Investigation of Crimes in the United States

By Brandon Mehlenbacher

March 28, 2020

Introduction

Background:

Crime is inevitable in almost any location regardless of whether it is in a big city or in a small city. Crimes can be mitigated by knowing potential indicators for cities that are nearing statuses that would cause a substantial increase in crime.

New York City (NYC) is the largest city in the United States (US) by population with 8 million people. However, the area that NYC expands over is only 784 km^2. In comparison, the second largest city in the US is Los Angeles with 4 million people and an area of 1300 km^2. Yet, the homicide rate in Los Angeles is higher than it is in NYC. It would be interesting to see whether the higher crime rate is universally spread across the city or is it localized to smaller areas due to other factors.

Business Problem:

Being able to determine factors that cause people to want or feel the need to pursue criminal acts is very useful in being able to help areas that are on the verge. Locations that have already passed the threshold are difficult to help because it becomes ingrained in the people a way of life. Preventing areas from reaching this point would attack the problem early.

Interest:

Crime prevention agencies would be interested in knowing these indicators as they can deploy their units effectively. People who are planning to move to new cities would also benefit from knowing this information since it is rare for travelers to know much about a city before they have lived there.

Data Collection:

Data on the crimes in New York City will be necessary for determining the locations of the highest density of crimes. Other data would include costs of apartments which can be found on street easy website, number of venues neighborhoods have to offer which would be acquired using the Foursquare API.

Data Cleaning:

There were some issues with the data when initially starting out. The biggest issue was connecting boroughs and neighborhoods with specific incidents of crime. There may crime related incidents that occur in several different neighborhoods as a result of chasing down the criminal or simply being on the boundary line between two neighborhoods. This meant separating crimes into individual neighborhoods was not simple. The way I handled this issue was to connect any crime that occurred in multiple neighborhoods to each of those respective neighborhoods.

Another issue that came up was lack of information on apartments for Staten Island for each individual neighborhood. This meant that connecting the information across all of the different neighborhoods between different boroughs was not able to be done. The way to get around this was to simply compare the overall boroughs to each other during the initial stages of the project to see whether there were any outstanding features that could potentially account for the different cumulative crime rates.

The last issue was getting reliable crime related data. There were few readily available sources on the crimes of each individual borough and even more difficult each individual neighborhood. Some referenced latitude and longitude values but this would be tricky to define as there are many redundant city names and correlating it to the original state would take several many days of work to sort out all of the crimes that take place in the country. To take care of that, I focused on only gun related homicides for this analysis.

Feature Selection:

There are many different features that can be related to a city including population density, population, number of venues, and cost of the different apartments. I chose to focus on the number of different venues in each neighborhood and the cost of different apartments for a few reasons.

The number of different venues an area has to offer is indicative of how much activity is going on in that area. A larger number of venues will indicate that neighborhood has more to offer for a incoming resident and will also specify where a policy force will need to focus most of their resources. It also indicates how desirable a location for real estate opportunities such as a new business owner coming into NYC and wanting to find the right place for a business. If there are very few businesses in an area that may indicate it is not a profitable location to be in.

As for the cost of different apartments, this is more useful for apartment seekers and business owners. People looking for apartments in one of the most expensive cities in the world want to find cheap and safe options that will also fulfill their outgoing desires such as having a good bar scene. For business owners it helps they tailor their business to their perspective clients. If a owner wants to start a high class restaurant, it may not be the smartest move to start out in the middle of a lower income area as fewer people are going to want to go to eat their minimizing their profits. On the other hand, a start up company may be looking for a good cheap office location and the cost of apartments in the area can be an indicator.

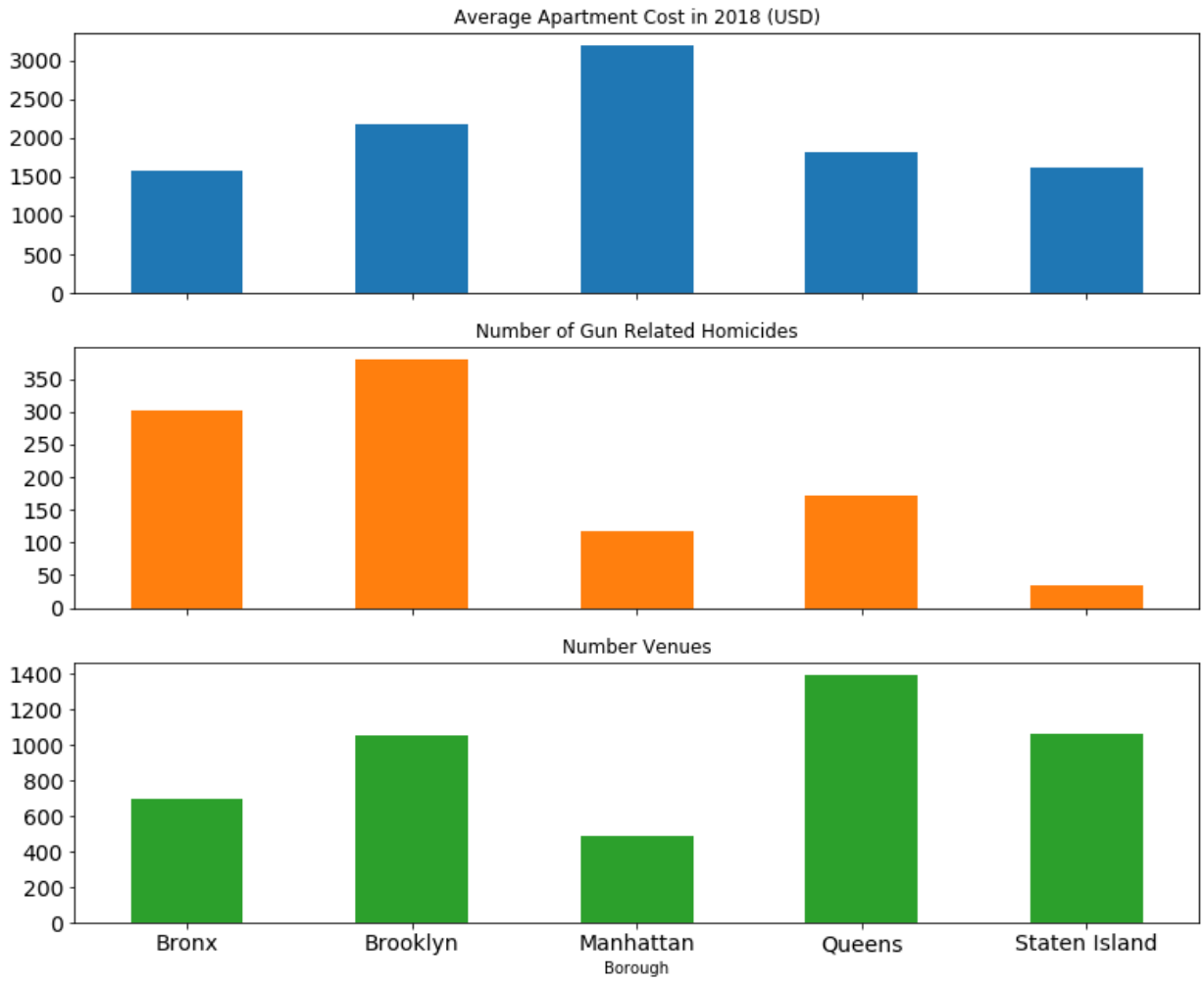

Figure 1 An overview of the features in the different boroughs

Exploratory Data Analysis:

Differences Across Each Borough:

The first initial insight into the data was to look at the boroughs of NYC. There are five distinct boroughs: Manhattan, Staten Island, Bronx, Queens, and Brooklyn. Figure 1 shows a three bar graphs each illustrating the different features for the boroughs. The first thing to note is the differences in apartment costs. It is very clear that Manhattan has a much higher average apartment cost than the other four boroughs. The lowest apartment cost is the Bronx. The number of gun related homicides is also the lowest in Manhattan as well most likely due to the much higher cost of living. The interesting thing to note though is that where the Bronx has the lowest apartment cost, it doesn't have the highest rate of homicides; Brooklyn has a higher number of homicides. This indicates that apartment cost is not the sole predictor for the data.

The number of venues is also plotted in figure 1. In this case, there is also the fewest number of venues in Manhattan. This is most likely caused by the much higher cost of real estate

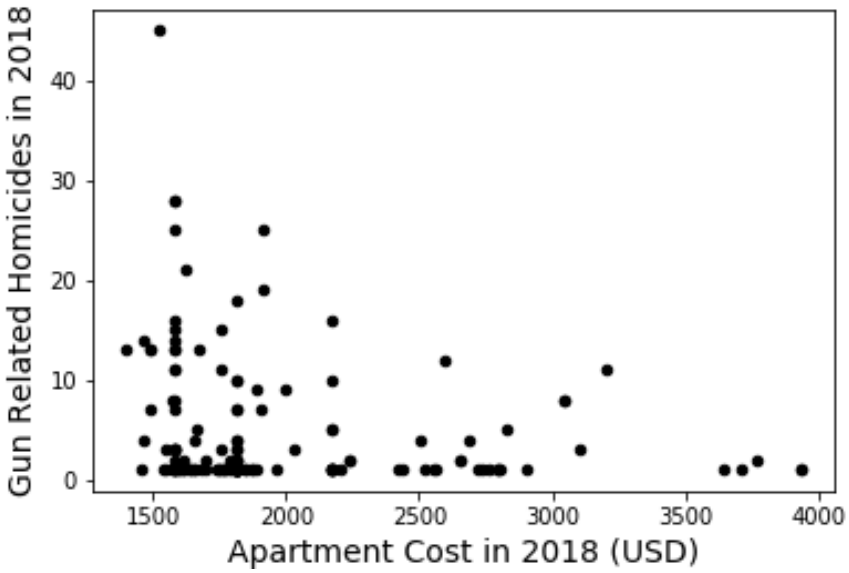

Figure 2 The average cost of an apartment versus the number of homicides in each neighborhood.

therefore making it more difficult to start a company there. On the other hand, Queens has the largest number of venues out of the boroughs. This is possibly due to the fewer number of crimes that occur in the area compared with the Bronx where there are a lot of crimes that occur.

Comparison of Different Neighborhoods, Apartment Cost versus Homicides

Like how there are different boroughs, there are also different neighborhoods that make up each of the boroughs. Each individual neighborhood may have their own unique features that are not accounted for in the overall different. Take for instance the Bronx, there will be some

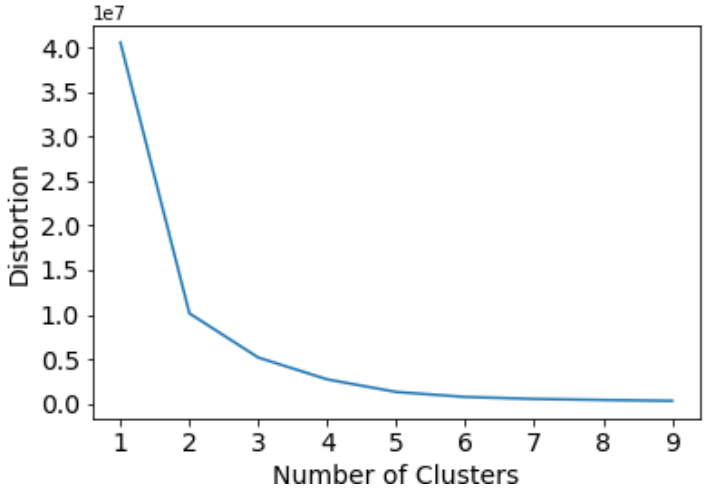

Figure 3 An graph showing the elbow method approach to determining the optimal number of clusters. The Idea behind the approach is to find the point at which the change in the distortion is considerably smaller than the change between the previous number of clusters.

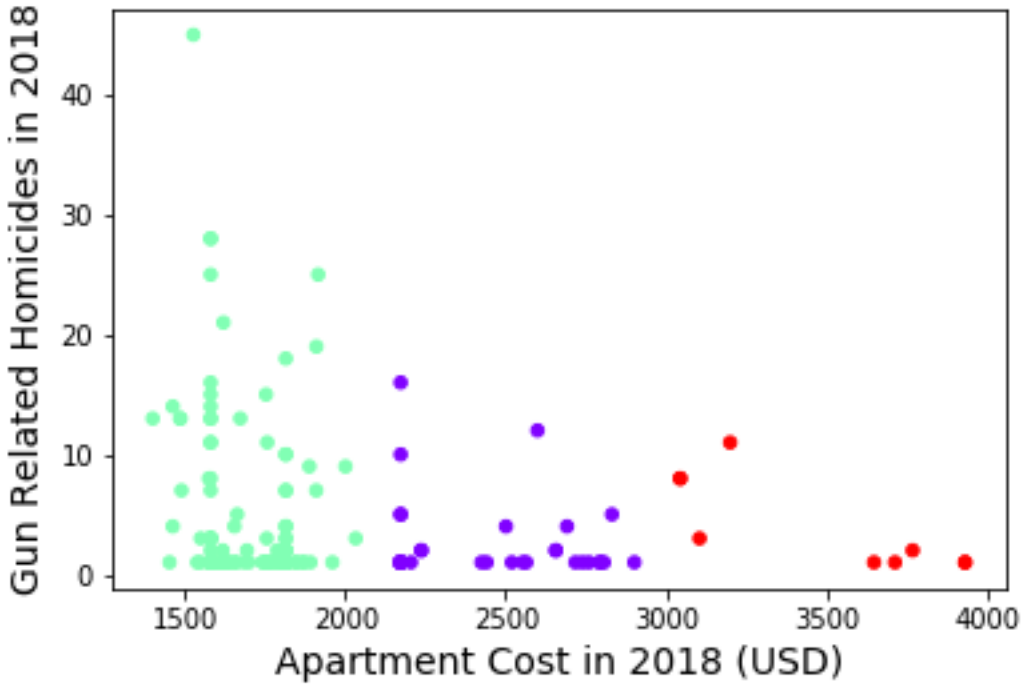

Figure 4 The different clusters that were formed upon applying the KMeans algorithm. There are three separate groupings, that were found. There is a grouping where the apartment cost is very high and in turn the crime is low. There is a second cluster where the apartment cost is moderate but the crimes are also moderate slightly more dangerous but safer in its own right. The last cluster is the lowest apartment cost. This is characterized with a large range of whether the area is dangerous.

neighborhoods that are expensive, and this may account for slight variances in the number of crimes that are committed.

Figure 2 shows a scatter plot of the average apartment cost versus the number of homicides in the neighborhoods. The first thing that is worth noting is that there is a higher number of neighborhoods that have cheaper apartment costs. This makes sense as there are more people in a lower tier of wealth who rely on less luxurious apartments or in general just need cheaper places to live. We also see a slight correlation (correlation coefficient -0.21) between the cost of an apartment and number of homicides meaning the cheaper the apartment the more likely there is to be crime nearby.

I then used a clustering-based approach to investigate different clusters and how they relate to the two variables. I used a KMeans algorithm approach to investigate the different clusters that may form. In this approach, it is imperative to determine the ideal number of clusters to be used. To do so, I used the elbow method which is shown in figure 3. It was found in this case the optimal number was found to be 3 clusters.

Figure 4 shows the different clusters that were formed. There are three different clusters that formed all in different price ranges. The first having a very high apartment cost but a much lower overall amount of crime. This could be a result of just fewer people who are living in that

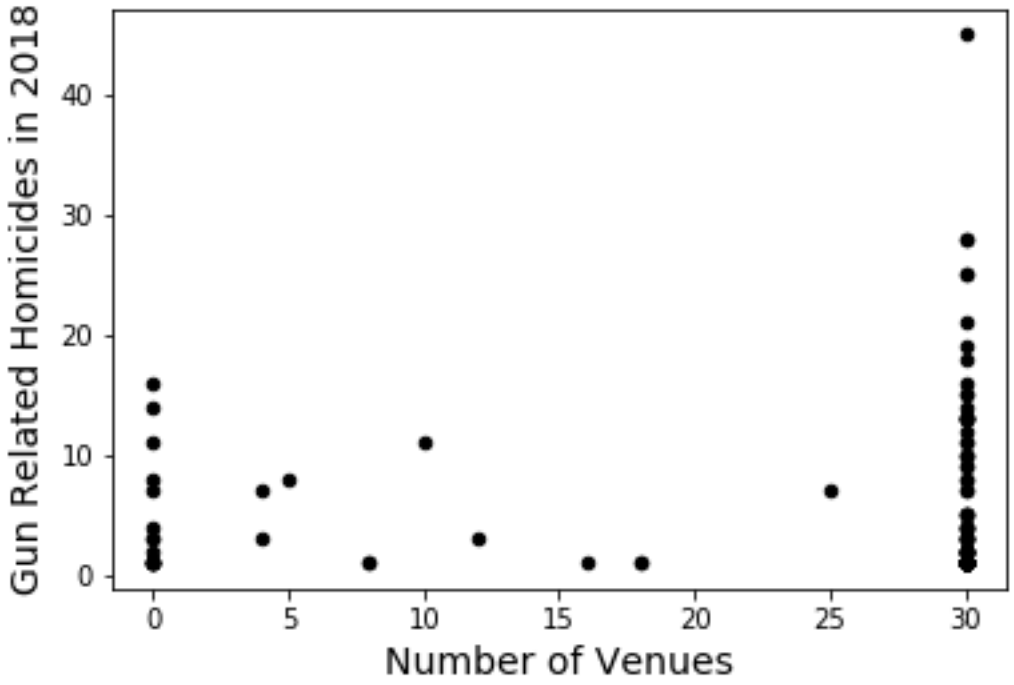

Figure 5 A plot showing the neighborhoods and their respective number of venues and gun related homicides.

area and therefore a smaller amount of people who would commit the crimes. Another plausible cause would be a higher police presence which would ultimately limit the amount of crime that would occur. The second cluster that formed is with a moderate apartment cost. This is characterized by a region where there is still a smaller amount of crime at a more reasonable cost. There are more crimes that occur than in the more expensive area but there are still many areas where the homicide amount is relatively low and could reasonably considered safe. This would be a good area for the middle class. The last cluster is the lowest apartment cost. Within this cluster there is a large range of housing options that may or may not be safe. It is characterized by the highest amount of crime that occurs in NYC. There are some areas that have had zero homicides that are recorded but that could just be due to the limited sample size used here (only looking at 2018).

Comparison of Different Neighborhoods, Venues versus Homicides

Venues offer a service to some of the areas and can create a high density of people so it is worthwhile looking at how the number of venues can affect the crimes in an area. I am going to use the same approach to explore the number of venues as I did for the apartment costs. In this case though the correlation between the variables is almost insignificant (correlation coefficient of 0.03). Figure 5 shows a scatter plot of the number of venues versus the number of crimes. It is worth noting the large number of data points around 30. Using the Foursquare API to explore the

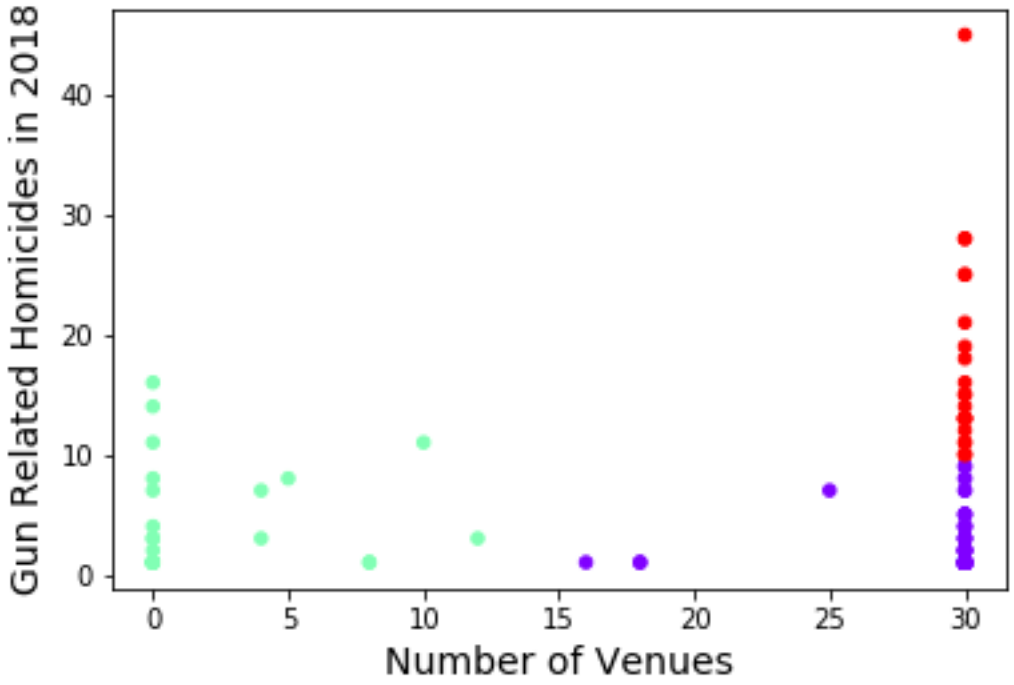

Figure 6 Clustering the number of venues and the gun related homicides. It was optimally found that three clusters fit the data best. The red cluster shows neighborhoods that have both high numbers of venues and high number of homicides. The purple cluster shows neighborhoods where there are high numbers of venues but also having low numbers of crimes. The cluster in green shows neighborhoods with low numbers of venues and a moderate number of incidents.

different area, it could be grouping some of the neighborhoods together due to proximity so there may be some neighborhoods with artificially higher number of venues. This does give a more accurate picture though as most venues cater to more neighborhoods than just one.

Figure 6 shows the clusters that were found when using the KMeans algorithm. The first cluster shown in red is showing the neighborhoods with high numbers of gun related homicides with high numbers of venues. This could be a result of lack of security in these locations. There is a weak police force in these areas and the businesses are geared toward low level incomes causing the incidents of aggression to be higher. These businesses could also more catering to the night life crowds of people since most crimes happen at night. The purple cluster shows areas with high numbers of venues with relatively few numbers of homicides. This contrasts with the previous cluster and could be related to a higher income class of people or just operating during safer hours of the day instead of during the night. The final green cluster us areas with few venues and a moderate amount of homicide incidents. This could be a result of small gathering of people in these areas or the fact that the area is dangerous and therefore business owners do not want to take the risks associated with the areas.

Conclusion:

In this study I investigated different attributes about the boroughs of NYC that would be indicative of higher rates of crime. It was found that looking at the price of an apartment is indicative of how high of crime activity there is in NYC and that finding a cheaper place to live will bring on a higher crime rate. In the case of venues, this was very weakly indicative of higher crime rates and although three separate groups were found, there would need to be further work done to prove this hypothesis.

Future Directions:

There are many other different variables that can be tested to see whether they are more correlated with higher crime rates. A few examples that could be looked at is population density, total population in the neighborhood, looking at the venues that are in the area. The search could also be expanded to other cities such as Los Angeles or Chicago which would also be rather interesting to see then how climate might play a role in the number of crimes or whether the same trends occur.

Sources for data:

https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

https://streeteasy.com/blog/data-dashboard/?agg=Total&metric=Inventory&type=Rentals&bedrooms=One%20Bedroom&property=Any%20Property%20Type&minDate=2010-01-01&maxDate=2020-02-01&area=NYC,Brooklyn

https://public.tableau.com/profile/yiqiao.li3563#!/vizhome/NeighborhoodShootingIncidencesin2018/NeighborhoodShootingIncidencesin2018