Evaluating Vulnerable Areas of New York City: Data

Jack Breingan

25/08/2020

1 Introduction

New York City is both the most populated, and most densely populated city in the world, with more than 8.3 million inhabitants. It is also one of the most expensive places to live in the world, particularly the Manhattan burough, and between these two factors it has a significant population of vulnerable people - whether that is economically, or to crime, or various other reasons. These populations tend to be clustered into neighbourhoods where the problem is prominent among the people there, and our goal will be to identify these vulnerable areas, and those at risk of becomming one. Our analysis will be on reported crime data from the year 2019.

In this project, we will attempt to uncover the areas of NYC most vulnerable in terms of crime, and the locations most affected by it using Foursquare location data and the City of New York's own cityofnewyork open data project (`https://data.cityofnewyork.us`). The results may be of interest to local governement, who allocate funding towards reducing the impact and level of crime, or the NYPD, who allocate funding and personel based on crime statistics. It may also be relevant to anyone thinking of opening a business in NYC, by showing them potentially hazardous areas of the city to set up shop in, and locating those that are less risky.

2 Data

The data in this NYC Open Data (`https://data.cityofnewyork.us`). Specifically, we used the NYPD Complaint Data found here `https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i` for the bulk of our data, and the geospatial data on New York's neighbourhoods found here `https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23` for our spatial analysis.

We have used the most recent data available at the time of writing, August 2020, and confined ourselves to a one year period - that is, we have used data from 2019. Crime data changes significantly year by year, as police and other

agencies act to suppress it and offenders concentrate around new methods, so earlier data is not likely to be as relevant, although certainly some kinds of crime, particularly domestic, are likely to remain in similar places with similar trends.

Our location data contained the name of each neighbourhood in NYC, the corresponding borough it was part of, and the latitude/longitude coordinates of the neighbourhood.

The crime data provided locations, dates, times and types of crimes. This data contained the following useful information:

1. CMPLNT_NUM: Randomly generated persistent ID for each complaint
2. CMPLNT_FR_DT: Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
3. CMPLNT_TO_DT: Ending date of occurrence for the reported event
4. CMPLNT_FR_TM: Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
5. OFNS_DESC: Description of Offense
6. PD_DESC: More granular description of Offense
7. CRM_ATPT_CPTD_CD: Indicates whether the crime was successfully completed
8. LAW_CAT_CD: Level of offense: felony, misdemeanor or violation
9. LOC_OF_OCCUR_DESC: Where a crime occurred in relation to the premises - in front of, behind, etc
10. BORO_NM: The name of the borough in which the incident occurred
11. JURISDICTION_CODE: Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port Authority, etc.
12. ADDR_PCT_CD: The precinct in which the incident occurred
13. PREM_TYP_DESC: Specific description of premises; grocery store, residence, street, etc.
14. VIC_AGE_GROUP: Victims Age Group
15. VIC_RACE: Victims Race Description
16. VIC_SEX: Victims Sex Description
17. Latitude of incident
18. Longitude of incident

We combined this with our location data, calculating which neighbourhood centre was closest to each crime, to associate each incident with a neighbourhood, then evaluated crimes on a neighbourhood level. We also examine some trends across the city as a whole. Using Foursquare, we examined the kind of locations prominent in these neighbourhoods, to get an idea of the most at-risk types of premises, and whether the crimes reflected the general trend of locations present or disproportionatley target something specific. This data allowed us to make profiles of the types of crimes that affect each neighbourhood, and also who they targeted. We were also able to make observations about the times crimes were likely to occur.

Using this data, we outline the most vulnerable to crime neighbourhoods in New York, and the types of crime which are prominent there. This data provides advice for anyone looking to handle these offences, whether government, law enforcement, or private businesses considering new locations.

3 Methodology

3.1 Data Preparation

In this project, we created a record of crimes by neighbourhood, and across the whole city, to get a view of incidents at both levels and so that we could track trends across the city, and by individual neighbourhoods. This required labelling our city-wide data with its corresponding neighbourhood. We used sklearn's Ball Tree algorithm to combine our neighbourhood data with our complaint data in order to determine which neighbourhood each incident occurred in. Ball Trees are useful for nearest neighbour problems like this, and are similar to decision trees with hyperspheres as leaves and containing distance measures between points - they allow us to compute nearest neighbours without having to compare every point to every other point in the data, which would be computationally prohibitive. This also allowed us to use the Haversine metric (or formula), which determines distances on a curved surface, making our measurement more accurate than a Euclidean metric would allow. This allowed us to find the closest neighbourhood to each incident.

3.1.1 City View

The city view was based on the NYPD Complaint Data. We created several new features to assist our clustering algorithm and analysis.

In this section, we analysed the rate of crimes across the city, and identified where offences were likely to take place, who they affected, and how long they took to resolve. These were plotted and analysed, see Results 4.1. The main part of this preparation was identifying data that was of no use and removing it. This included duplicates, rows with missing values, rows with records in the wrong columns, and the like. As we had plenty of data, we simply dropped this kind of row rather than correcting it, leaving us a total of 322812 valid rows for our analysis.

3.1.2 Neighbourhood View

To create the neighbourhood view, we aggregated our city data, getting the count of incidents in each neighbourhood, the mean duration of those incidents, and the count of each of the three categories of incident - Felony, Misdemeanour or Violation. We also counted how many times an incident was handled by the police, housing authorities or another external agency. While data on which specific external agency handled each case, it was considered superfluous for our purposes, as we were most interested in police incidents. As such, we considered all such agencies 'External'.

We used FourSquare to retrieve informations on venue types within each neighbourhood. A venue was considered to be 'within' a neighbourhood if its latitude and longitude were within 500m of the central latitude/longitude coordinates of the neighbourhood. This method is likely to have overshot in some cases, and undershot in others, but we lacked data on the extent of NYC neighbourhoods and the metric was adopted as a reasonable compromise between misssing and mis-assigning venues. The venues were aggregated, and the top 5, i.e. the 5 most common, venues per neighbourhood were appended to our neighbourhood data.

We used these dataframes to analyse crime rates accross the city. Our analysis, and the conclusions thereof, are presented in the next section.

4 Results

4.1 City View

We first examined the distribution of incidents across the city. We find that there are more incidents in Manhattan than any other district, and that the neighbourhood reporting the most incidents overall is Central Harlem. The breakdown of incidents by borough can be found in Figure 1.

Examining the number of incidents by neighbourhood, as shown in Figure 2, we can see that most are in Mahattan, or the Bronx. This makes sense, considering the extremely high population density and income disparity present in that area of the city, and in line with that we see that the lower income areas of Manhattan in the north of the borough have higher crime rates than the wealthier south. When we also examine Brooklyn, Queens and Staten Island, we continue to see evidence of this trend - the higher income areas of the city (Manhattan, Staten Island and Queens have average incomes above the $64,000 average for the city as a whole, though the former two are far above it and Queens is close) experience lower incident rates than poorer areas (Brooklyn and the Bronx have average incomes below the city average, particularly in the Bronx where incomes are approx. 60% of the city average).

Examining the areas of the city with the lowest reported incidents, see Figure 3 we see the that they are in higher income areas in keeping with our abve conclusions.

Figure 1: Count of incidents by Borough

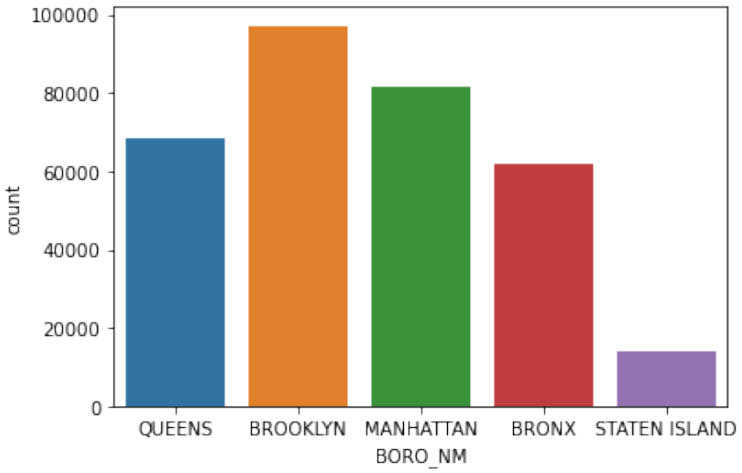


Figure 2: Top 10 Neighbourhoods by Incident Count

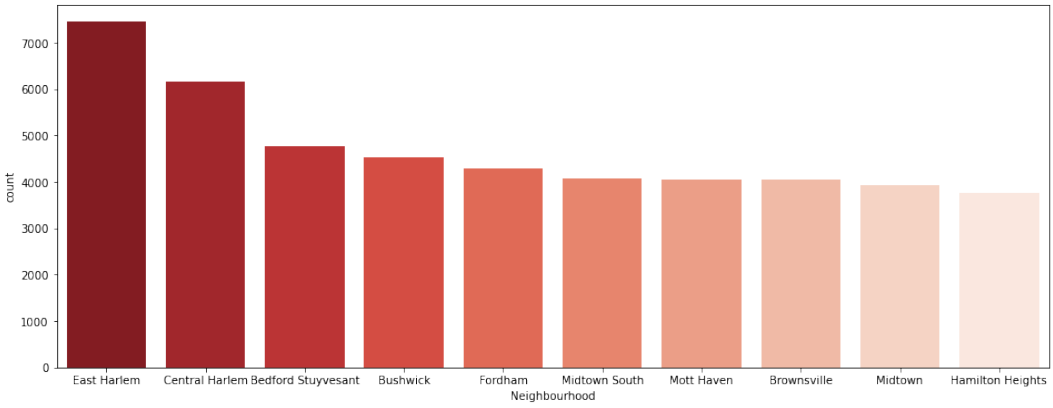
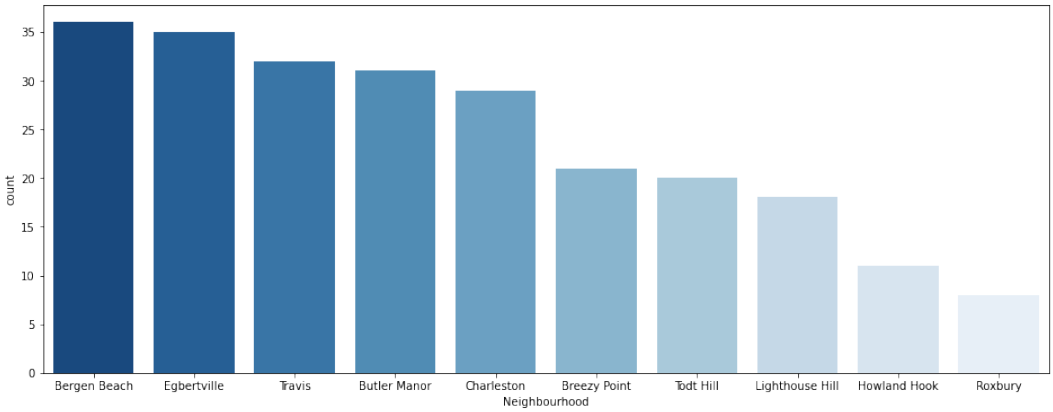


Figure 3: 10 Lowest Incident Count Neighbourhoods



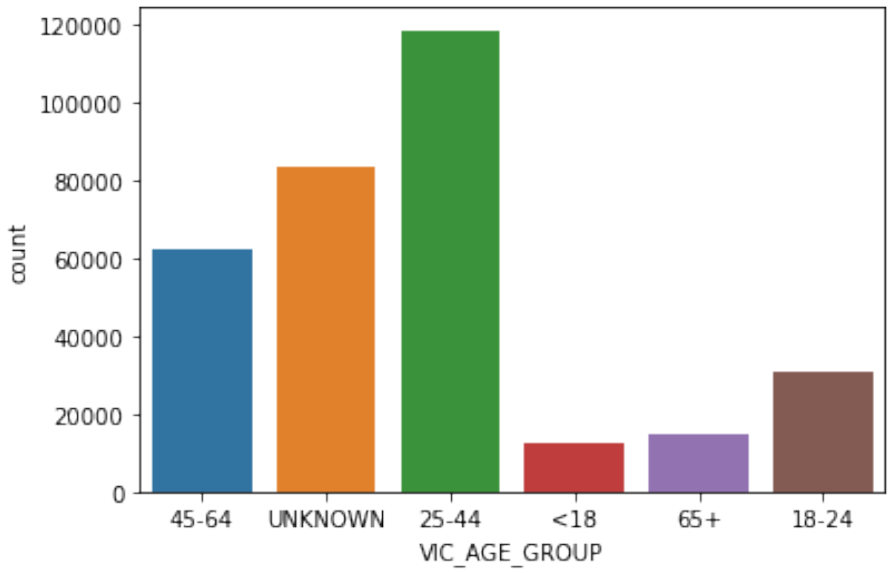
From this, we conclude that a major issue facing New York City is income disparity, and suggest that measures to reduce the crime rates in these areas take that into account - this might mean programs of community enrichment, repairing and enchancing local buildings for example, investment in local businesses, or consideration of improvements to existing welfare systems.

We would like to see who is most affected by crime in New York City, so we have counted incidents among age and race groups, as well as determining the distribution of crimes by age and race. The first, age groups, can be seen in Figure 4. In this figure, we can see that the majority of victims are 25-44, with 45-64 the next most at-risk category. It should be noted that a substantial number of victim's ages were unknown. This represent both crimes without a singular victim, and where ages were not recorded, and we must keep in mind that the graph may be significantly affected by this.

We then note that the most common racial group among victims are Black, followed by White Hispanic and then White, see Figure 5. This is disproportionate with the population demographics of New York City, and so we can conclude that Black and White Hispanic individuals are at higher risk than white. Again, we must keep in mind the high number of unknowns when drawing conclusions, but in this case our conclusions would not change even if all unknowns were assigned to any single other group, so we can draw this conclusion with some confidence.

Next, when we examine the violin plot in Figure 6 we notice some trends. White victims are more likely to be elderly, and less likely to be below 18 years old, while the opposite is true of White Hispanic, Black Hispanic and Black victims, who are more likely to be younger. Most notably, victims below 18 are most common in these groups. American Indian/Alaskan Native victims display a unique trend, with few elderly or young victims, and a more even distribution of victims in the 18-64 range. We can conclude certain groups are at higher risk

Figure 4: Count of Incidents by Victim Age Group



than others, and note that while in general victims are more likely to be elderly, Black, Black Hispanic and White Hispanic youths are at higher risk than others.

By considering where crimes occur, we may identify high risk businesses and areas. We have access to information on where crimes occurred (Figure 7), as well as whether they were indoors or outdoors (Figure 8). We can see that incidents usually occur indoors, and that in particular they occur in residencies. Incidents also frequently occur in the streets. Certain types of stores are also common incident locations, with chain stores, gorcery stores, drug stores and commercial buildings accounting for the majority of remaining incidents.

Finally, we consider how long incidents typically take to resolve. Figure 9 contains a box plot of incident durations accross the city. We see that the longest durations are in Manhattan. Staten Island, Queens and Brooklyn display no appreciable difference in duration. Figure 10 shows us the city-wide distribution of incident durations in the case of incidents taking more than one day. Incidents taking 1 day or less to resolve are by far the most common, and were excluded from the figure so that the pattern of durations could be percieved without those counts reducing it to indistinguishability. We observe that the vast majority of incidents are of very short duration, and see an approximately exponential fall off of durations. This implies that most crimes are simply resolved, or that no further action was required, as most complaints are not serious issues.

Figure 5: Count of Incidents by Victim Race Group

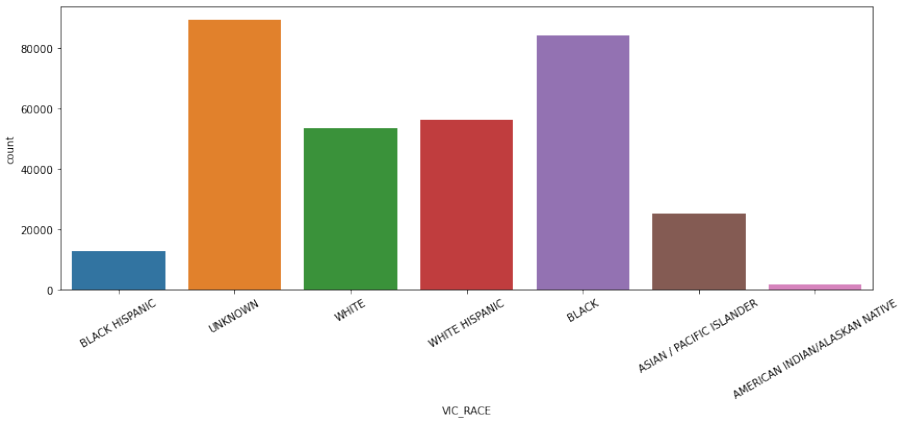


Figure 6: Distribution of Incidents by Victim Age and Race Groups

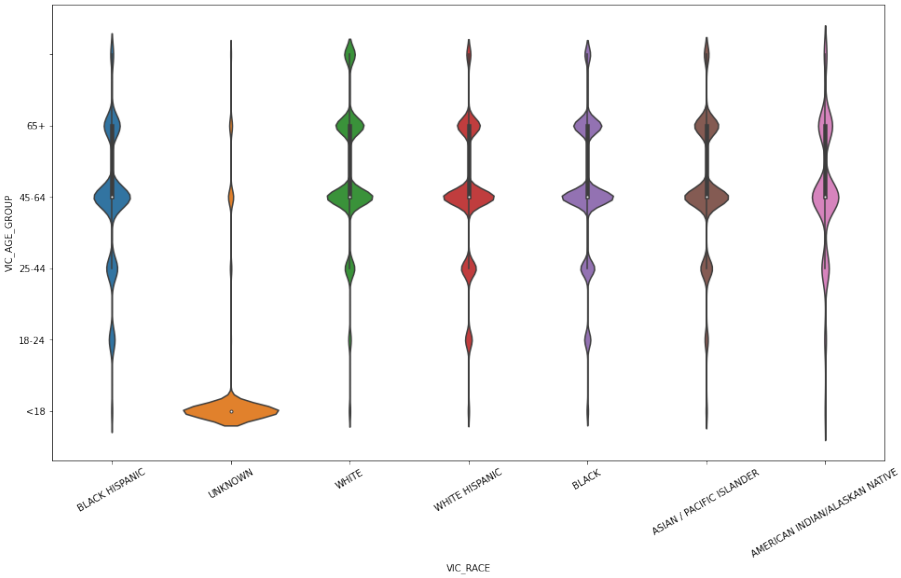


Figure 7: Distribution of Incidents by Location

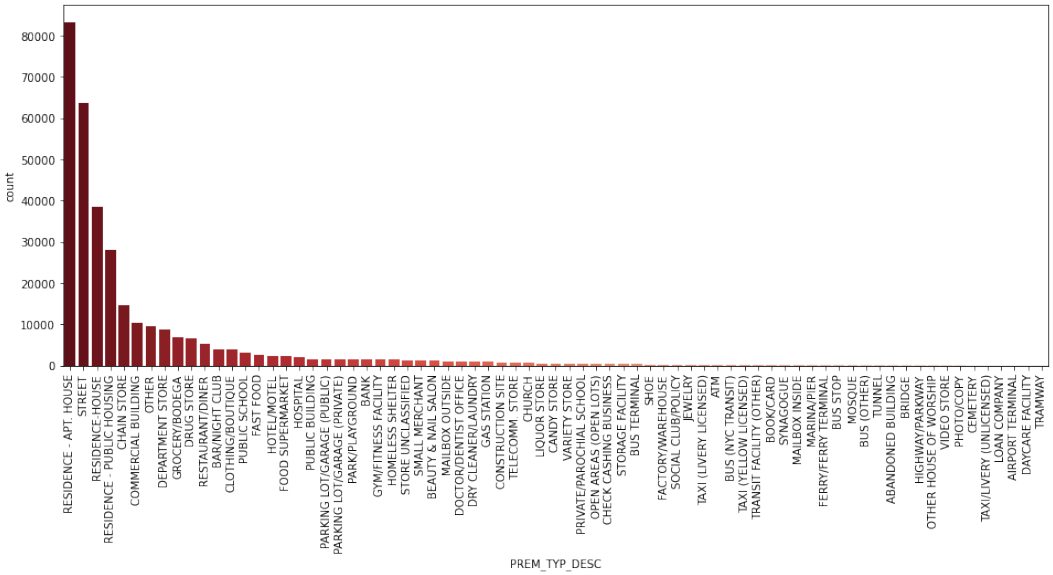


Figure 8: Distribution of Incidents by Victim Age and Race Groups

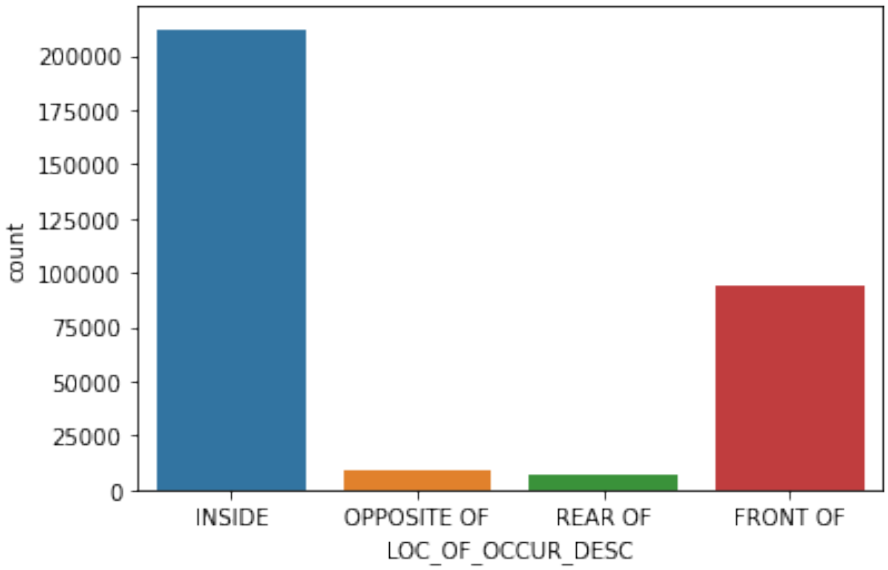


Figure 9: Mean Incident Duration by Borough. The plot is truncated at 50 days because while there are outliers above that level, they are few and obscure the differences in distributions.

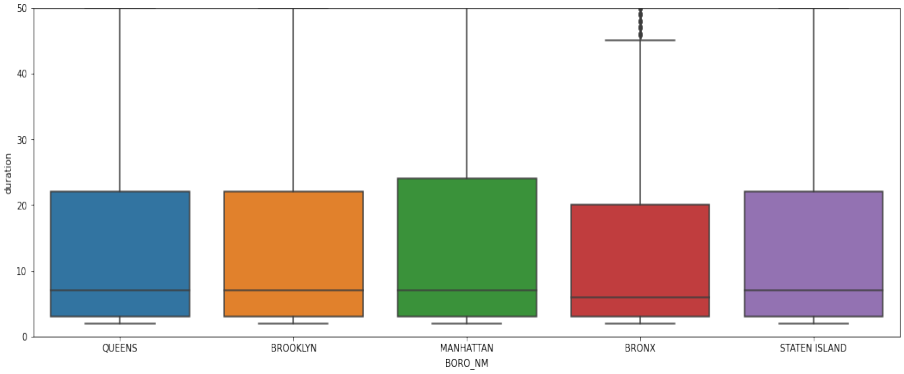
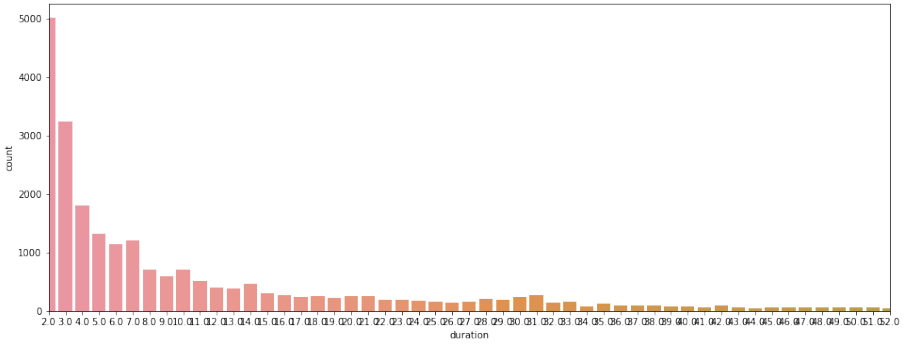


Figure 10: Mean Incident Duration across NYC. 0 and 1 day durations are not included as they dominate the plot completely. 52+ day incidents were excluded as outliers due to their rarity.



4.2 Neighbourhood View

We then clustered neighbourhoods according to their similarity, making use of location, incident count, average durations and the types of incidents present. We can see the results of our clustering algorithm in Figure 11. The method we used was DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which as the name suggests groups points into clusters based on their density, rather than on their distance from a pre-defined cluster node. This allows the algorithm to be used in cases where the total number of clusters is unknown, making it suitable for our project. The main considerations for this process are the spacing of points and the number of points that formed a cluster. The algorithm randomly selects a point, then looks for any others within that spacing distance around it. If it finds none, the point is labelled an outlier, and otherwise the two are considered a cluster. Points may also become outliers if between them there are not enough to match or exceed the minimum cluster size.

We used a spacing of 0.82, and a minimum sample size of 3 for our clusters, which were found to give the best results. Our data was found to contain 4 clusters, and a number of outliers. Two of these clusters were areas where incidents seemed to be handled in an above average manner compared to their surroundings - they had low incident counts, incidents were resolved quickly, and they saw a lower ratio of felonies. The vast majority of neighbourhoods fell into a cluster we identified as representing our standard neighbourhoods - those with typical properties for thier location. Finally in our clusters, we identified a group of neighbourhoods in central Manhattan with high incident_counts. This cluster makes sense as this is an area of the city popular with tourists, containing landmarks like the Empire State building and Times Square. Such areas usually see high rates of petty crime, and elevated rates of more serious incidents.

As the outliers described neighbourhoods with some significant difference to their neighbours, they were of interest to us and we did not discard them. Instead, outliers with an incident count below the average of normal neighbourhoods were concidered to be low-incident neighbourhoods, and above that were considered high-incident neighbourhoods.

5 Discussion

Based on our results we have identified a number of areas of concern. We find a cluster of neighbourhoods in central Manhattan with high crime rates, are able to identify some areas of low incident counts, and were able to detect outliers among the city's incident pattern to identify both neighbourhoods which are doing well, and those that require additional assistance. We have identified the neighbourhoods with the most serious issues, and those with the least. We were able to determine how age and race groups related to who is targeted in these incidents.

Based on our results, we can reccomend that steps be taken to address

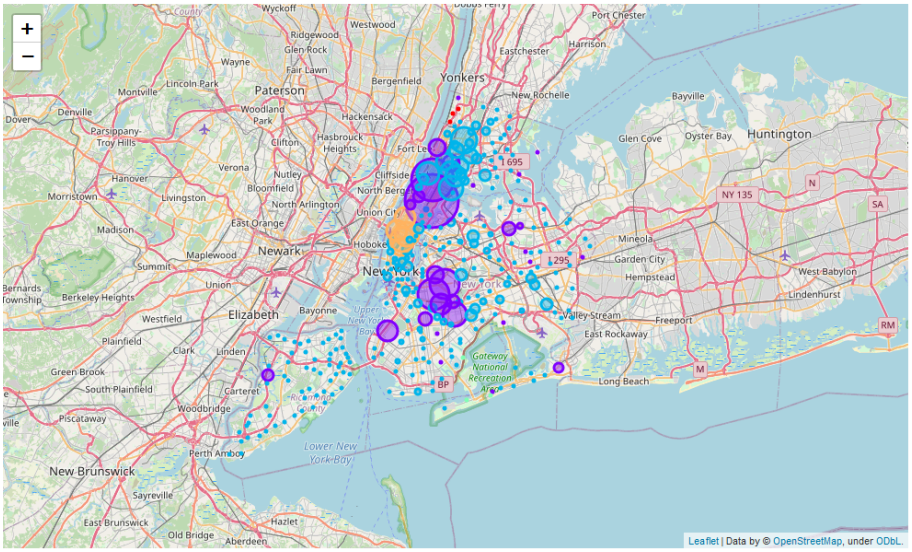


Figure 11: Clustered Neighbourhoods in NYC

Cluster Legend

- Outliers
- Typical Neighbourhoods (Cluster 0)
- Low Incident Neighbourhoods North (Cluster 1)
- High Incident Midtown Neighbourhoods (Cluster 2)
- Low Incident Neighbourhoods South (Cluster 3)

income inequality in the city through local and city-wide initiatives. We reccomend that immediate action be taken in the 10 identified neighbourhoods in Figure 2 with the highest crime rates especially. Understanding specifically how to help these areas requires further exploration and access to more data, but reasonable suggestions might include assistance for local businesses, improved police and local authority funding, and improved welfare in low income areas. For some of these areas, safety advice for tourists might also be useful.

Residential crime is high, with homes and residencies eclipse all other areas in terms of incidents reported. In these cases, raising safety awareness with residents, including advice on how to recognise common hidden crimes like domestic violence and of access to help for these issues, may bring some benefit. Investment in local communities could also help to reduce the incident rate, in keeping with our findings that low income areas are disproportionately affected by crime.

While we identified that elderly members of some groups are at risk, and that there are clear differences between racial groups, it is impossible to make any reccomendations based on victim information from the data available to us. Follow up on this issue should consider who the offenders are as well, in addition to requiring more information on both parties.

Our results could have been improved in a number of ways. Access to more detailed data on the city's neighbourhoods such as population demographics, average income, police prescence and funding, funding granted to the neighbourhood and so forth would have allowed a much more detailed analysis of the causes of crime in New York City. With more time, a more thorough statistical analysis could have been carried out, and that will comprise a future project. There is scope to expand this project into a comparison of crime in cities across the USA, and similarly a comparison between cities in different countries could produce interesting results. In the latter case especially, the different approaches each country has to addressing these issues could provide an interesting comparison of their effectiveness, but such an analysis must not ignore cultural differences and their role in all aspects of life before advising the implementation of anything derived from that process in another country - look to Walmart's disastrous attempt to enter the German market while acting like a US company would at home for an example of the risks here.

We would like information on the offenders as well as the victims for further analysis. This would allow us to compare and contrast the groups, as well as identify common problems behind these incidents. Reccomendations made with access to such data would be far more useful.

From these results, a reasonable follow-up might be examining complaint data over a period of several years, which would allow us to identify neighbourhoods which are improving, and those which are increasingly at risk. This may allow us to develop an understanding of which methods work to reduce crime, and help people who need it.

6 Conclusion

In this project we examined the issue of crime in New York City. We analysed the common types of incidents, and looked at them in terms of location, victim profile, location and incident duration. We were able to cluster neighbourhoods by incident details, and find several distinct clusters of neighbourhoods with common issues. Based on these clusters, we identified the areas of most concern in the city, and looked at the kinds of incidents present in them. We made a series of reccomendations based on the data, in the hopes that they might result in reductions in these incidents. Finally we made suggestions for several follow up projects that could expand on the results of this project, and develop a better understanding of these issues, and their underlying causes.