Faraaz Beyabani, Andrew Vagliano, Harkirat Gill
November 3, 2020

Checkpoint 4: Machine Learning

For this checkpoint, our goal was to use machine learning to identify areas of Chicago that were being overpoliced. For this purpose, we used linear regression with independent variables such as year, number of officers with at least one complaint, and the ratio of Use of Force and Illegal Search allegations in each district over time. We examined these 'moderate severity' allegation types in particular because they struck a balance between frequency and severity. Other allegation types such as Personnel Violations are too common and don't provide much information about harsh policing. On the other hand, more severe allegations such as Bribery and Criminal Misconduct are far too rare in every instance to be considered for overpolicing.

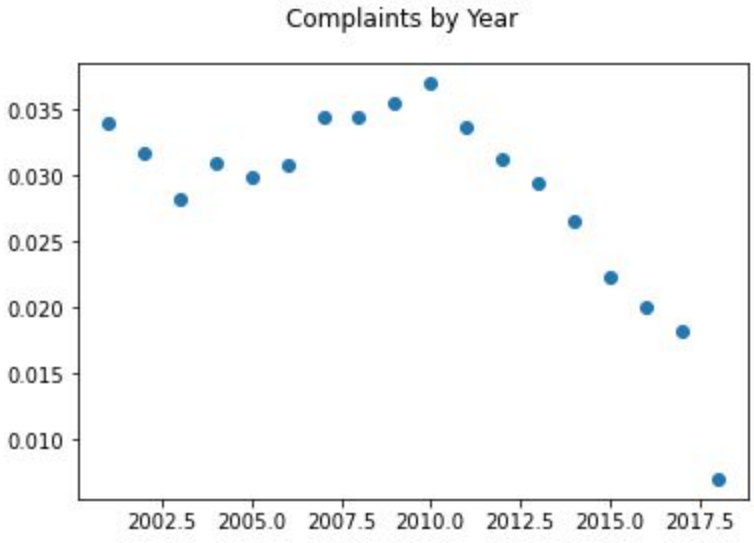
Our previous analyses involved the visualization of demographics and other statistics on a district-by-district basis. In order to consolidate data and facilitate comparison, we created a superset of regions, which contained 5-6 districts each. These are the questions that we aimed to answer with this assignment:

- How do the numbers of these moderate severity complaints change over time for each region?
- Do regions with higher officer counts have more occurrences of moderate severity complaints?
- Do higher ratios of Use of Force and Illegal Search complaints have a direct relationship with the number of moderate severity complaints?
- Given the relationships gleaned from the data and the previous questions, which regions, if any, are overpoliced?

Number of Complaints Over Time

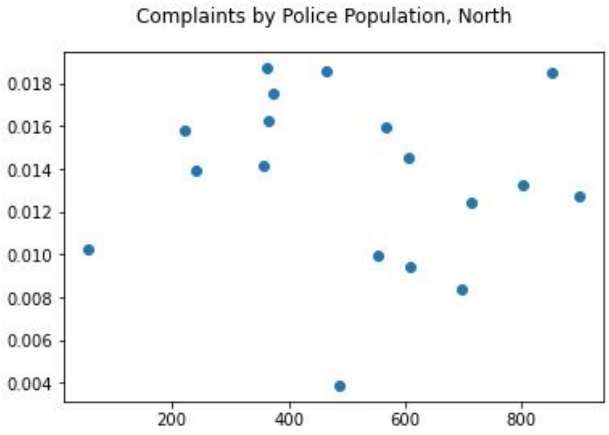
For this question, and every question hereafter, we normalized the number of complaints per year by the total population of each area, which gives us the average number of complaints per person. From the graph below, which shows the complaints per year for the North region, we can see that there is somewhat of an inverse relationship between the number of complaints and the year. In other words, since 2001, the number of complaints has been on a downward trend. This is the case for all regions, but it is also true that some regions have a consistently larger number of complaints over time. This is true for the Central and South regions, which have maximums of 0.5 and 0.6 complaints per person, compared to the 0.18 and 0.35 for the North and West regions.

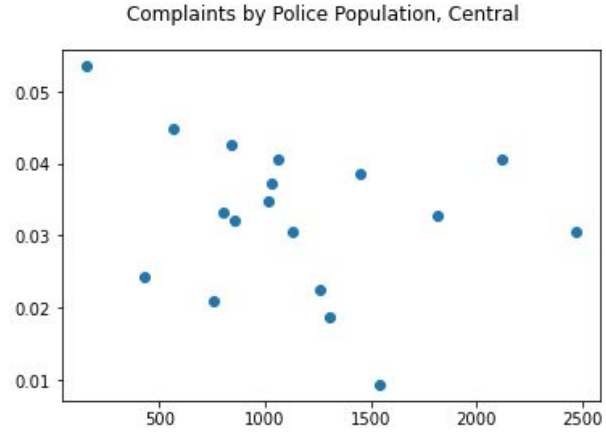
This implies that some districts in the Central and South regions might be overpoliced, considering that the average complaint per person over time is also higher in these regions. However, this may just be a side effect of a larger civilian and police population, therefore it is necessary for us to address the other factors that may play a role.



Number of Complaints Versus Police Population

The feature of ‘police population’ refers to the number of police officers who had at least one complaint filed against them in each year. We chose this feature because we believed that a higher number of officers accused of some complaints would directly translate to a larger number of complaints per person in each region. This statistic would then assist us in determining overpoliced regions. Our results for the North and Central regions are shown below.

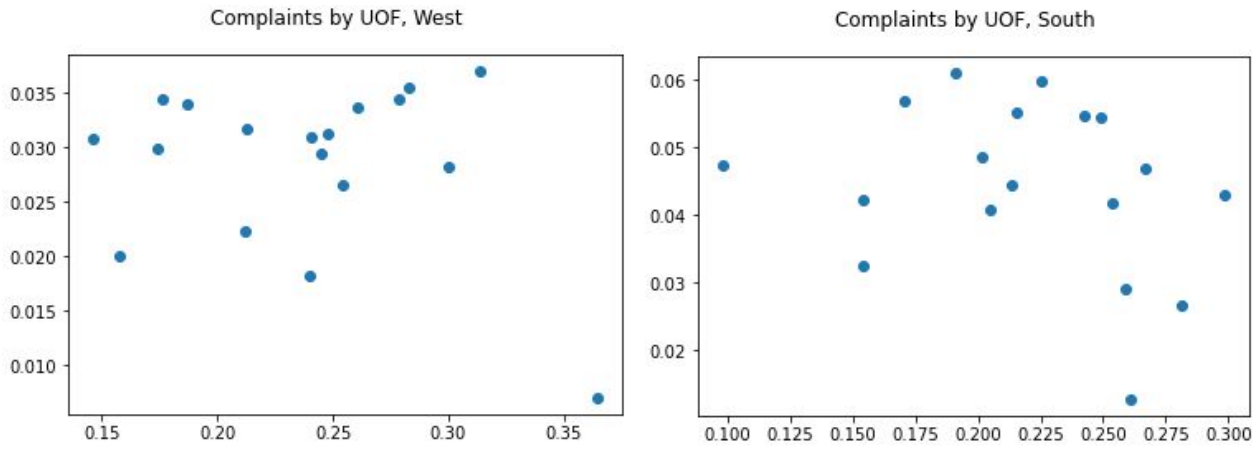




There is virtually no relationship between the number of complaints and the number of officers with complaints within a single region. This is unfortunate, but we were still able to understand some information by comparing the plots for different regions. For instance, with the above two regions, we can see that the maximum number of complaints and maximum number of officers are significantly lower in North than in Central. Therefore, this data might have been better represented via clustering, rather than regression. Using this evidence, we can consider the possibility that our hypothesis for this feature was correct, but the method used to analyze this feature was not ideal.

Number of Complaints Versus Use of Force Ratios

As aforementioned, the ratio of an allegation category refers to the ratio, or percent, of all allegations that fell under that category. For example, if a region has a Use of Force ratio of 0.18 in 2013, this means that 18% of all allegations were categorized as Use of Force in that year. Having said that, our goal with this feature was to show that larger allegation type ratios would also result in a larger number of complaints This would imply that officers would go out of their way to use excessive force against groups or individuals they disliked. The data for Use of Force allegations in West is depicted below.

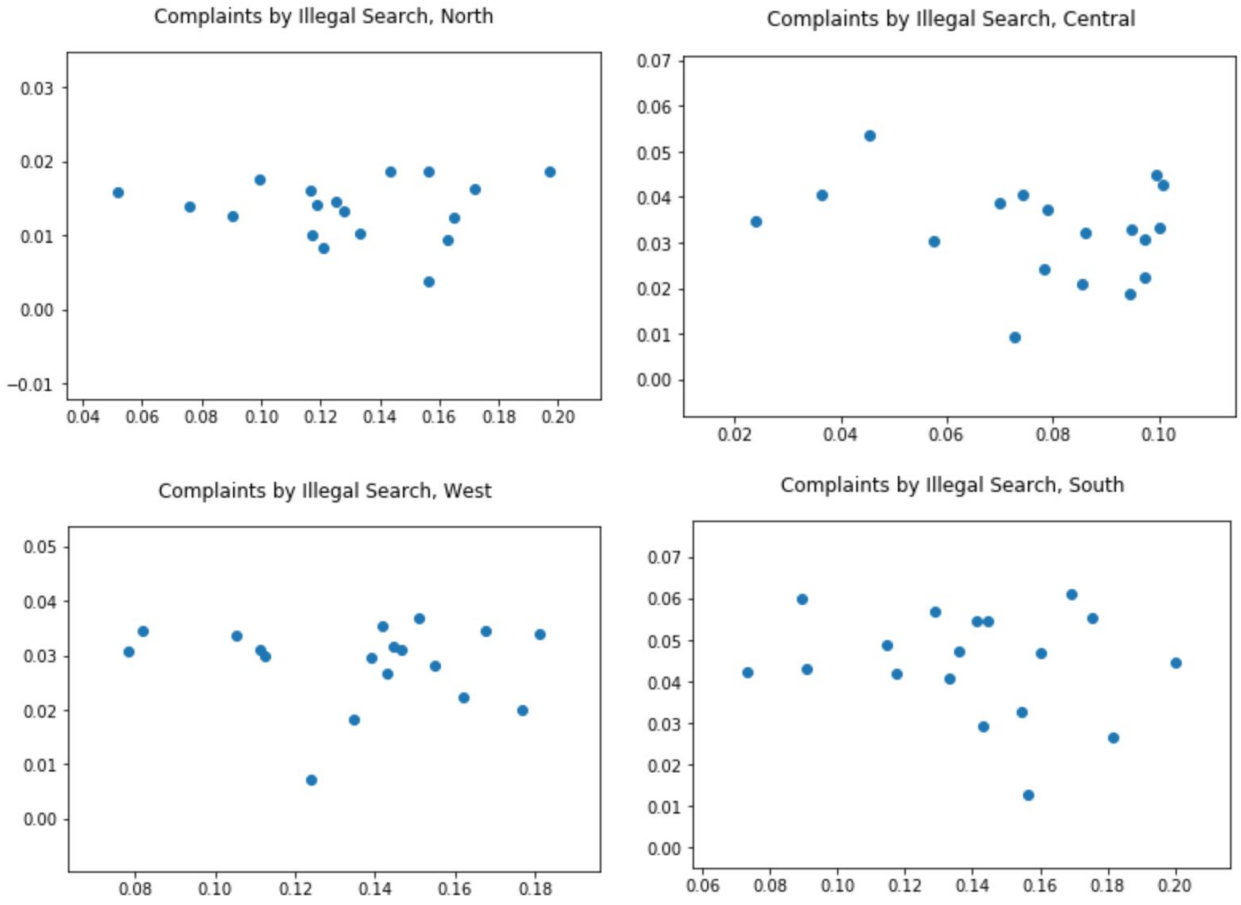


As can be seen, we actually found somewhat of a static or even weakly negative correlation between this type of allegation and the total number of complaints. This means that our hypothesis was incorrect. However, we can see that in the South data, we do have a cluster of points that have both a high UOF ratio as well as a high number of complaints. While this might mean that complaints per capita and UOF ratio are weakly correlated in the grand scheme of things, the average rarity of this allegation ratio and our likely incomplete data for each district over time result in us not being able to make this claim with confidence.

Number of Complaints Versus Illegal Search Ratios

Similarly to the UOF allegation ratios, Illegal Search (IS) ratios represent the portion of the total allegations that is made up of Illegal Searches. In this exploration, we hoped to understand whether there was any discernible relationship between the per-capita number of total

complaints and the IS ratio, potentially theorizing that police would be more likely in higher-complaint regions to illegally search community members, or vice versa.



As the data shows, there is a slightly inverse relationship between these two features, with per capita complaints decreasing as illegal searches increase. This could be the result of more consistency city-wide in illegal searches than total complaints, meaning that in higher-complaint regions the illegal searches are dominated by other forms of complaints.

Conclusion

Overall, our analysis for this checkpoint was not very successful compared to our previous studies. While we were able to see an overall downward trend for UOF and Illegal Search complaints per year, we were unable to meaningfully apply our features towards finding regions that are overpoliced. While we believe that certain districts in the central or southern areas are overpoliced, we could not find evidence that explicitly showed this.

If we were to iterate on this analysis, we believe it would be useful to explore alternative avenues of feature prediction, such as k-means clustering over all regions or non-linear regression methods. It would also be beneficial to obtain and organize new features, such as

the crime rate of each district per year, true civilian population statistics over time, and the number of sustained allegations per district over time.