Settlement and Discipline in the Chicago Police Department

The Wicked Roadrunners

Introduction

Our goal this quarter was to explore how settlements and disciplinary action relate to police behavior and career. We want to explore both settlements and disciplinary action because they are two ways of acknowledging wrong-doing by the police. If we can identify what officer and allegation characteristics make discipline or settlement more likely, we can gain insight into how the CPD handles allegations and misconduct. We can also get a sense of the CPD culture, seeing if there are any biases within the institution toward certain groups or practices. For example, if we notice black officers are disciplined more than white officers per allegation, that would suggest racial biases within the CPD. We can also see how certain behaviors shift with changes to unit or rank.

How does rank relate to discipline?

We wanted to look at how rank and seniority relate to discipline in the CPD. We wanted to answer questions like ‘are officers of higher ranks disciplined more?’ or ‘do officers of higher ranks think they can get away with using offensive language at victims?’. The impact of rank is important for understanding CPD culture because higher ranks typically serve as an example or role model to lower ranks. One limitation was the CPDB database did not contain information on an officer’s rank history. We could see a history timeline for an officer’s units, but we don’t have this information for ranks. In the future, this would be valuable and interesting data to acquire and look at.

Before diving deeper into the nuances of rank and discipline, we first looked into the average discipline rates per rank (Figure 1). One caveat is that several police officers play multiple roles in the department. For instance, there is Police Officer (PO) as Detective, PO Assigned Canine Handler, PO Substance Abuse Counselor, etc. The most prominent feature in this graph is PO Per Arbitration Award, which has a staggering average of 3.4 disciplines per officer. This may not be surprising since an arbitration award is provided towards officers as a quicker alternative to having to go through a court of law. In other words, it would be surprising for a cop to have this award provided to them without having a complaint that would lead up to a legal situation where a cop needs such an award. Otherwise, ranks of interest would be the ranks of first deputy superintendent, canine handler, detective, and assistant superintendent. This implies that the roles of leadership may have some impact on how many complaints an officer gets. This aspect is also notable when noticing that lieutenants, sergeants, and field training officers have comparable averages to police officers despite being a smaller portion of the police force.

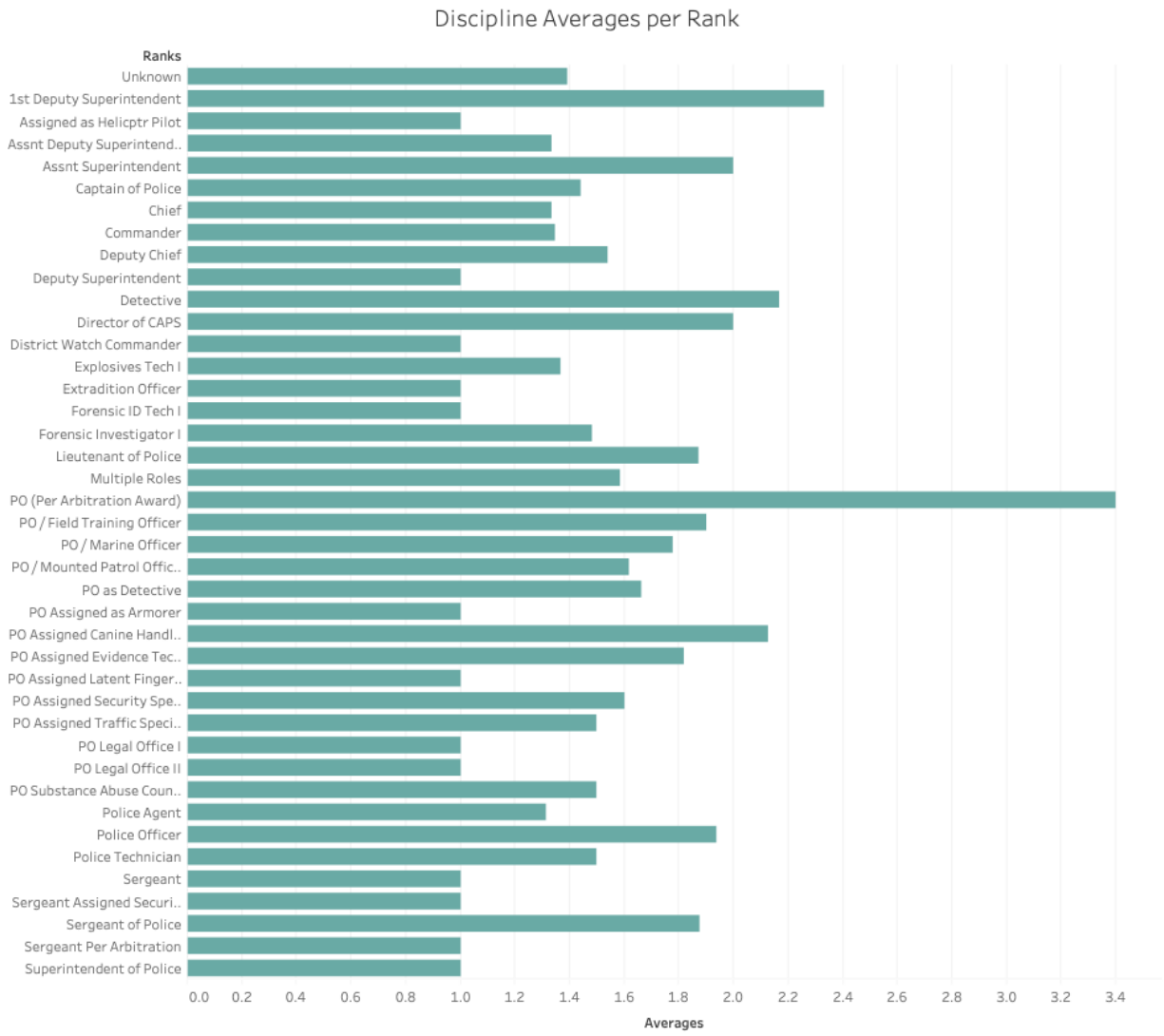


Figure 1

We also were exploring how much settlement costs do officers grouped by their common allegation cause. We created two bar charts that look at the average and total costs of officers groups by their most common allegations. In Figure 2, we can observe that in terms of the total cost, the officers who have a complaint history in the use of force and verbal abuse then constitute the largest portions of total costs. But we need to take into account that verbal abuse and use of force may be more common than other categories, which is why we have the graph with average costs. Here it is seen that while the use of force and verbal abuse is still costly per person, it's not comparable to the far larger average cost for Supervisor Responsibility. Thus, we can see more is expected of Supervisors, and there are major financial consequences when they mess up. This finding shows some promise for the CPD. A supervisor's job is to create good officers by rewarding good behavior and correcting bad behavior. If we know there is a cost to supervisor failure, that should motivate supervisors to monitor their officers more effectively.

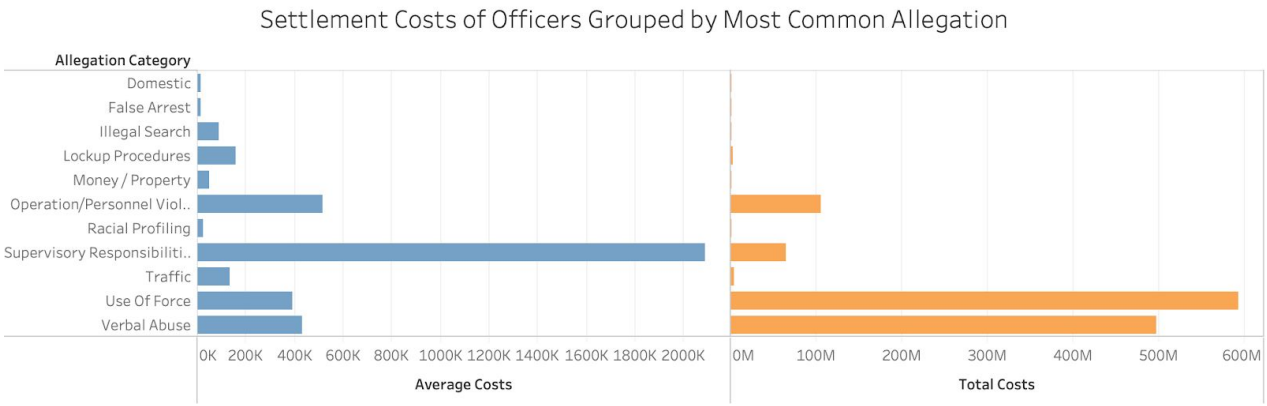


Figure 2

How does unit relate to discipline?

Since unit members spend a lot of time together, it makes sense to explore trends among and between units. Why do certain units cost the CPD more money than others? Why do certain units have higher discipline and complaint rates than others?

One theory we had was that disciplined officers would switch units more often. For instance, if a cop has a high number of allegations/disciplines, they might be moved to a different unit that deals with less risky calls and patrols. We calculated the average number of unit changes over a total career for disciplined and not disciplined cops. We found that cops with discipline have an average of 4 unit changes over their career compared to 3 for non-disciplined cops. While there is a difference between these numbers, the difference is small enough that there might not be practical significance to it.

We also found a surprising relationship between career success (awards) and discipline. We grouped officers by their unit and plotted the log of awards per unit vs. disciplines per unit. We can see in Figure 3 that there is a non-linear positive relationship between awards and discipline. Also, we noticed there is an increased spreading of points the further upper right the points are. This brings up the question if there is a third variable at work in these high-award, high-discipline areas. It could be that these units may view a more radical methodology that the community disapproves of as a positive. But it could also be that these units are bigger or work in more dangerous situations leading to differing perspectives on these units' bravery and anxiety-inducing presence.

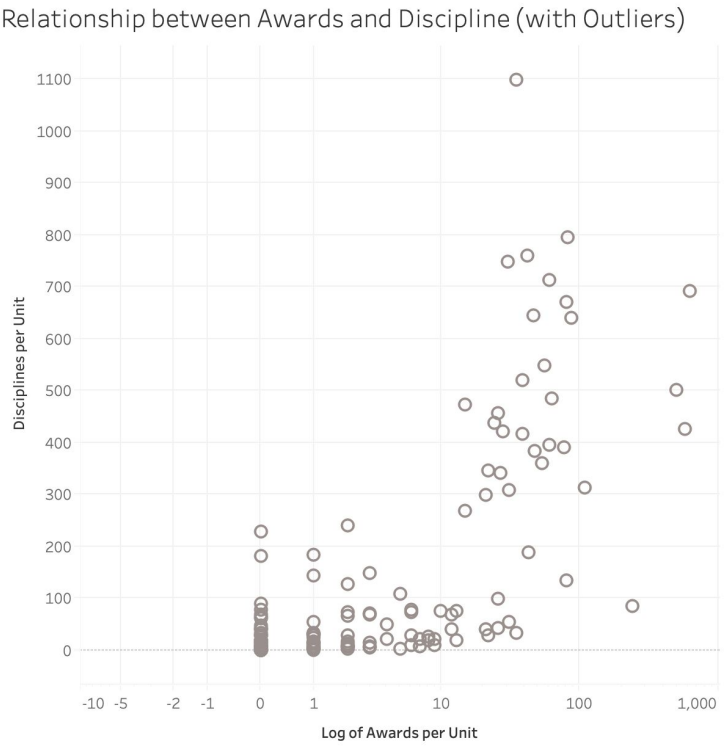


Figure 3

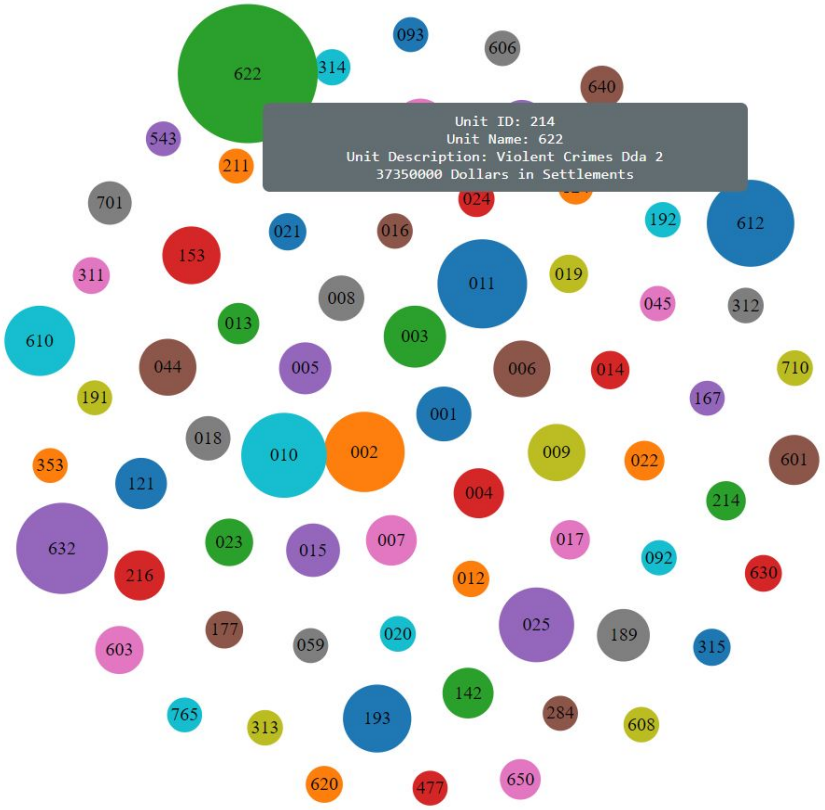


Figure 4

Using an interactive D3 bubble chart, we were able to explore which units have the highest settlement costs (Figure 4). Each bubble represents each unit, and the size of the bubble represents the settlement costs (total overtime). Hovering over the bubbles will display the unit name and settlement costs. The largest contributor is Unit 214 (Violent Crimes Dda 2), costing the CPD almost 40 million dollars. The other police units do not even compare in settlement costs. The runner-ups are Units 217 (Violent Crimes Dda 3), 12 (District 11/Harrison), and 211 (Violent Crimes Dda 1), each costing around 20 million dollars over time. To some degree, this isn't surprising since the most expensive units are those dedicated to violent crime or work in high-crime neighborhoods. The presence of more violent crimes likely contributes to additional aggression, abuse, and general improper behavior that could lead to more settlement costs.

We also were interested in seeing how certain units handle allegations over time. We looked at police districts since they correspond to geographic regions and units 2-26 map to districts 1-25. If we see the ratio of disciplines to complaints decrease over time, that could suggest officers increasingly not being held responsible for their behavior. First, many districts have incredibly high percentages for data before and during the 1980s (80% or more). Then, a couple of data points later, the percent drops significantly (20% or less). This may be a result of limited data points during that time. In many of the charts, we see a decrease in percentage over time. This suggests that these units don't take complaints/allegations seriously, as they don't lead to discipline. See Figure 5 for examples of these trends. In the future, it would be valuable to explore neighboring districts and similarities between their line plots since they are geographically close together. It could suggest that behavior among cops can spread through geographic proximity to each other.

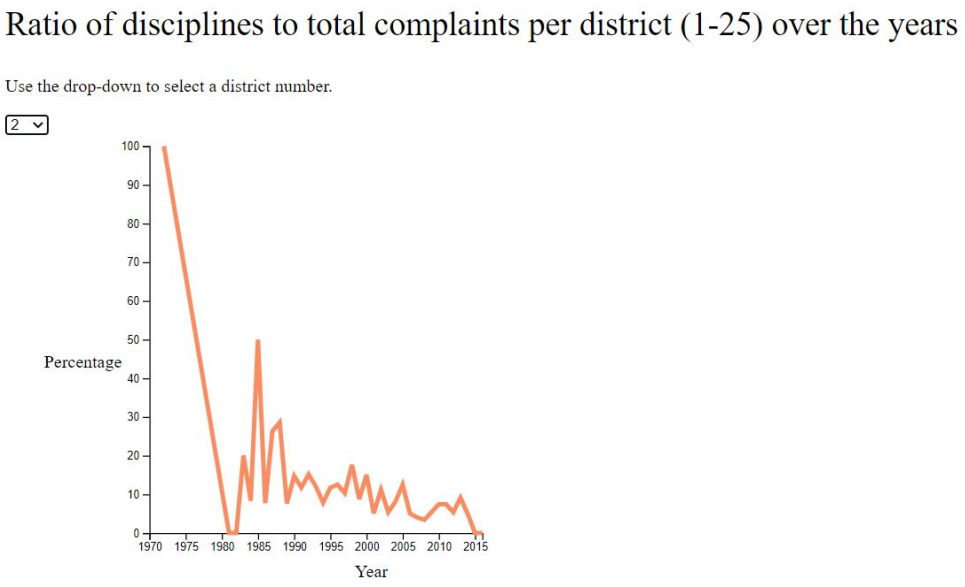


Figure 5

Can we predict discipline or settlement amounts?

Next, we wanted to utilize machine learning to see if we can predict discipline or settlement amounts. First, we investigated whether we could predict whether an allegation would lead to discipline based on officer and allegation data. If we can determine which combination of officer and allegation attributes lead to discipline, we can then further explore why some allegations are taken more seriously than others. Seeing which characteristics create the best predictive model also offers other unique insight into the CPD. We found that the features of race, rank, complaint percentile, allegation count, and misconduct category gave the best performance for predicting discipline. As previously shown in Figure 1, we see not all officer ranks get disciplined proportionally; thus, it isn't surprising that using rank as a feature strengthens our model. Allegation characteristics make sense for the predictions as misconduct like excessive force should lead to more discipline than an operation violation. On the other hand, a relationship between officer race and disciplinary action implies racial biases within the Chicago police institutions. We found black officers are disciplined more than white officers per allegation, meaning white officers can get away with more misconduct.

After some preprocessing like fixing class imbalances and encoding labels, we created a Random Forest model with a hyperparameter of 15 trees.

We evaluated our model in two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy (Figure 6). Second, we used 5-fold cross-validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.88695522 0.88499464 0.89054067 0.89162476 0.88187941]. Thus, we can confidently say our model was a success and can predict with 89% accuracy whether an officer will be disciplined for an allegation. However, we were surprised by the impressive performance of our model. Since very few officers are held accountable for misconduct, we expected the data to be more random and noisy. Instead, we see that there is a pattern to officer accountability.

	precision	recall	f1-score	support
False	0.961	0.811	0.879	43302
True	0.837	0.967	0.897	43407
accuracy			0.889	86709
macro avg	0.899	0.889	0.888	86709
weighted avg	0.899	0.889	0.888	86709

Figure 6

Second, we wanted to see if we could predict the settlement amount of an allegation lawsuit based on officer and allegation characteristics. Since a settlement amount difference of a few dollars is insignificant, we tried to predict the magnitude of the settlement amount. For example, $10 has a magnitude of 1, $100 has a magnitude of 2, $1000 has a magnitude of 3, etc. The reasoning behind this is if we can see which officer or allegation characteristics cost the CPD

the most money, then we can also find which areas in the CPD and the lawsuit process are the most costly, too. This model would also be valuable to civilians, as they could assess whether it is worth it to try to settle for money in court. If the predicted settlement amount is small, then the process might not be worth it with all the time and legal fees involved.

The input features we used for the model were: interactions (e.g. traffic stop, home invasion), outcomes (e.g. killed by an officer, drug charges), misconduct category, gender, and race. As previously shown in Figure 2, we saw that settlement costs varied highly by the misconduct category. Thus, it makes sense that this attribute enhances our model. We also found that male officers have around twice the average settlement cost as females. We suspect this is because most cops are male, and male officers tend to commit more violent misconduct. We also found that black officers have the highest average settlement cost, shortly followed by white officers. Other racial groups have significantly lower average settlement costs. The other features are related to the allegation and are more intuitive. For example, we would expect being killed by an officer to result in a higher settlement than resisting arrest.

After identifying important input features, we performed data preprocessing, including transforming settlements to their magnitude and fixing class imbalances. We then created a Random Forest model with a hyperparameter of 10 trees. We liked Random Forests for both models because they are less influenced by outliers, more resistant to overfitting, and still work with correlation among input features.

We again evaluated our model in two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy (Figure 7).

	precision	recall	f1-score	support
3.0	0.940	1.000	0.969	109
4.0	0.925	0.916	0.920	107
5.0	0.989	0.860	0.920	107
6.0	0.925	1.000	0.961	98
7.0	1.000	1.000	1.000	97
accuracy			0.954	518
macro avg	0.956	0.955	0.954	518
weighted avg	0.955	0.954	0.953	518

Figure 7

Second, we used 5-fold cross-validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.80694981 0.88223938 0.90926641 0.85521236 0.84555985].

One thing we noticed was that for our discipline model, the classification report and cross-validation numbers were the same (both around 89%), but our settlement model classification report shows a slightly higher performance (95% vs. 85%). This implies a small

amount of overfitting. However, since the cross-validation numbers are still quite high (between 80 and 90%), we still would say our model is successful in its ability to predict settlement money magnitudes.

While both models show good performance evaluation, these models are only as good as the data used to train and test them. Because we had to resample due to class imbalances, our data might not be truly realistic. However, these models provide good starting points for further explaining trends in the CPD relating to discipline and settlement.

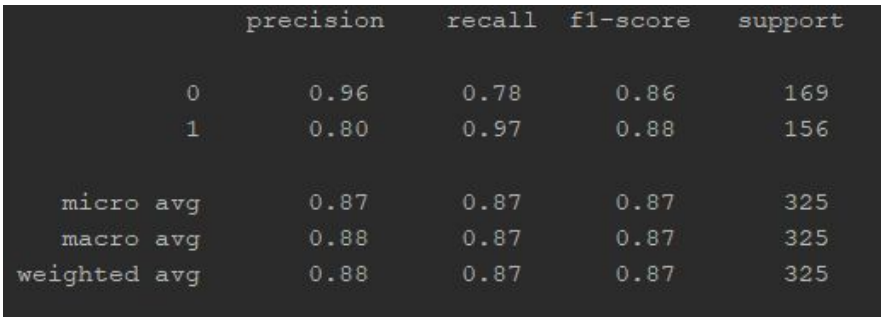
Can we predict verbal abuse and do certain cops use it more than others?

Lastly, we utilized natural language processing to classify verbal abuse, profanity, and offensive remarks in complaint report narratives. Then, we connected it back to our theme to see how verbal abuse relates to other officer characteristics.

We noticed many allegations are mislabeled. The summary may contain instances of verbal abuse, yet the misconduct category does not list it as such. This is because allegations only list the primary misconduct. This makes it more difficult to identify cases of verbal abuse since the data is not properly tagged. We manually went through the narrative summaries and identified keywords that imply verbal abuse. We then assigned weights to these keywords for how strongly they implied verbal abuse. For example, the f-word or n-word would have high weights since they are highly offensive terms. For each summary, we identified keywords that appeared, totaled up the weights, and marked it as verbal abuse if the total was above our threshold of 2. We set a threshold of two to prevent false-positive tagging (tagging something as verbal abuse when it is not). We did some manual verification to make sure the tagging was generally comprehensive.

After, we tokenized, lemmatized, and removed stop words. Then, we converted the data into TF-IDF feature vectors. Once we processed the input data into a more usable form, we split our data into training and test data sets using an 80/20 split. Then, we fit the training data with a Multinomial Naive Bayes classifier. Because Naive Bayes is a fast and scalable model, it made training and testing much easier.

We evaluated our model in two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy (Figure 8).



	precision	recall	f1-score	support
0	0.96	0.78	0.86	169
1	0.80	0.97	0.88	156
micro avg	0.87	0.87	0.87	325
macro avg	0.88	0.87	0.87	325
weighted avg	0.88	0.87	0.87	325

Figure 8

We also used 5-fold cross-validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.83435583 0.78703704 0.84876543 0.77777778 0.83024691]. We do see a bit of variation here with the CV accuracies, but the values are all still pretty high, around 80%. Thus, we can confidently say our model was a success and can predict with about 80% accuracy whether a complaint report narrative depicts verbal abuse or other offensive language.

There are some limitations to our data and model. First, many allegations don't have summaries, meaning there was less data to work with here. Second, while we tried to make our keyword list and tagging comprehensive, there is likely some human error. There are probably a handful of mislabeled cases. Third, Naive Bayes does assume independence between the features, but some words are more likely given the presence of another word.

Moreover, we connected our exploration of verbal abuse back to our theme. How does verbal abuse relate to certain officer characteristics? Do officers of higher or lower ranks think they can get away with verbal abuse?

For every summary for an officer allegation, we totaled the counts of verbal abuse, organized by rank. However, since some ranks are more popular than others, we normalized these totals by the frequency of a particular rank. The results are shown in Figure 9. One thing to note is that the high bars on the right are due to limited data on those ranks. But we can still see higher ranks like captain, sergeant, and lieutenant tend to have lower rates of verbal abuse than police officers. This relationship doesn't tell us anything about causality though; do individuals with more self-control or a better filter tend to get promoted to higher ranks? Something that surprised us was that the pattern among ranks here differs from our Figure 1 results (average disciplines by rank). In Figure 1, we see police officers, lieutenants, and sergeants had similar average discipline numbers. Here, we see police officers are higher than lieutenant and sergeant. However, these differences in the grand scheme of things are quite small − only a fraction of a percentage. So while there technically is a difference, the difference might not be practical enough to make any major claims.

Conclusion and Future Work:

In just a couple of months, we were not only able to learn new tools like Tableau and SciKit Learn, but also able to apply them to relevant issues on police misconduct. We were able to identify biases within the CPD in terms of how they handle misconduct, discipline, and settlement. Certain ranks are more prone to be disciplined than others, like police officer per arbitration award or 1st deputy superintendent. The misconducts use of force, verbal abuse, and supervisor responsibility cost the CPD the largest amounts of money in settlements, suggesting these are areas they should focus on for improving training, discipline, and culture. Certain units also cost the CPD lots of money in settlements, but these are units dedicated to violent crime or are in high-crime neighborhoods. Unit data also shows us that many districts are taking allegations less seriously over time, as we see the discipline to allegation ratio steadily dropping. This, plus the fact that higher disciplined units tend to have more awards, implies a

corrupt sense of values. Members of the CPD should think twice before rewarding or overlooking improper behavior. This only perpetuates the “Blue Wall” or “Code of Silence” among cops.

We additionally found that machine learning models let us predict discipline and settlement with high accuracy. They also revealed racial biases within the CPD, as black officers are disciplined more than other racial groups, and they have the highest settlement costs. We can also identify verbal abuse and offensive language in allegation summaries and see that higher ranks are less prone to verbal abuse. Thus, newer, less-experienced officers should be the target for reform. If inexperienced officers were better taught how to de-escalate situations avoiding offensive language or excessive violence, the department would save lots of money and public trust in the long run.

Due to time and data limitations, there were certain things we were not able to explore over the past few months. As previously mentioned, having access to rank history for officers would open up new opportunities to explore how rank relates to discipline and settlement. We also would like to spend more time tagging the complaint report narratives to improve the dataset for training. An improved dataset would make it easier to see how verbal abuse relates to other officer characteristics. Lastly, much of our research involved viewing discipline as a binary thing. An officer was either disciplined or not. However, it would be interesting to see if we could predict the severity of discipline as accurately.

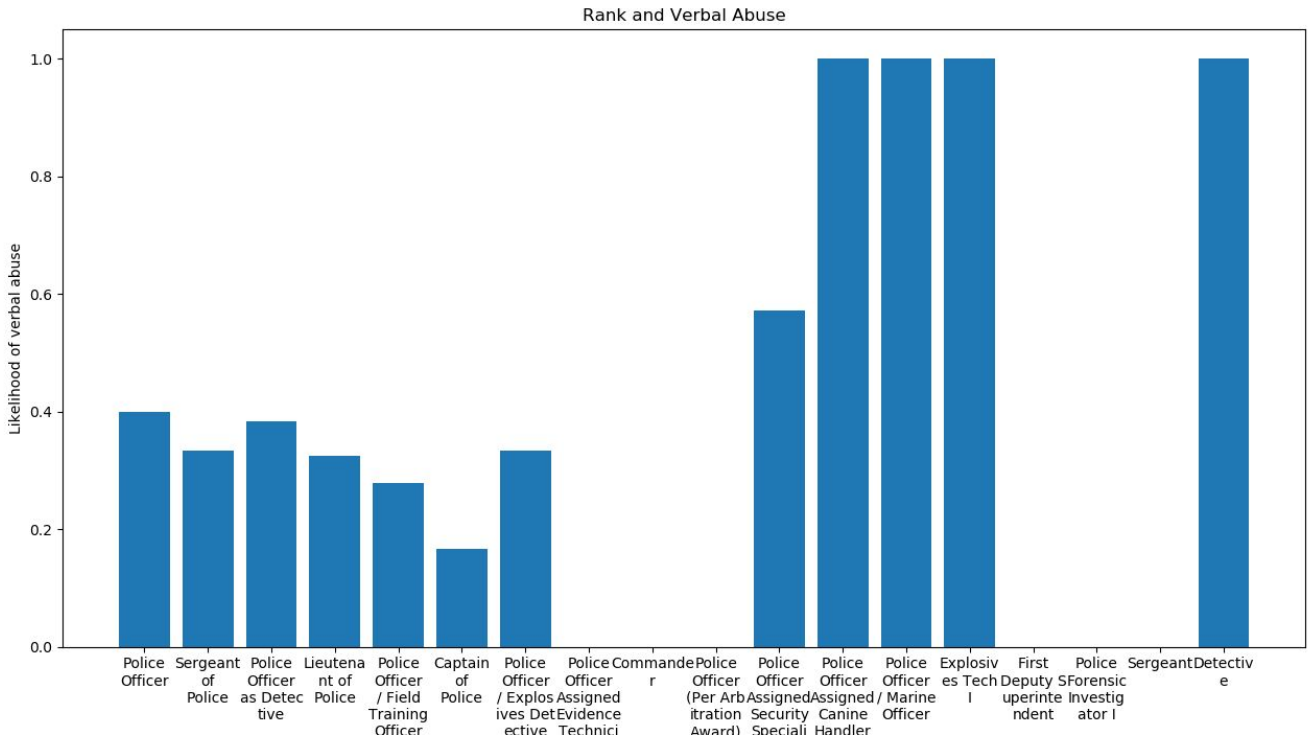


Figure 9