New York Police Department Complaints Data Analysis

1. Objective of the project

In the presentation slide, we stated that our project would explore, analyze and visualize the New York City Police Department complaints data, enriched by using spatial data. Our objective was to go in depth in the Visual Analytics methodology in all the stages of a project, taking all the decisions by ourselves and extracting some useful results. Also we wanted to use this project to learn more about the geo-visualization methods, since it is a very interesting area from our point of view.

The challenge of this project was to extract useful facts about crimes in New York, focusing on three parts. First, analysing the profiles of the victims and suspects (by clustering the data), getting an insight of the profiles more propense to suffer and commit an offense. Second, analysing the type of offenses and visualising the complaints distribution using filters, comparing the data against genders and other variables. And finally, visualizing maps showing how the complaints are distributed among zones, being able to identify which zones of New York are more dangerous, by adding the spatial components.

Therefore, at the end of this project we would be able to know how the suspects are and victims profiles, which crimes are most committed and other useful facts about crimes in New York that would come out while working on the project. The main opportunity of this project is to identify which areas are more dangerous, to know where more police should be placed, and who should be more guarded (potential suspects) or protected (potential victims) to guarantee citizens security.

2. Description of the project

In this project we wanted to perform a deep study of New York's complaint data from different viewpoints and using different methods. To do the analysis, we have followed the Analytics workflow seen in class.

After gathering the data, a Dataset Exploratory Data Analysis has been done for a better understanding of the data. In this step, we have analyzed the main characteristic of the variables, removed the non useful ones, managed the null values and corrected the datatypes when needed. In this way, the dataset was ready to perform further analysis.

Then, we plot some distributions to look for interesting information, such as the time distribution of crimes, the most common crimes and the stations with more crimes. Also we compare the crimes committed during the day and night, seeing that there are not very big differences among them. Also a special emphasis is placed on the violation crimes. All those analyses are described in the notebook itself.

After studying the data and once we had more knowledge about it, we implemented an unsupervised model. We performed a clustering regarding the characteristics of victims, suspects, and the type of complaints. By doing k-means after encoding the variables, we have added the cluster column to the dataset that we export in a csv for a better visualization using Tableau, extracting useful information about the profiles of victims and suspects, which was one of our goals.

Also, we tried to predict the race of suspects that were not identified by using GradientBoostingClassifier. The resulting accuracy of this model is not very high, however it is better to know the race of the suspect with a 61% probability than having no idea at all, since this information can be useful for the police. For a deeper understanding of this part, see the notebook.

The final part of the project takes advantage of the spatial variables of each of the complaints to implement a more profound analysis of our data, with the creation of some maps, even some of them animated to see the time passing of the complaints throughout one day's hours or the days of the year. The last part of this section changes the perspective of the project, focusing more on the neighborhoods rather than the complaints themselves, providing a deeper study of the spatial location of the crimes recorded in New York city. We create a second dashboard published in Tableau and finish up with some Carto maps visualizations that are really interesting and representative.

To end up the project, and gathering all the analysis and results we have obtained, we get to the conclusions of the project in the next part of this report, explaining them all together with the final overview.

3. Results, Visualizations and insights

In this first dashboard (https://public.tableau.com/profile/n.ria4479#!/vizhome/finalProject_16077875349710/Dashboard1?publish=yes), we can see the profiles of victims and suspects related to each type of complaint (felony, misdemeanor and violation).

In the first page there are two visualizations, one that shows the amount of suspects per race and age group, and another one that shows which races tend to attack other races. Both visualizations can be filtered by cluster or type of offense.
In the second page, there are two plots. One shows the victim races divided by the sex of the suspects and the other one shows the age of the victims depending on the races. Both can be filtered also by cluster and offense category.
The third page shows the three clusters with the corresponding ages of the victims and suspects.

From these visualizations, we can see that the age group with more suspects is 25-44 and the most prone race to commit crimes is the black one, followed by white hispanic. We also see that races tend to proportionally attack their own races, meaning that for black victims, most suspects are victims, for white victims more suspects are white, and same with white hispanic. Regarding the sexs, it is seen that males tend to commit more crimes than females, and females tend to be more victims. We can emphasize that black victims of men offenses are mainly females with a very big difference, which does not occur in other races.

Regarding the clusters, we see that they have been distributed mainly by ages. So in the first cluster there are victims only older than 56 years, whose suspects are mainly people from 25-44. The second cluster contains victims of all ages, but crimes only committed by suspects older than 45. And the third cluster includes the young victims, being all younger than 44. In this case, the suspects are mainly between 25 and 44 years old, but also from 24 to 18.

Once we have added the spatial features to our dataset, we have decided to study more visually the records of the complaints according to their location. For that, we have created a second dashboard on Tableau. https://public.tableau.com/profile/aran3436#!/vizhome/NewYorkComplaints/Dashboard2

On the first page, we see the Neighborhood Tabulation Areas with most complaints, colored by their type. We see that the majority are located in Manhattan's district, with 6 out of the 10 represented. Bronx and Brooklyn, with two neighborhoods each, round out the top ten. Although their situations are different, the two Brooklyn neighborhoods are only behind Midtown-Midtown South, Manhattan, in the total number of complaints. Just below, and in a similar way, a bar graph with the 5 most common (specific) crimes, we observe the same neighborhoods of the previous graph, just in a distinct order, but is something that was somehow expected, the proportionality of the complaints' types.

On the second page, we observe the parks with the most complaints, the most famous, both in Manhattan, the Washington Square Park and Central Park, reaching a hundred complaints. We also see that they are usually categorized complaints from felony or misdemeanor. And just below, a graph with the most common places for complaints, categorized by districts, where a difference is seen between the number of complaints in the apartments and the street, with respect to the third most common place, the house as a residence. We notice the difference between the borough of Queens and the rest.

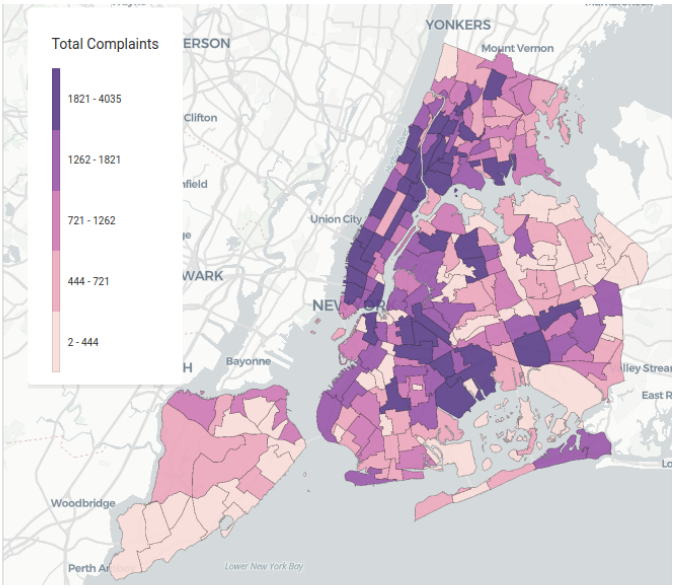
Turning to the third page, we try to see if there is a pattern that differentiates between the race of the victim and the suspect depending on the neighborhood where the complaints were recorded. Nonetheless, we found that there are not too many differences in the scope of location, nor does it happen in the age graph of the suspect and the victim, where the districts do not seem to play a crucial role.

On the last page, we reach to similar conclusions, seeing that the percentages of sex of suspect and victim do not have a great change depending on the district.

Summarizing everything up, we see that Manhattan is the district with the neighborhoods with most complaints, followed by Brooklyn and Bronx, and finally by Queens and Staten Island. When it comes to differentiating the characteristics of suspects and victims according to neighborhoods, we see that they follow a more general pattern throughout New York City.

TOTAL COMPLAINTS MAP:

One of our most explanatory analysis is the map with all the neighborhoods painted according to the total number of complaints recorded within their area. This provides us a better understanding of the data, thanks to the spatial variable.



In the map we observe that Staten Island (bottom left), as a district, is the one with less recorded complaints in its boundaries, followed by Queens (in the middle right), which has some areas with really smooth colours, but some others have the darkest color, meaning it is on

the highest rank of complaints. For the next districts, we have to do some calculations to know which has more complaints. It makes sense that Brooklyn (bottom center) is the top 1, since it is also the one with more neighborhoods within its boundaries, but if we take the mean of all of them, we notice that it is indeed the third one, taking similar mean values with Bronx (top center). At the end, and by looking at the map again, we see that Manhattan (center) is the district with more complaints per neighborhood, all of them being on the top level or close in our Total Complaints' scale.

All in all, with this project we wanted to extract how are the suspects and victims profiles to know who should be more guarded (potential suspects) or protected (potential victims), and identify which areas are more dangerous, to know where more police should be placed. After several analyses, we can conclude that most crimes are done by black people, followed by white hispanic and white. The same distribution is followed in victims, where black is the main race, followed by white hispanic and white. Regarding the ages, most suspects are males, while regarding victims there are more women than men. The most common age group for both suspects and victims is the one between 25 and 44, which includes the big range of citizens that do more activities in general, since they are in their adult life.

As an interesting insight, it has been seen that for violation crimes there is a marked pattern showing that males tend to attack women. This can be related with the discrimination women suffer in society.

Regarding the spatial distribution, it has been seen that Manhattan is the district with more complaints, followed by Brooklyn and Bronx. We also wanted to see if the profiles of victims and suspects depend on the geographic zone, so if one neighborhood was attacked mainly by one race, for example. However, we have seen that this hypothesis is not true since the age, sex and race characteristics for both suspects and victims follow very similar distributions for all the municipalities.

4. Next steps

With more time, to make a really complete analysis of the complaints in the city of New York, our database could be joined with demographic and socioeconomic data of all the neighbourhoods, so that crimes can be related with the number of inhabitants and the citizens status. This would complement and enrich our conclusions with the real situation of every county of New York.