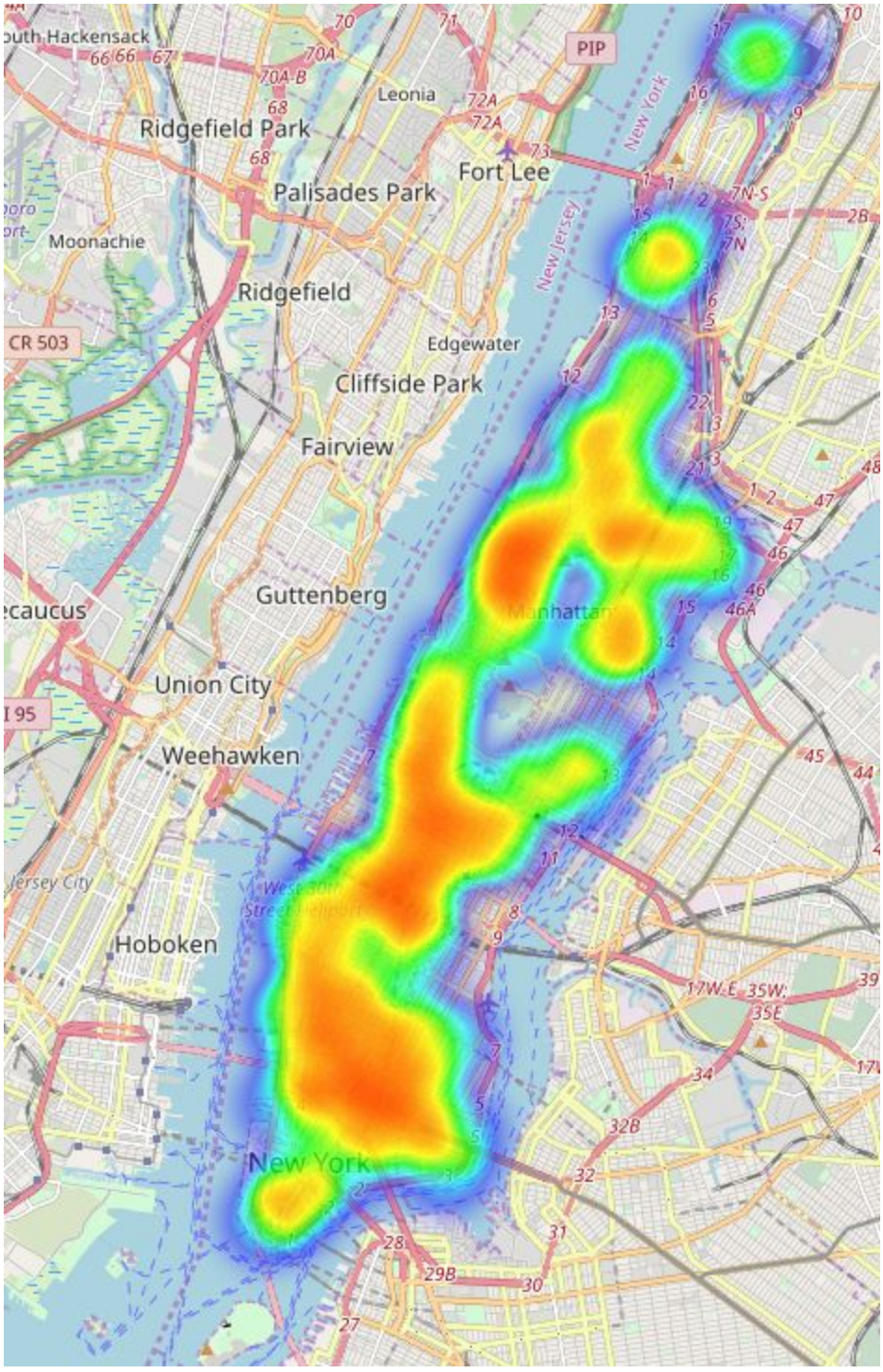

Figure 2. Complaints generated during time window 2

Figure 3. Complaints generated during time window 3

3.2 Relationship between venue category and event occurrence

To determine if a specific venue category influence the rise of criminal complaints, with the clusters made i downloaded the top ten popular venues around the zones with the highest criminal activity.

	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	40.739274	-73.997819	Rubin Museum of Art	40.740110	-73.997807	Art Museum
1	40.739274	-73.997819	Bar B	40.739471	-73.999074	Tapas Restaurant
2	40.739274	-73.997819	Coppelia	40.738819	-73.999908	Cuban Restaurant
3	40.739274	-73.997819	Yanni's Coffee	40.739768	-73.998884	Coffee Shop
4	40.739274	-73.997819	da Umberto	40.739473	-73.995864	Italian Restaurant

Figure 4. Venues dataset

Then the common venues categories are grouped to get the count of venues of the same type.

	Venue Category	Venue
32	Italian Restaurant	5
28	Gym	5
53	Theater	5
43	Pizza Place	5
14	Coffee Shop	5
17	Cuban Restaurant	3
18	Dance Studio	3
5	Bakery	3
6	Bar	3
50	Taco Place	3

Figure 5. Venues count by category

However as the venues in Manhattan are so diverse is difficult to generate conclusions about this relationship.

4. Geospatial Data Cluster Model

A KMeans cluster model is used in this investigation to help the grouping of similar events based on the longitude and latitude of each point.

For this purpose, 10 clusters were created. This way we can make conclusions using these groups as a reference.

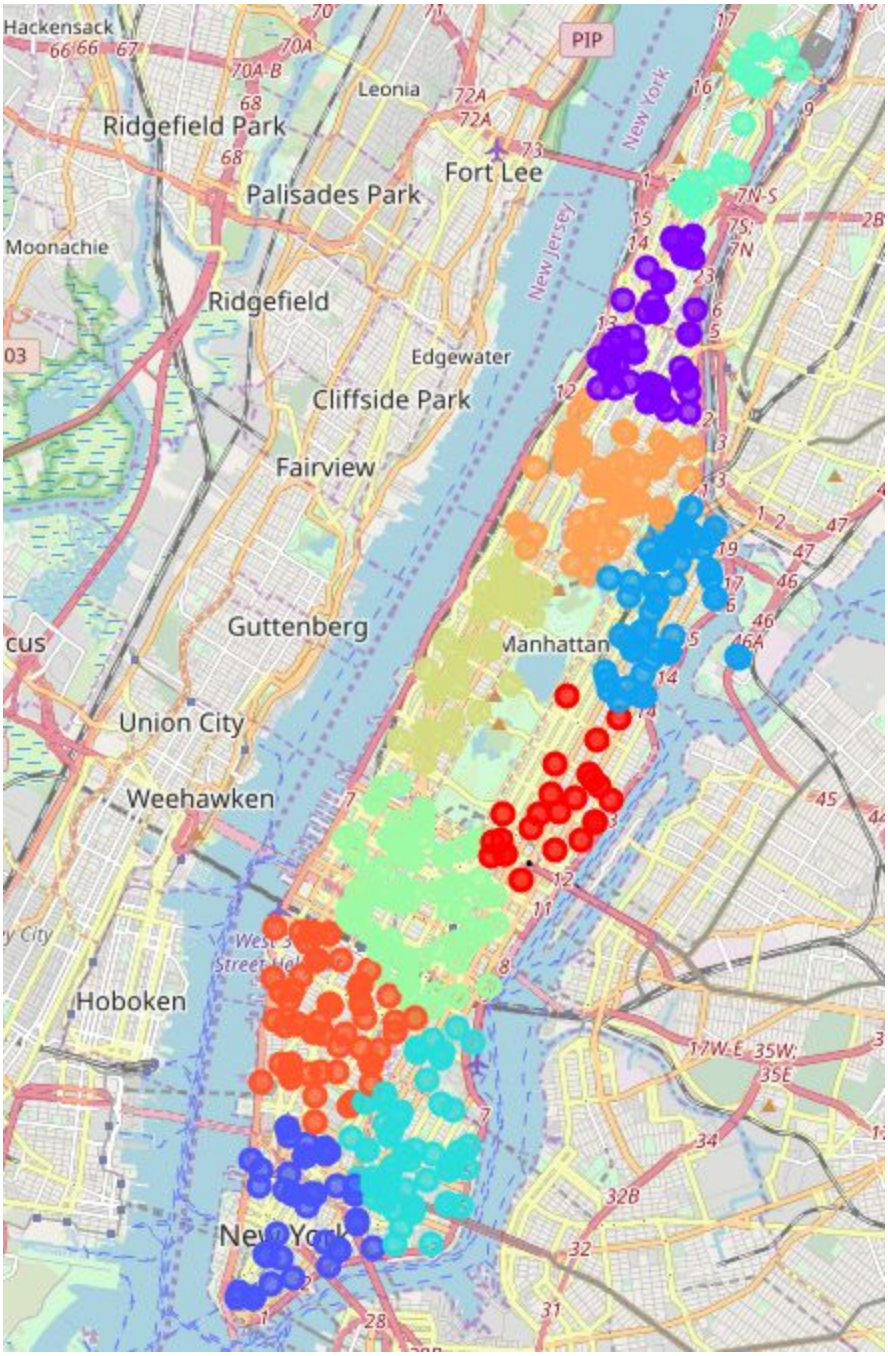

Figure 6. Complaints Clusters

Then I calculated the incidents rate by clusters to determine if an area is safer than other region in Manhattan

	Incidents_Rate	Cluster	Longitude	Latitude
0	0.109279	9	-73.997819	40.739274
1	0.083239	1	-73.942302	40.829724
2	0.070389	7	-73.974556	40.786797
3	0.193768	6	-73.984566	40.756282
4	0.108597	8	-73.951042	40.810280
5	0.102456	4	-73.984331	40.725526
6	0.119627	3	-73.940346	40.795409
7	0.057426	5	-73.929363	40.856558
8	0.091199	2	-74.004067	40.714472
9	0.064021	0	-73.958713	40.770263

Figure 7. Incidents Rate Table

Note that some clusters have higher incidents rate than other clusters

5. Conclusions

With the analysis of the complaints reported to the NYPD i can see a relationship between several factors like the time of the day of the occurrence and the location of the complaints generated.

However the higher diversity of venues in the borough of Manhattan makes difficult the binding of venues types and complaints reported.

6. Future directions

With the use of more information about the complaint (sex, gender, racial group) out of the scope of the exercise harder relationships may be found.