MSAI 339 Project Report

Group 4: Amit Adate, Mayank Malik, Souvik Bagchi

December 10, 2018

1 Introduction

As the gun violence in the USA is at all time high, resulting in tens of thousands of deaths and injuries annually, we were intrigued to work on the CPDP project that gave us the opportunity to focus on weapon usage by officers and subjects in this project. The Citizens Police Data Project ,CPDP, has records of police and public interactions which would otherwise be ignored in huge local databases. CPDP opens these records up to make the information useful to the public, creating a permanent repository for every police officer and a public record for every civilian complaint.

From those public record, we tried to infer some really interesting facts, but not limited to :

1. What are the type of weapons discharged by police and victims.
2. Were the weapons discharged lethal or non lethal.
3. Police and victims of which color, race, sex, position are more prone to weapon discharge.
4. Which weapon a police officer can discharge in a given circumstance . (predictive model)
5. Free text analysis on TRR reports
6. Performed visualization using tableau for many of the above analysis.

2 Relational Database - Checkpoint 1

2.1 Description and summarization

Study of the conditions that lead to a weapons discharge in the CPD and type of weapons used. Also, study the effect of weapon discharge on settlements. The below mentioned analysis is in 3 parts:

2.1.1 Relationship between Weapons discharges and Police Officer's Race

As one can infer from that White and hispanic cops are more prone to weapon discharge than black officers. (we are ignoring asian police officers due to small sample size).

Race	count police by race	times weapon use	Number of times weapon used per person
white	20671	5269	0.25
black	7630	1402	0.18
hispanic	4579	2225	0.48
Asian/pacific	539	234	0.43
rest	Insignificantly small	Insignificantly small	

Figure 1: Relationship between weapon discharge and Police officer race

2.1.2 Relationship between Police officer ranks and weapons discharged

Although Sergeants are fewer in number, they are clearly more prone to weapon discharge. Note - Due to lack of sufficient time, we couldn't find out if sergeants used weapons more often before the promotion or after. This is something we would love to figure out when we have more ample time once the quarter ends.

Rank	count police by rank	times weapon use	Number of times weapon used per person
Police officer	22966	6240	0.27
Sergeant	3495	1859	0.53

Figure 2: Police officers vs Sergeants vs weapon discharge

2.1.3 Analysis between race of the victims of weapons discharged

	count bigint	race character varying (50)
1	59298	Black
2	12406	White
3	10626	Hispanic
4	6290	
5	747	Asian/Pacific Islander
6	120	Native American/Alaskan ...

Figure 3: Race of the groups (victims) that got most affected.

Although Chicago has approximately only 30% of black residents, they account for 66% (59 K/ total complaints) of the complaints. The complaints by blacks are significantly more and hence is point of deliberation.

2.1.4 A few other common deductions:

The average of the police officers is 38.3 years, which we think is on lower side. Similarly, their average Salary is approximately 75 K, which is lower too with respect to their age. These observations were important in the machine learning model creation.

3 Data Cleaning and Integration - Checkpoint 2

Data cleansing is a valuable process in all data science projects, helping save time and increasing the accuracies.

1. Data Wrangling with Trifacta allowed us to discover, wrangle visualize complex data quite quickly. We used Trifecta to perform a clean dataset, that has features pertaining to our use cases.
2. We learnt the following about the data during the completion of this checkpoint:
   - The police officers who are ranked lower in the hierarchy are more likely to fire weapon than someone who is higher up in the chain of command
   - The officers who discharge the most weapons belong usually to the white race.
   - The race of the victims is predominantly black.
   - The most settlement has been with black subjects.
   - The area which is most affected seem to be South Chicago. With the highest going to tasers.
3. This checkpoint was essential to provide us an insight within the data and Trifecta was an accessible and noteworthy tool that we used in our multiple other course projects.

4 Workflow Analytics - Checkpoint 3

4.1 What are the similarities between individuals that were involved in confrontations with the police?

4.1.1 Analysis:

In almost all the cases under observation, the victims had confrontation with the police, so we thought it is better to see the patterns among all the victims. We have implemented K means clustering to make clusters of the victims on the basis of the following features - gender, race, and age. We also created the correlation matrix between count(the number of times the weapon is used against victims), gender(male is 1), and age.

4.1.2 Results:

1. As expected the Age is negatively correlated(-.56) with count(the number of times the weapon is used against victims), which means that young people have been the victim of weapon discharge more than older people (Might be because their aggressive nature) .
2. Similarly, Being male is positively correlated (.4) with the count, which was also anticipated, which means that males have been the victim of weapon discharge more often than females. This is expected because males are considered more aggressive in general.

4.2 What are the similarities between different individuals who faced weapon discharge?

We tried to find the similarities or pattern or correlations in the victims who have been faced with the weapons by the cops. The KMEANS did a good job in creating the clusters- It creating two clusters: 0 is majority males and young people and 1 is more females or older people. It is also corroborated with the fact that most people who were the victim of weapon discharge were males and young and are larger in number. Hence, both K Means and correlation analysis get to the same point.

4.3 Mapping out the regions where weapon discharges happen?

For this analysis, we used K means for clustering and random forest machine learning model to find the feature importance. Random forest algorithm predicted North and South region as the most important feature for predicting for Lethal weapons usage.We found out that weapons discharged were mostly in SOUTH and very low in NORTH, which is supported by the fact that more crimes happen in the south Chicago.

5 Machine Learning - Checkpoint 4

5.1 What kind of weapon a police officer is likely to use under given circumstances ?

we have used to classify the weapons in the following 3 categories

1. NO WEAPON
2. TASER
3. LETHAL WEAPON

The Algorithm used was Random Forest Classifier.

Age, gender and subject weapon discharged were highly important in our model prediction. Other features such as location, officer-rank and others which did not help the models were dropped during the iterative model building process, improving our model accuracy from 60% to the current 86%.

5.2 If a Police Officer has an allegation against him can we predict if there will be action taken against him or not ?

The Algorithm used was Random Forest Classifier. Initially we thought we wouldn't be able to predict with high accuracy if an alleged police officer will be reprimanded or not, but eventually we were able to predict this with 86% accuracy. The analysis was carried out by adding a single column - Reprimanded or Not Important indicators - Weapon type used, Age of subject

6 Modelling with Neural Networks - Checkpoint 5

We performed a free-text analysis on the TRR Reports provided.

Initially, we extracted data from the TRR Reports and it was intended to be classified into 8 categories:

1. Resist
2. Obstruction of Justice Battery
3. Bodily Harm Battery
4. Physical Contact Resisting
5. Obstructing Aggravated Assault
6. Issuance of Warrant Assault
7. Simple Domestic Battery Insurance
8. Operating without Insurance

We were getting spurious results and the baseline model (bag of words) was giving about 58% accuracy. After tuning a lot of parameters, we concluded that the amount of text files that are available for analysis and comparison are not enough for vector representation and nearest neighbour computation.

In our experiments we went on to perform document classification on 3 different models, and the common metric for comparison being multiclass logarithmic loss. Our experiments tell us that attributes of TRR-CHARGE, Description and Statue are the top two most important features to classify based on the TRR-ID.

The classification models that were most effective to least effective were:

1. Word2Vec
2. TF - IDF
3. Bag of Words

Best classification accuracy attained is 62% by Word2Vec. To view our experiments performed with neural networks, visit this **link**.

7 Visualization - Checkpoint 6

We worked on a tableau workbook to investigate the following visualizations:

1. Visualizing the race and gender of the officers given the type of weapon discharged
2. Visualizing the types and counts of weapons used given the age of the subject
3. Visualizing the types and counts of weapons discharged by the police officers
4. Visualizing weapons discharges on the map of Chicago

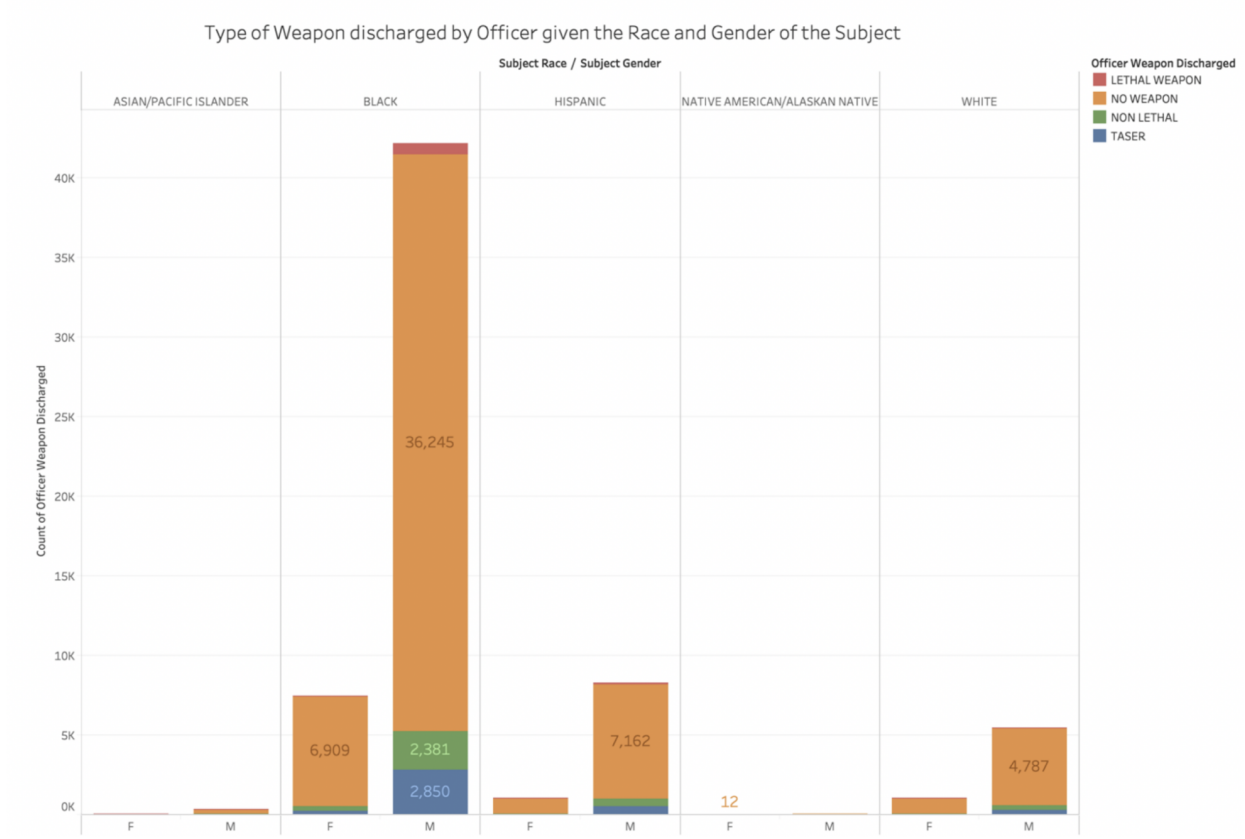

Figure 4: Type of weapon discharged given the race and gender of the subject

As clearly seen from Figure 4 if you are a male then it is more likely that you will face a weapon discharge by an officer.

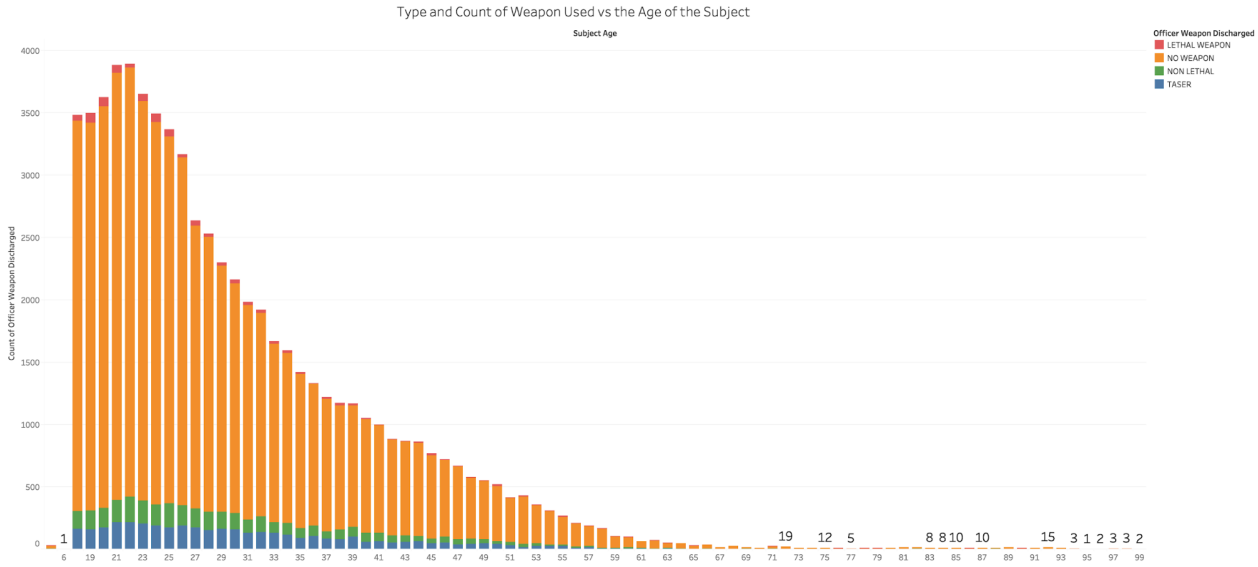

Figure 5: Type and count of weapon used against the age of the subject

It can be clearly seen from Figure 5 that as the age of subject increases we steadily see the drop in the weapons being used. At first we see the increase in the number of weapons being used with the peak ages around the 20s and then it keeps dropping.

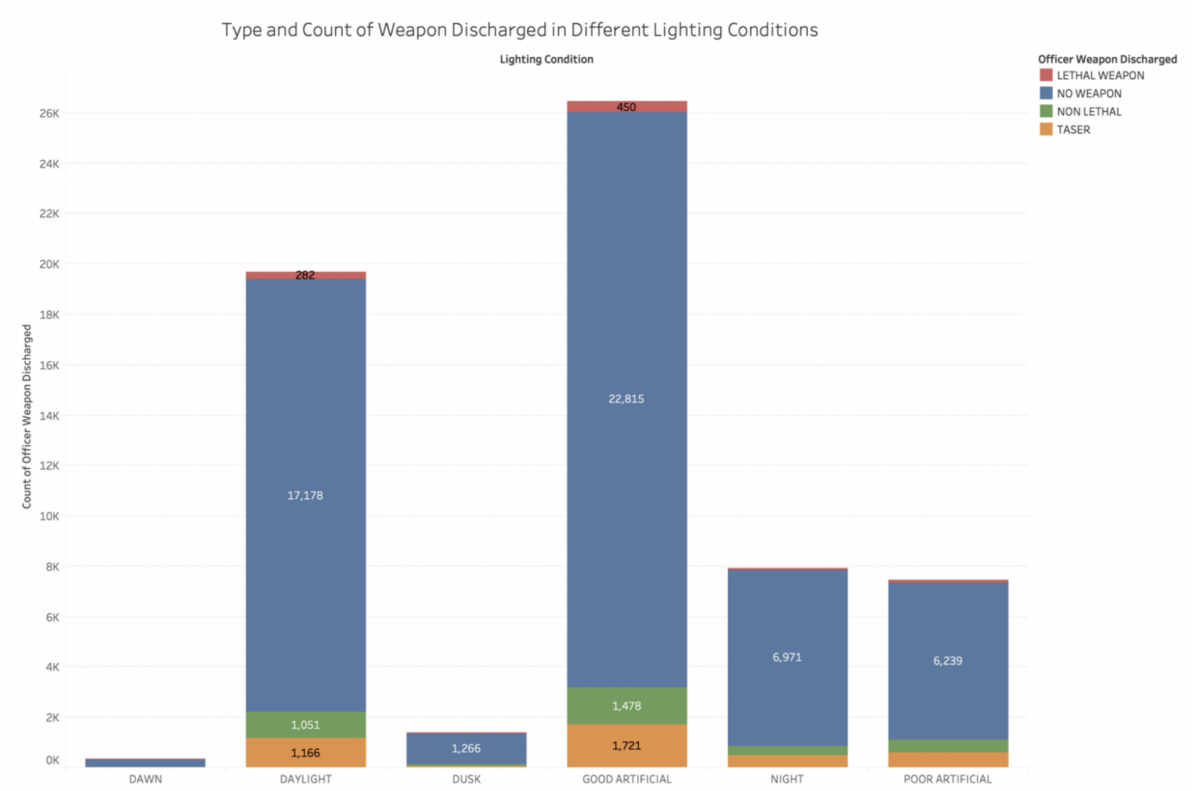

Figure 6: Type of weapon discharged in different lighting conditions

As we can see from Figure 6 that the weapons being used by the officers seem to be more during artificial lighting. This can be attributed to the fact that artificial lighting is usually present indoors or in poorly lit areas which might make an officer apprehensive of any subject.

In the next couple of visualization, we try to see how it looks on maps of Chicago. The following visualizations will show the type of weapon discharged mapped to the areas in Chicago.

The map area has been kept constant to see overlapping areas. The legend of individual areas also show the count of cases which help us understand the relative use of these weapons.

Legend

1. Green - Non-Lethal
2. Red - Lethal
3. Gray - No weapon used
4. Blue - Tasers

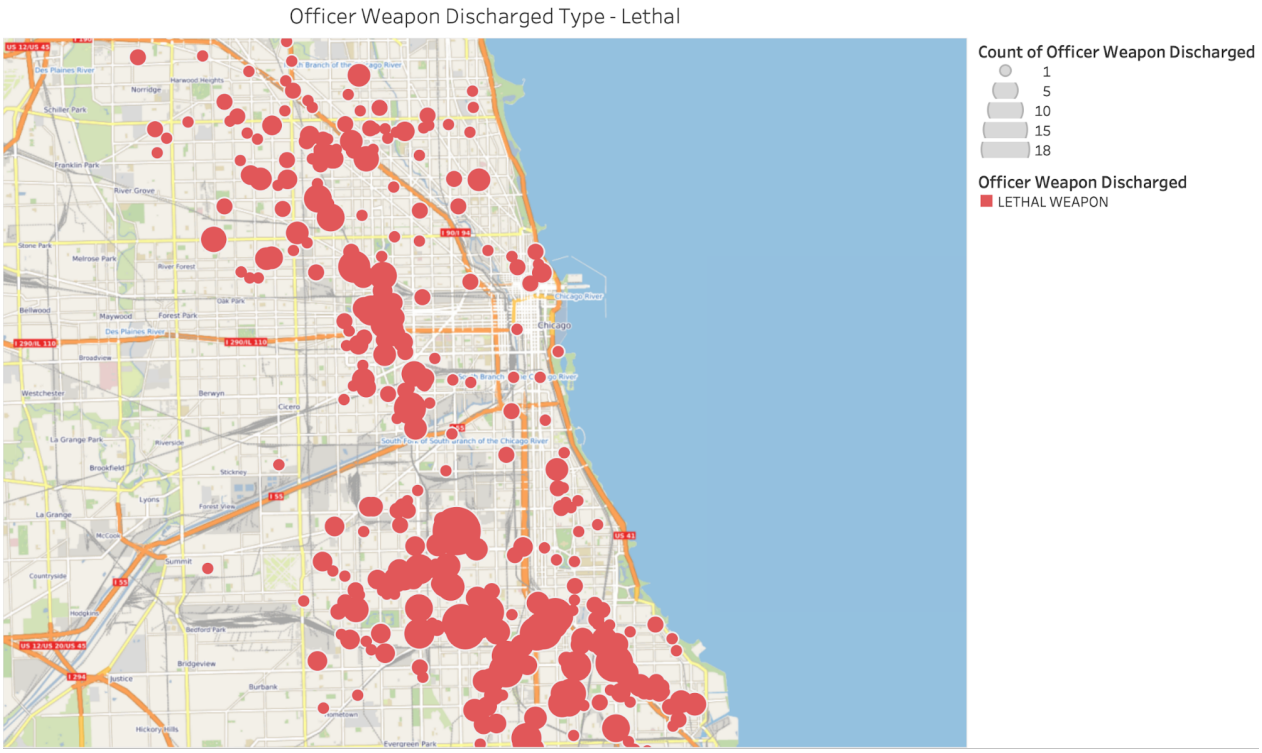

Figure 7: Officer weapon discharged type: Lethal

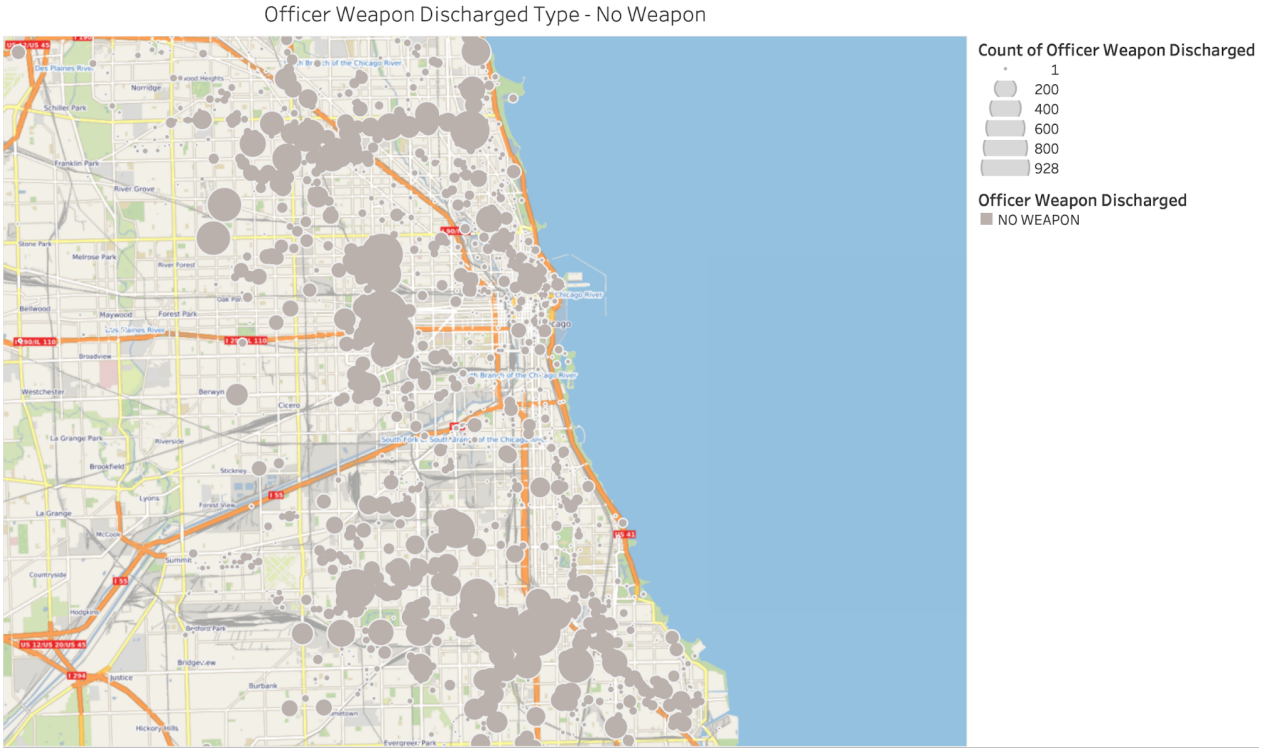

Figure 8: Officer weapon discharged type: No Weapon

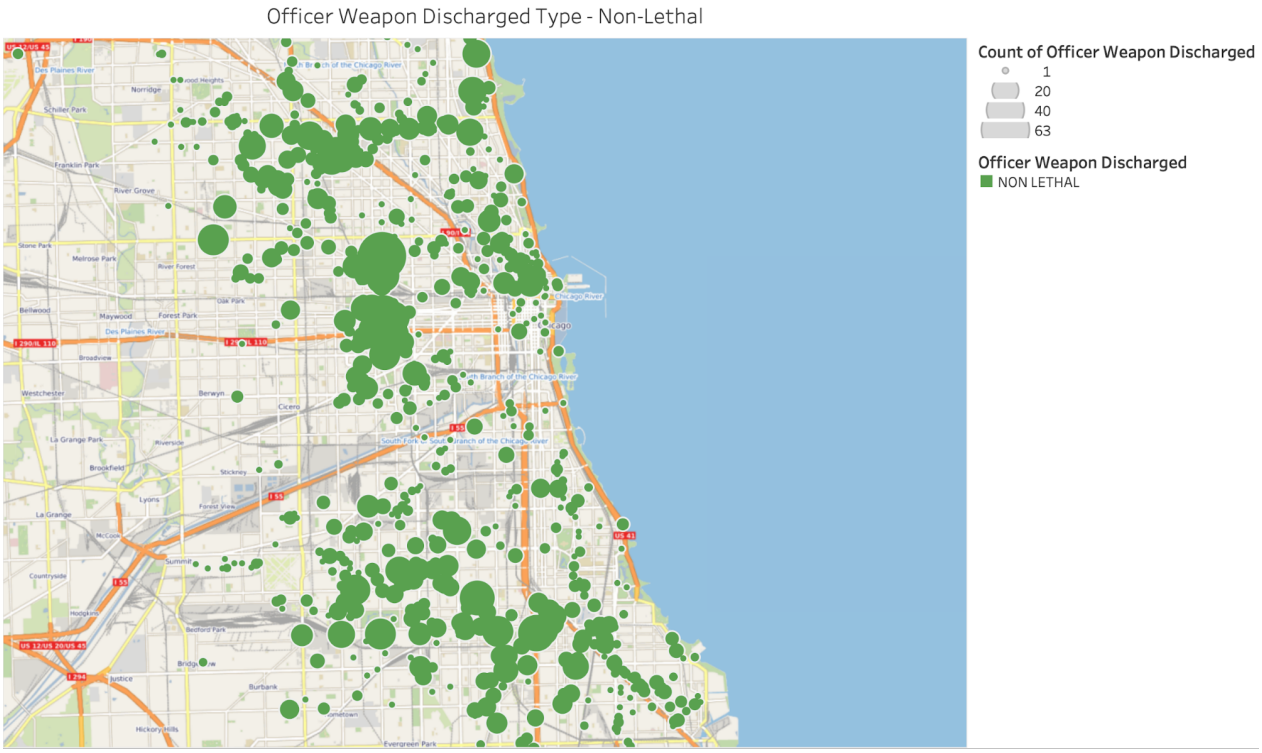

Figure 9: Officer weapon discharged type: Non Lethal

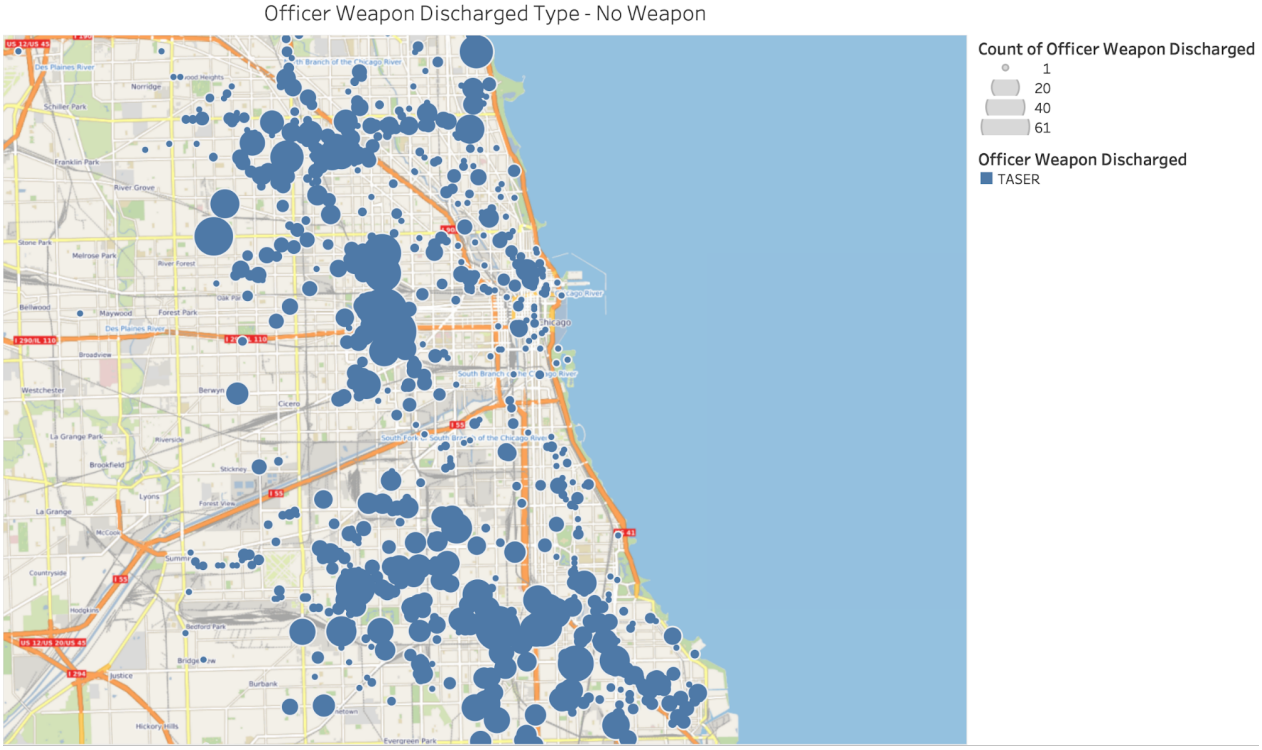

Figure 10: Officer weapon discharged type: Taser

As we can see from the above analysis that the number of lethal weapons discharged are the least when compared to other weapons used. This is a good indication that the police officers are showing restraint with the subjects. We can also see that lethal weapons tends to be used more in

the south and north side of Chicago.

The comparative map which combines all the types of weapon discharged by the police officers:

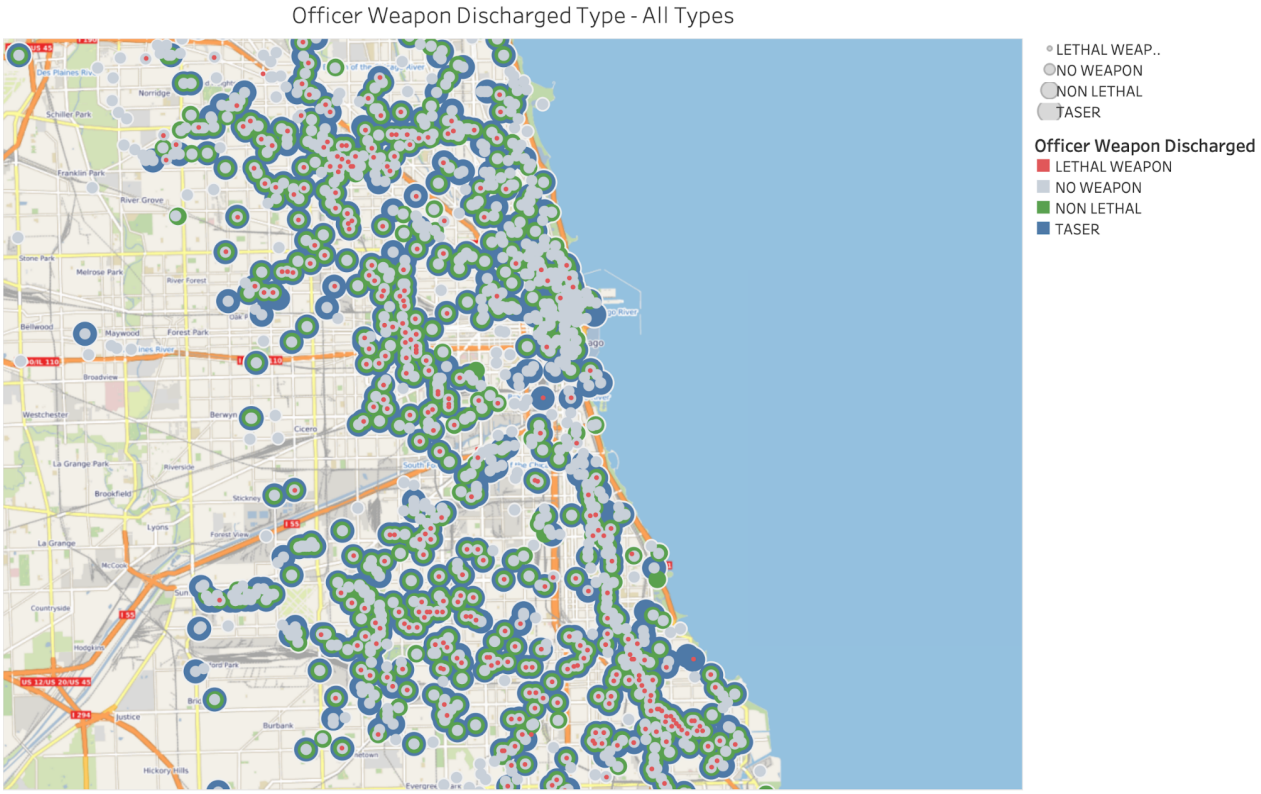

Figure 11: Officer weapon discharged type: All Types

Furthermore upon further analysis we found that the firearms used by the subjects are concentrated in pockets in the north and the south east. Most offenders dont use any kind of weapon, while people have tried to use their own vehicles as the least priority weapon. Here the usage of vehicle refers to the subjects trying to hurt the police officers with vehicles.

8 Conclusion

In conclusion, we are glad we were able to bring light to some critical inferences about discriminatory practices, which is point of deliberation for all of us. One of such key conclusions is that officer weapon discharge is dependent on the neighborhood. This infers that officers behave diversely in every district across the city. Also, the officers with higher rank are more likely to discharge weapon. We also found that there have been anti-discriminatory practices against the younger population by the officers of Chicago.

We believe CPDP has been a great learning opportunity for all three of us, rendering myriad of inferences. Nevertheless, we still believe there is definitely a scope of improvement in our work and we hope to continue working on it as diligently as we did before.