Are All Allegations Listened to the Same?

William Ansehl
Nicholas Easton
Krish Seth

Professor: Dr. Jennie Rogers, Northwestern University Department of Computer Science
December 1, 2020

Overview

Protests recently raged across the United States due to the tragic event of a white police officer - Derek Chauvin - kneeling on George Floyd's throat for 8 minutes and 46 seconds as other officers within arms-reach watched. No one can argue that this was an event that should not have happened. His name rekindled the fevered discussion on police reform and systemic racism in the country.

In light of continued effort to analyze the recently published police data and create insight on a significant issue today, our class and group have partnered with the Invisible Institute of Chicago. In examining tactical response reports (TRR), our group originally sought to explore the relationship between an officer's time spent policing a specific area and the bias they develop against the community they swore to protect and serve. However, time constraints and data limitations pushed us to explore another question - are all allegations listened to the same? In essence, are there discrepancies in the outcome of an allegation made against a police officer based on the complainant's race, gender, or location of residence? In our analytics, we aim to explore the potential for systemic racism, systemic sexism and discrimination against beat (an area which an officer polices) of residence.

SQL Analytics

Goals/Motivation

We began with an interest in analyzing the relationship between tenure policing a specific region and frequency of complaints. Namely, do officers who police one particular area over extended periods of time accrue more complaints than they used to? Is it possible that an officer becomes more confident and more reckless after years of arresting individuals in a particular neighborhood? In ethnically dense areas such as the south side of Chicago, it might make some sense for an officer to develop detrimental biases after arresting BIPOC for 40 years. To answer this question, we looked at individuals who had frequent transfers and their complaint history. Our SQL analytics aim to glean initial insights from the datasets provided by the CPDB database.

Methodology

We broadly had 4 queries that we attempted to use to gain this insight. These were: the total number of allegations per officer per year, the number of allegations per beat per year, total number of transfers per officer, and a breakdown of the total number of allegations per officer per year in terms of sustained vs not sustained (SU vs NS). We chose these queries because we believed an officer's timeline of allegations would be an important indicator of developing bias. The thought behind including the beat allegation data was to normalize individuals. If an officer spent a majority of their time working in neighborhoods with higher rates of allegations, either because of increased crime rates or some other factor, it might not be unusual to have a large number of allegations. But an officer who has more allegations than the beat average in which they serve is one we are interested in studying further. To this data, we included the number of transfers for each officer, the relationship we hoped to study. What we thought would be something of a side project, looking at the differences between SU and NS based on facts about the officer and complainant, would go on to become our main focus down the line.

Results

From this data we could begin to deduce some interesting facts. For example, we found that the officer with officer_id 1 held a constant rate of 1 complaint per year over a 10-year period. We can see that beat 1 had 3 complaints total in 2016. If from the officer's history, we know that this officer was assigned to beat 1 during 2016, then we know that the officer contributed ⅓ of the complaints beat 1 experienced. This particular example doesn't seem particularly drastic due to the low numbers overall. However, if we translate this analysis to beats that experience high rates of complaints and officers who experience high rates of complaints, then we can formulate an initial understanding of a particular officer's behavior over time compared to their peers in the same beat.

Alternatively, we can track an officer's overall behavior and its correlation with transfers. If we seek to understand if officers who transfer more do not develop bias, then we must understand the circumstances of transfers. Are they highly correlated with individuals who misbehave or are transfers representative of something else? From the data we collected, we see officer_id 604 has 0 allegations against them, but 27 transfers over the course of their tenure as an officer. This data gives us the ability to view officers who experience a high number of transfers and stack their complaint records against officers with low numbers of transfers. Now we can begin to understand if officers accrue more complaints if they remain in the same area for longer periods of time.

If we assume that sustained complaints indicate the guilt of the officer more than not sustained complaints, then we can isolate particularly troubled officers from the entire population for analysis. Similarly, we can compare rates of sustained to total or sustained to not sustained complaints across officers in the same

beat, officers who transfer in and out of a particularly troublesome beat, and track growth in sustained complaints for any particular officer of concern.

In sum, we believed the queries we constructed would allow us to analyze officer complaint records over a time series, compare records among peers, compare individuals who transfer units a lot to individuals who don't, understand transfers as a mechanism in policing and track sustained vs not sustained complaints across time. This proved far more difficult than expected, as we will continue to show throughout this report. There are quite a few tasks which, while easy to describe, are challenging to actually implement.

For example, transfers don't take place only on January 1st of a given year, they happen throughout the year. As such, keeping track of which beat an officer was assigned to over time, while simultaneously tracking the allegations was difficult. We struggled to construct a way of turning the raw calendar into something more manageable for further analysis. Another struggle we faced had to do with transfers. There seemed to be two competing forces that pushed transfer numbers up. The trend we hoped to capture would be officer's with bad behavior being transferred to hopefully reduce their beat's numbers and maybe to mollify those officers' behaviors under different circumstances. This is contrasted by officers doing their jobs well and getting promoted. Or maybe officers moving within the city so their children can attend a different school and thus want to work closer to home. There are so many reasons for a transfer, and the majority of the reasons we discovered were positive or neutral at worst. It was therefore difficult to disentangle the truly problem-prone officers from their counterparts that just did the job.

Static Visualizations

Goals/Motivation

Our original motivation for our visualizations was on exploring the relationship between the number of transfers and the number of allegations at the officer level, such as whether race, gender, or rank had a causal effect on this relationship. As such, we constructed our Tableau dashboard with six main visualizations. A useful way to traverse this figure is to start by looking at Figure 2A. When we find an officer of interest, we can then look at Figures 2C-D to see how their number of allegations and transfers compare to the rest of the officers in the CPDB. Next, we can move on to Figure 2B. Likewise, once we find an officer of interest, we can compare their allegations and transfers to the rest of the Chicago police officers by reading Figures 2C-D. This provides insight into if allegations and transfers are correlated. The visualizations are shown below.

Data and Methodology

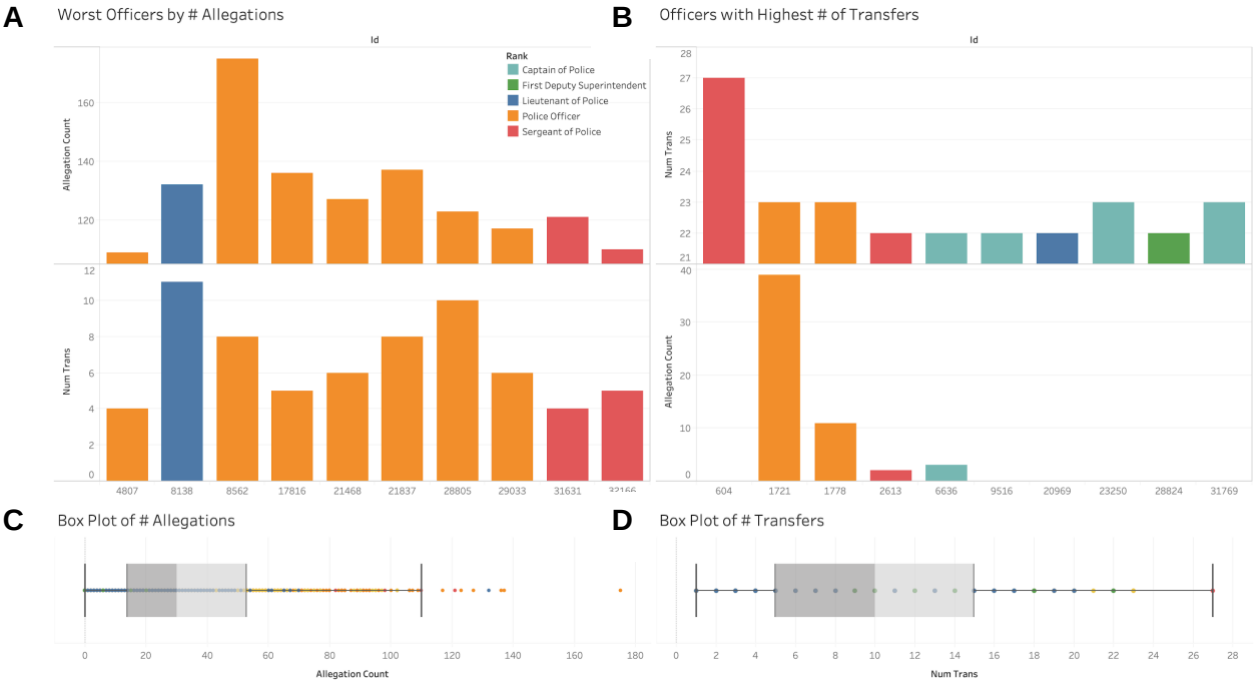

Figure 2A-D. (A) Officers with the highest allegation counts and their corresponding number of transfers, color-coded by officer rank. (B) Officers with the highest number of transfers and their corresponding allegation counts, color-coded by officer rank. (C) Box and Whisker Plot depicting the distribution of allegation counts across all CPDB officers. Data points are color-coded by officer rank. (D) Box and Whisker Plot depicting the distribution of total number of transfers per officer across all CPDB officers. Data points are color-coded by officer rank.

The dashboard is color-coded, whereby colors are indicative of a particular officer's rank. For example, blue indicates an officer who is of rank "Lieutenant of Police." Similarly, via hovering one's mouse over a respective officer's data point (or bar), the user can see the officer's gender (M or F).

Findings

From Figure 2C, we see that officers with higher ranks tend to have fewer allegations. The left-hand side is dominated by the ranks of Lieutenant and Captain. The right-hand side is filled with ranks of Police Officer and Sergeant. This is expected, as higher ranks have fewer interactions with the public in the course of their duties. Furthermore, this is likely a selection pressure within the department. We assume that officers with more allegations levied against them are more likely to be passed over for promotions. This also plays into the political nature of advancements; these high allegation count officers are more riskier promotions given the public backlash towards them we've seen in recent years.

While Figure 2A might suggest that officers with a high number of allegations tend to be transferred frequently, closer inspection suggests otherwise. These officers are sitting right around the median number of transfers, shown in Figure 2D. Furthermore, the officers with the greatest number of transfers, Figure 2B, are typically lower than average in their respective allegation counts. As such, these measures do not seem well correlated with each other. It seems that officers with many allegations levied against them do not tend to transfer at a higher rate. Similarly, officers who tend to transfer frequently do not seem to do so because they have been accused of wrongdoing.

Our current results indicate that the officers transferred the most are accruing fewer allegations than their peers. We originally aimed to further connect the relationship between allegations and transfers among individual officers and better understand the usage of transfers as a mechanism in police administration; however, we were met with difficulties in being able to analyze and visualize transfers at the officer level. As such, we shifted to exploring how allegations were being handled for each beat, either sustained or not sustained.

Interactive Visualizations (A New Hope)

Goal/Motivations

After we decided to shift our focus, we constructed a choropleth heat map showing the number of allegations per beat, as well as a stacked bar chart showing the number of sustained and not sustained allegations per beat over time. These interactive visualizations can be viewed by clicking on their corresponding links.

Data & Methodology

Allegation Map: To create this visualization, we began with a Geojson containing only the polygons of beats and a json with the number of allegations in a beat for a given year. From there, we filtered allegations based on the year, which is selected via an html drop down field. This allowed us to join the polygons and the corresponding counts for a year, scaling the color by the number of allegations. There is another layer of polygons which are the neighborhoods. Unfortunately, beats are not confined to neighborhoods, so the mapping isn't perfect. It does allow us in broad strokes to compare neighborhoods though. We chose to not pin the maximum as the maximum over all years, and instead let the color scale vary according to the year. This makes it harder to compare the numbers of allegations across years, but it makes it significantly easier to compare relative relationships.

Stacked Bar Chart: To create this visualization, we utilized a csv file consisting of 4 columns: the year, the beat, the number of non-sustained allegations in that respective beat and year, and the number of sustained allegations in that respective beat and year. From this information, we filtered the allegations based on the beat we aim to further analyze. In the visualization, the user can toggle the sliding scale to select the specific beat they aim to analyze. From there, the user can see the number of sustained complaints in comparison to the number of non-sustained complaints. For example, for beat 165 in 2009, there were 6 sustained allegations and 2 not-sustained allegations. This ratio clearly flips in 2011, in which there is only 1 sustained complaint in contrast to 5 not-sustained complaints. Data can be visualized upon availability. In the case of beat 165, there is no data in 2010 that provides the number of sustained or not-sustained complaints.

By itself, this visualization makes it easy to analyze trends of sustained and not-sustained complaints within a single beat. Similarly, it is easy to compare two beats and their number of sustained and not-sustained allegations over time. This visualization, however, lacks in geo-spatial awareness. Beats are an intrinsically spatial data type. The beat number refers to a specific area in Chicago. As such, it is impossible, from this visualization alone, to understand proximity between beats. Similarly, it is impossible to compare neighboring beats in their trends over time. For this reason, this visualization pairs nicely with the interactive choropleth map introduced earlier in this paper.

The choropleth map gives a user an "at a glance" understanding of beat proximity and allegation hotspots over different years. Utilizing the choropleth map, a user can narrow in on a particular year, identify a hotspot of interest, identify neighboring beats, and ultimately turn to the stacked bar chart for further analysis. For example, in 2009, beat 261 is a hotspot of allegations. We can also identify that its numbers - beats 217, 253 and 257 do not seem to be as "hot." Now we can proceed to the stacked bar chart for trend analysis.

Findings

Over time, the number of allegations resulting in sustained or not sustained consistently goes down, from peaks over 600 in the year 2000 to peaks of 60 in the year 2018. This is consistent with the findings outlined by the Invisible Institute, where in recent years the number of allegations has been decreasing. As the maximum number of counts decrease, the deviation also decreases. The allegations are considerably more evenly distributed in 2016 than the early 2000's. Interestingly, there is a single hotspot right in the center of the city, right around the loop. This beat (#129) is persistent over the years, generally presenting as the brightest point on the entire map.

As mentioned previously, beats don't necessarily fall within a single neighborhood. This is frustrating as we would like to attribute allegations to a single neighborhood. If we aggregated in a preprocessing step this would be doable, however, it would be preferable to track both beats and neighborhoods at the same time. As it stands, we can only guess at where the allegations fall. One would expect that neighborhoods represent sufficient granularity to capture trends across the city as they present natural borders. From this map, we gain the understanding that some neighborhoods have considerable variation. This is more pronounced in the south where neighborhoods are generally larger. It perhaps suggests that even beats are too large to accurately describe the underlying distribution.

In 2009, beat 261 had approximately 15 sustained allegations and 18 not-sustained allegations. In 2010, beat 261 had approximately 14 sustained and 5 not sustained. And so on. In contrast, beat 257 had approximately 2 sustained complaints and 14 not-sustained complaints. Beat 257 also experienced 7 not-sustained complaints despite 0 sustained complaints. And so on. Clearly, the ratio of sustained complaints to not-sustained complaints this area tells very different stories according to the beat you are looking at.

We can also compare non-neighbor beats to understand how ratios might change across the city landscape. For example, beat 170 is a relatively low allegation count beat in congruence with its neighbors. In 2009, there were only 6 not-sustained complaints and 0 sustained complaints. The number of allegations only continued to decrease as time progressed.

Machine Learning

Goals/Motivations

Our approach towards machine learning was twofold. We began by attempting to reason about patterns within the allegation data using forms of time series analysis. The same problems that had been plaguing our project up until this point reared their heads once again. Poor results and the difficulties of working with data that was dependent on time made this analysis challenging. Because of this, we began to alter our approach to the data and what we would focus on. Instead of discerning the relationship between allegations and transfers, we would try and determine what about an allegation made it more or less likely to result in a finding of SU. We will first discuss the approach to the time series analysis and some of the specific difficulties that we ran into, and then talk about our approach to determining which allegations are more likely to be listened to.

Linear and Polynomial Modelling

We began with a similar set of data as the allegations per beat per year from SQL Analytics. That is, a table in which we have a timeline of all allegations in a beat, over the entire city. Using the data as a (time stamp, allegation) pair doesn't work because that data is incredibly sparse, most days would have 0 allegations. Instead, we tried to reformulate this into a space where the data wasn't sparse, such as looking at the time between allegations. Instead of data points being the day an allegation took place, we looked at the time between allegations. Doing this also allows us to free ourselves from the calendar, as we can look at time as being days since the first allegation rather than a date.

Methodology

Our first pass at analysis learned linear fits to randomly selected beats. This wasn't the best decision for a few reasons. First, because of the way we created the data (using wait time) it is concentrated in toward the beginning of the time span, when there are a lot of allegations. However, as we progress in time, the wait times become larger and whatever potential predictive power we had is lost. We should expect a lot of noise, but none of the models manage to do much better than randomly guessing.

Findings

Another issue to note, since these are linear models, the wait time will continue to grow forever. We'd like to live in a world where the time between allegations is extremely long, or even nonexistent, but that isn't realistic. Instead, we'd expect our model to show some approach toward an equilibrium. In an attempt to achieve better results, we decided to forgo linearity (which is an admittedly strong assumption we can't make) and try a polynomial fit. We tried tuning the polynomial order as a hyperparameter without huge success. These models did better than their linear counterparts, but still managed meager performance. Neither these models, nor the linear ones manage to have predictive behavior we want. As soon as we exit the data and extrapolate, we begin to see wild results. As seen in 6 of the 9 models visualized, predictions turned toward negative wait times, a nonsensical result. We tried to be clever and avoid issues of sparsity and did. However, instead of eliminating problems, we just traded them out for other ones.

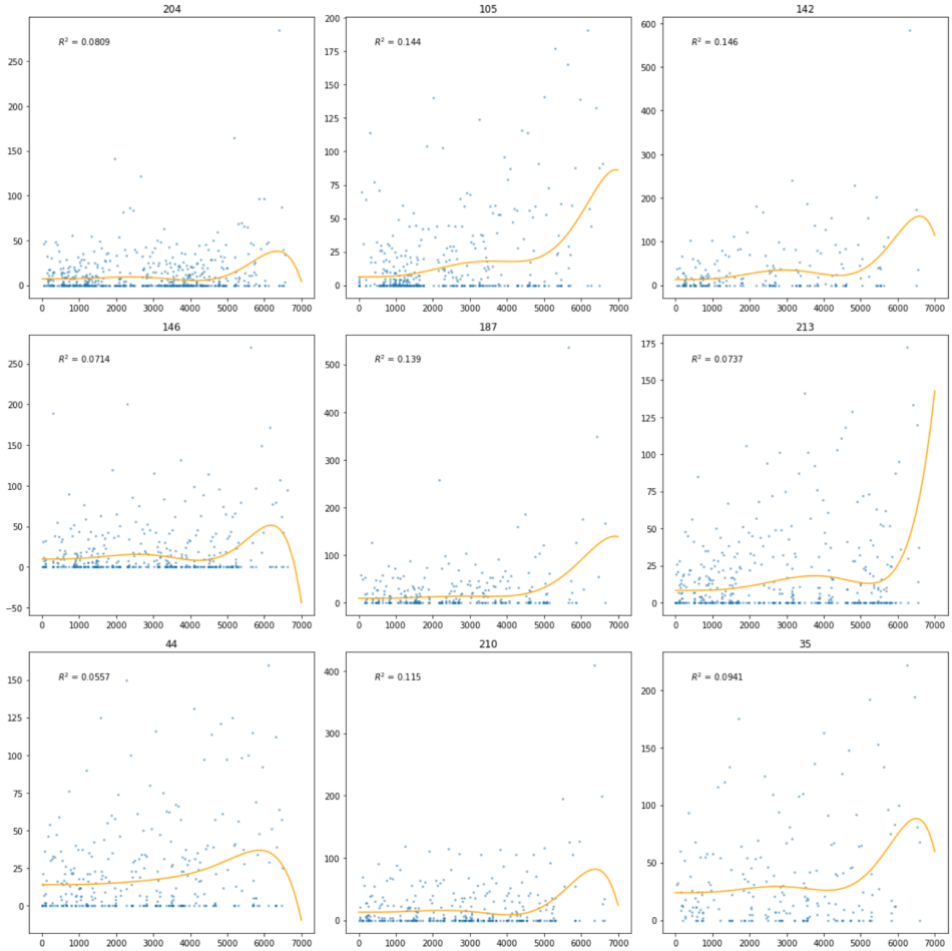

Figure 4A: Polynomial models learned from individual beats, randomly selected. None of the models here performed well and many produced wildly inaccurate predictions.

Our next attempt focused on preventing the issue of extrapolation. Instead of passing a single feature as input, we created a vector of the last 5 wait times and used that as our feature vector. This allowed us to "generate" more data from a single beat. For this analysis, we only did 1 beat, however it is the same as the beat in (1, 1) from the previous analyses. By applying this windowing function to our linear fit, we achieve much better results. This approach is akin to using a moving average as the prediction. It produces a noisier plot, but the predictions are much better. Furthermore, because we include some history as context for each data point, this model should do better in the long run. It doesn't appear like it has any of the behavior problems exhibited by the previous fits. The early days were well fit while later time periods didn't fare as well. In this model, the predictions seem to do equally well across the entire time period.

We attempted to improve upon this analysis by training on the entire city rather than a specific beat. To do this, we broke every beat down into windowed data vectors with the appropriate label. Doing this caused us to lose information about the beat it came from. While this should improve generalizability, the beat we've been using to visualize has poor performance. This larger training set does smooth out the predictions.

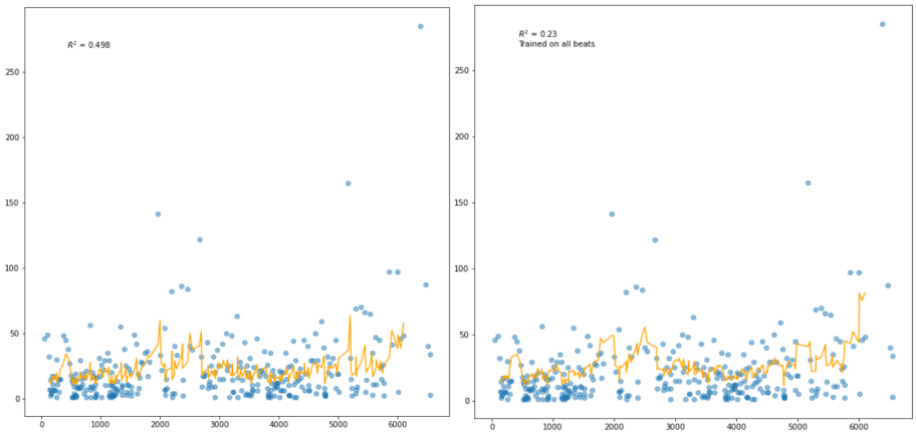

Figure 4B: Left the windowed linear fit trained only on a single beat 204. Right, this model was trained on all the beats in the city and with beat 204 withheld for testing.

Logistic regression of Outcomes

We begin our new implementation with a new dataset comprising the year, beat ID, allegation ID, gender of the officer and complainant, race of the officer and complainant, and the final finding for the particular allegation. In particular, cases were limited to allegations that resulted in NS or SU. Furthermore, cases that had missing information were removed, as well as duplicates. These actions resulted in a dataset of 9,173 rows.

Findings

Since most of these are nominal (categorical) variables, and not ordinal, a series of indication variables were used to indicate the presence or absence of a value. These substitutions were made because unlike ordinal variables, nominal variables don't have any relation from one value to another. After dropping the column for allegation IDs, the data frame had 292 columns. In order to balance the dataset such that there were equivalent cases that resulted in sustained findings to that of not sustained, we down sampled the cases that ended in NS. These actions led to 2,811 examples from each class, each having 292 features. From this dataset, we implement a logistic model. The log model relies on the sigmoid function to output a probability and classify inputted data as one of two outputs (in the case of binary classification).

In conducting dimensionality reduction, our group ran into issues with variable significance. Specifically, every variable was computed as being statistically insignificant. This is counterintuitive and impractical to the aim of our analysis. As such, we proceeded with no dimensionality reduction. That said, we

can analyze the relationship of each variable to the outcome SU/NS. For example, Complainant_Race_White has a strong positive relationship with a complaint being sustained. Alternatively, Complainant_Race_Black has a negative relationship with a complaint being not sustained. There are certainly potential confounding variables that have not been included in the model, but we did seek to account for basic ones such as complainant gender, beat of alleged origin, and officer biographical information. Further analysis is needed to determine if this can lead to conclusive evidence as to the role race, gender and location play in a sustained vs not sustained allegation. For example, the allegation type is another potential confounding variable.

Other findings included a positive relationship between certain beats such as beat 2 and 3 and an allegation being sustained, as well as a negative relationship between beat 1 and final allegation result. Asian/Pacific Islanders had a positive relationship with sustained finding while Native American/Alaskan Native and Hispanic had negative relationships. Further correlations can be found via indexing the regression results.

The result of our model can be visualized with a confusion matrix and an accuracy report. In the confusion matrix, our model performs relatively equal with classifying true positives and true negatives. The algorithm correctly predicted approximately 38% of the data as not sustained and 38% of the data as sustained. Consequently, approximately 24% of the data is misclassified. Both sustained and not sustained classes have decent precision, recall and f1-scores (mid 70s). Compared to other classification models, the logistic model also performed the best in terms of accuracy

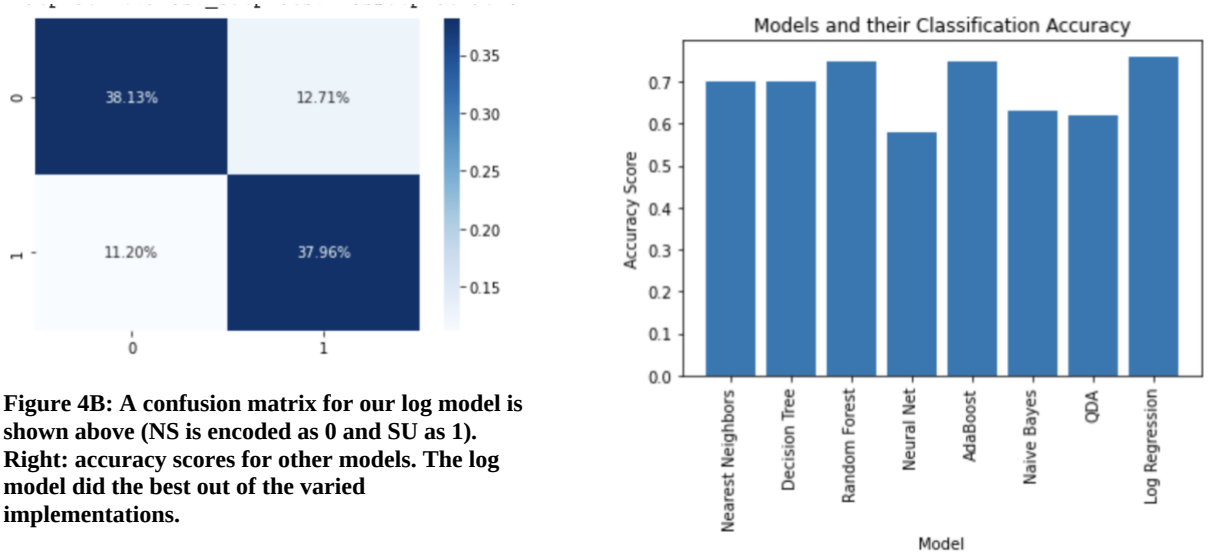

Figure 4B: A confusion matrix for our log model is shown above (NS is encoded as 0 and SU as 1). Right: accuracy scores for other models. The log model did the best out of the varied implementations.

Our logistic model can be used to predict the likelihood of a case being sustained or not sustained based on certain features of the case. Consequently, variable analysis can indicate bias and room for growth in the response to certain allegations. More specifically, allegations made by black complainants might experience negative bias while allegations made by white complainants might experience positive bias. As such, black complainants ought to receive further attention equivalent to other races. This is one example of a conclusion we can draw from the log model.

Natural Language Processing - Sentiment Analysis

Goal/Motivations

In prior analyses, we explored the relationship between ethnicity, gender, and geographical region to the outcome of an allegation report. Conclusions drawn include a negative relationship between a sustained outcome and the complainant being black. Alternatively, there existed a positive correlation between a sustained outcome and the complainant being white. Despite these results, we hesitate to make the claim that “systemic

racism exists in the response to allegations." The legal system is complex and even more so considering that individuals who participate might have their own inherent bias. To explore this facet is an entirely different problem deserving of its own due attention. However, it is possible to explore signals within these TRRs that indicate the prevalence of discrepancies in how justice is delivered.

One such signal might be the relationship between the sentiment of the summary report of a given allegation and the final finding (sustained vs not sustained) of the allegation. More specifically, is there a disconnect between the sentiment of a summary and the final finding, relative to certain groupings of individuals divided by ethnicity, gender, and location? In this section, we aim to answer these questions via exploring allegations that result in sustained filings vs allegations that result in not sustained. We will analyze the sentiment scores of the summary reports and evaluate discrepancies between sentiment scores and an allegation's outcome. Distributional differences between sustained vs not sustained outcomes and sentiment scores may potentially serve as an indicator for systemic issues underlying the delivery of justice.

Data (SQL Query)

Our dataset had 719 rows and consisted of 4 columns:

1. Allegation_id: the id of the allegation
2. Summary: the summary report on the TRR
3. Beat_id: the beat an allegation was lodged from
4. Final finding: the outcome of the allegation

Cleaning the dataset involved limiting cases to 'not sustained' and 'sustained' (NS or SU) and removing null values and duplicates. In total, there were 554 'not sustained' complaints and 165 sustained complaints. The not sustained complaints were down sampled to promote a balanced dataset for sentiment analysis.

Sentiment Analysis

The library implemented for sentiment analysis was NLTK. Specifically, only the polarity score compound value was tracked. Each summary was fed into a sentiment analyzer and the results were plotted. Each label on the x-axis corresponds to a bin of compound values. 'Very Negative' corresponds to scores between -1 and -0.6. 'Negative' corresponds to scores between -0.6 and -0.2. 'Neutral' corresponds to scores between -0.2 and 0.2. 'Positive' corresponds to scores between 0.2 and 0.6. 'Very Positive' corresponds to scores between 0.6 and 1.

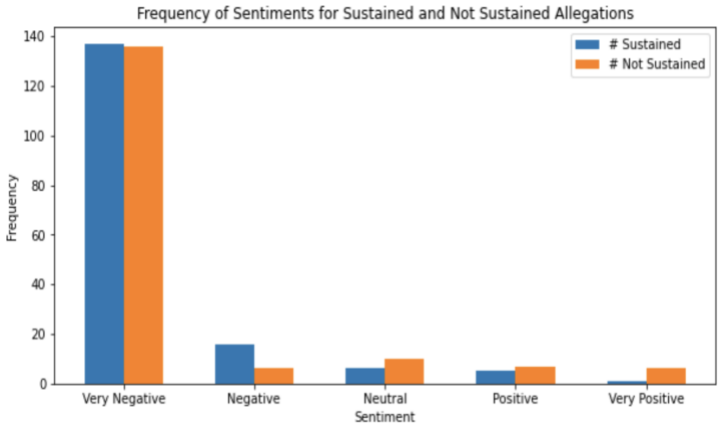

Figure 5A. Depicts the sentiment distribution for allegations that resulted in 'SU' (sustained, blue) and allegations that resulted in 'NS' (not sustained, orange).

Our group then proceeded with two methods of cleaning the summary reports. We believed the NLTK sentiment analyzer would be more accurate with cleaner summaries, resulting in more representative

distributions. Our first summary cleaning method implemented syntactic parsing. We filtered each summary and kept only the adjectives and verbs, words we believed to more strongly correlate with the sentiment of the given summary report. This method removed nouns such as "October," "2000" and also prepositions. The resulting bar chart can be viewed below.

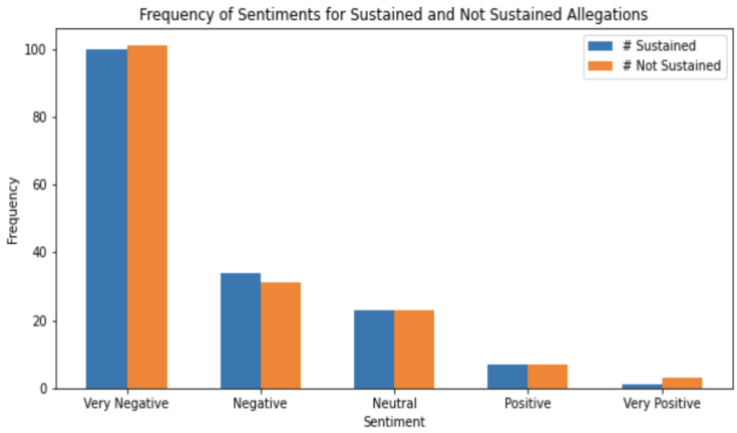

Figure 5B. Depicts the sentiment distribution for allegations that resulted in 'SU' (sustained, blue) and allegations that resulted in 'NS' (not sustained, orange). In this case, the summary reports were cleaned to include only adjectives and nouns.

Our second method of cleaning the summary reports involved lemmatization. Similar to syntactic parsing, we kept only adjectives and verbs. The results of conducting sentiment analysis on the lemmatized text can be viewed below.

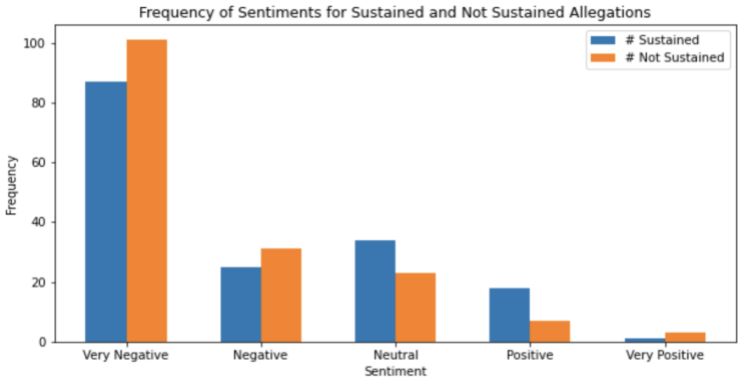

Figure 5C. Depicts the sentiment distribution for allegations that resulted in 'SU' (sustained, blue) and allegations that resulted in 'NS' (not sustained, orange). In this case, the summary reports were cleaned via lemmatization.

In the next section, we will discuss the results of our analysis.

Findings

In Figure 5A, we see that the distribution of sentiments for sustained and not sustained complaints are roughly the same. There appear to be more negative sentiments for 'sustained' complaints and more positive sentiments for 'not sustained' on a grouping-by-grouping basis. However, the differences are not that drastic, and it is hard to say if these results are statistically significant enough to warrant such claims (especially given the low amount of data). Surprisingly, there appear to be allegation summary reports that fell in the "Very Positive" bucket. Such reports are more likely cases where the sentiment analyzer failed to properly grasp an accuracy sentiment metric for the given information.

Syntactic parsing also failed to further distinguish the populations, as seen in Figure 5B. Both of the distributions for 'sustained' and 'not sustained' complaints shifted more to the right. More complaints registered as 'Negative', 'Neutral' and 'Positive' than before. Consequently, fewer complaints for both distributions registered as 'Very Negative.' However, the cases we analyzed as "Very Positive" still appeared to be rather negative, signifying fault in the sentiment analyzer once again.

In Figure 5C, we see that lemmatization seemed to more affect the sustained complaints' distribution rather than the not sustained complaints distribution. The sustained allegations' distribution experienced another right shift towards the direction of more overall positive sentiments and fewer extreme negative sentiments. Regardless, the overall distribution of the two populations is still rather similar. It is still hard to say if there is any significant difference between the two.

Unfortunately, due to a lack of data, it doesn't make statistical sense to further divide the data and analyze the resulting distributions based on the gender and race of the complainant or officer, or the beat of origin. These actions would further reduce the amount of data we have to work with for analysis.

Ultimately, we found that there is very little distributional difference between the sentiments for sustained vs not sustained allegations. This implies that there is no correlation between the sentiment of an allegation and the likelihood of the outcome being sustained. Until more data is collected and/or cleaned for analysis, it is not possible to create conclusive statistical claims about discrepancies between sentiment of a summary report and allegation outcome based on race, gender or geographical location.

Conclusion & Limitations

Our analysis demonstrates the existence of relationships between race, gender and location of residence to the outcome of an individual's allegation.

1. There exists a positive correlation between a white complainant and a sustained outcome
2. There exists a negative correlation between a black complainant and a sustained outcome
3. There is a positive correlation between a black police officer and a sustained outcome
4. There is a positive correlation between an Asian/pacific islander officer and a sustained complaint
5. There is no correlation between sentiment of a summary report and likelihood of an outcome being sustained or not sustained

Despite these claims, there certainly exist limitations in our work, and room for improvement.

1. Despite our efforts, time series modelling of this data met with disappointing results. Even training over the entire city did little to increase performance.
2. None of the variables in the logistic model appear to be that significant. These correlations above exist, but there still is a question of how relevant they are.
3. A lack of summary data limits the insightfulness of our sentiment analysis. Similarly, we are unable to further expand our understanding into specific demographic relationships with summary sentiments since there isn't that much data available.
4. Our models and visualizations mapped data to beats. However, beats don't necessarily fall within a single neighborhood. This problem complicates our ability to attribute allegations to a single neighborhood. We can only guess at where the allegations fall by neighborhood.

Further work includes added model tuning for better classification results and feature selection. There are other methods for time series analysis we could attempt, such as true convolutions with a CNN or a Gaussian process type model. Additional data collection and/or cleaning is required to expand our sentiment analysis model. Perhaps added summary data can be sourced from lawsuit datasets for further sentiment analysis.