Machine Learning

1. Quantile dot plots: When discussing quantile dot plots, I reference this code about making them. Pulling from their paper, Matthew Kay, Tara Kola, Jessica Hullman, Sean Munson, created this tool to think about distributions of event likelihood. I am able to explore the question of officer density in areas. I was curious if we could predict or better understand the range of officer density at the district level within a day? Using quantile dot plots, I will predict/estimate how many officers are in a district on any given day using the number of officers in relation to the population size, district area, and time spent.

The first part of this question is to organize the data around officer density. I am thinking about density across space and time within the scale of square miles of Chicago and time within a general sense of hours worked. I used SQL to create a table that had distribution for each of the police districts. What I saw was that district populations ranged from 59,458 to 247,373 residents. The number of officers ranged from 271 to 451. The size of districts varies from 4.1 square miles to 34.4 square miles. These were not entirely correlated, or at least the largest district area was not the district with the most officers nor residents in the district. The rate of residents per square mile ranged from 4,094.73 to 23,772.5 residents per square mile. There were between 9.36 to 80.14 officers per square mile. Finally, I saw the range of average officer hours worked in a year to go from 360,101.77 hours to 698,876.57 hours

From these values, I broke down the hours worked into each day which ranged from 986.6 to 1,914.7 hours in a day per district. For instance, this means on a given day district 1 has 1,403.83 hours worked across their on duty officers for that day. I assumed shifts would be around 8 hours based on past explorations with the data, so I could estimate the number of officers working each day as the number of hours worked across all officers in a district divided by 8 hour long shifts. The estimated number of officers working per district on a given day ranged from 123.32 to 239.34 officers per district. This seemed to be less than half of the total number of officers within a district, which makes sense given days off from work and differing schedules across police officer employees. Taking district 1 as an example again, what this tells me is that across the district there are about 175.4 expected officers working (8 hour shifts) on any day in the district. This calculation does not take into account holidays, large events, or other unique occurrences, but it does provide a sense of how many officers are working at a given time in the district. My next thought revolved around how those individuals might be spread out through the district. I standardized the expected number of officers working by the area of the district to calculate the expected number of officers per square mile within a district. The number of officers within one square mile (given a specific district) ranged from 3.8 to 41.98.

Notably, the 1st, 11th, 18th, and 15th districts have the highest number of officers per square mile on a given day from 34 to 41 officers per square mile. These districts were also the ones with significantly higher officer per 10k residents rates. The potential significance of these rates is to be seen when putting this into perspective. Consider that Northwestern campus is 231 acres which translates to one-third of a square mile. Or, we could say that 1 square mile is almost 3 times the size of Northwestern University campus. Then imagine 41 officers in that area or take ⅓ of 41 to get about 13 officers to be on Northwestern campus. And consider that

the surrounding areas would have similar rates of officers. How does that ratio seem? You can compare this theoretical ratio with the case such as district 17 with 3.8 officers per square mile, which would mean 1 officer for all of Northwestern campus, and more officers across neighboring square miles of areas.

In my example, I am imagining that the number of officers per square district is evenly distributed and independent. This raises interesting questions about the actual distribution of officers at the level of beats and how some areas might have more attention than others. It is interesting to consider how some police districts may be distributing officers based on their area and how that might be influenced by histories of events in those areas.

I tried to capture more of this information within the quantile dot plot. I used the idea of quantile dotplots to explore police presence in Chicago. Within a R markdown notebook, I saw the potential distribution of the number of officers per square mile within a district on any given day. Given the data, I described above, I saw that the values had a mean value of 21.6 officers per square mile with a standard deviation of 10.77. Inputting these values into the R code. I could see the theoretical distribution of values and their frequency of occurrence. It ended with the creation of the following quantile dot plot. The purple line represents the number of officers I should expect in any Chicago police district on a given day to have about 18 officers across a square mile in a district throughout the day around 10% of the time.

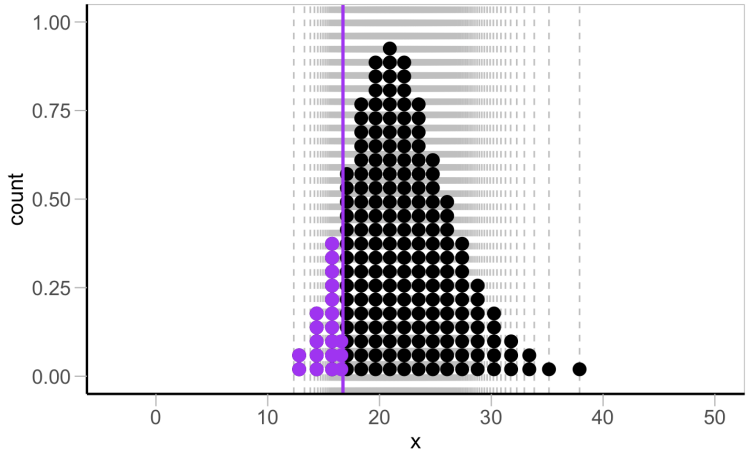

Fig 1: FINAL PLOT FROM R CODE TO SHOW EXAMPLE (above)

This could be used as a theoretical interpretation of how often people could expect to see officers if their routes are evenly distributed across all spots in the district. However, it might be interesting to explore how these theoretical probabilities differ from the actual way that officers are distributed throughout their district.

2. From part 1, I am exploring how officers might be distributed across a space over time. I am then curious how might officers be spending their time. Although we do not have an in depth look into where they were, a dimension of the time they spent in the district is reflected in the number of allegation counts they received. I will use a decision tree model to predict whether officers received at least one allegation in their time based on their average number of hours spent on duty in a given year.

For this question, I did several data manipulations in SQL before transferring the data into a csv file and uploading it to a google COLAB space for the machine learning aspect. In SQL, I started by calculating the number of hours worked by each individual officer and getting the shift lengths across all officers. I joined information across different tables to get the officer allegation counts. With this I had a table about each officer and the number of allegations they had, the number of hours worked by the officer in the year, and included a few additional features about their race, gender, and birth year.

Within the COLAB notebook, I read in the data, and plotted the distribution of allegation counts and the hours worked by an officer in a year. I decided to go with a binary output and created a column to show if officers had at least one allegation count. After a small bit of data cleaning, I used this as my y value to predict with my model. I used a decision tree model to predict whether officers received at least one allegation in their time based on their average number of hours spent on duty in a given year. The original events for the y value are split with 27% being false and 72% being true. The accuracy of the model on the train set was 74.2% and on the test set was 75.02%, which seems better than just predicting true for all individuals, given that I stratified the dependent variable when doing the train and test split. I plotted the decision tree to see where there were specific breaks across the values. The first break was in the total hours worked being more or less than 2182 hours. To put this number in context, if someone is working 40 hour weeks for 52 weeks of the year, then they will work 2080 hours in that year. In this case, one of the interesting noticings is that individuals are working more than 40 hours per week to get totals per year over the expected amount of 2080, this is aligned with the likelihood of having at least one allegation going up.

I followed up this decision tree with a random forest classifier that had a randomized cross validation search to find the ideal combination of features. The best hyperparameters for this model were found and the accuracy score was 72.7%, which was less than the simple decision tree model, which might be a reflection of the few amounts of features that went into the model. Thus, I consider the implications of the decision tree over the random forest model. Overall, what this model told me was that officers without any allegation seemed to work fewer than 2180 hours most often. However, it also seems that those without any allegation also had more than 559 hours worked in a year but less than 1815.95. This shows that part time working individuals may be more likely to not have any allegations. This also might explain an aspect of police presence being that the more time they spend in the district then the more likely they are to have at least one allegation. However, those who receive an allegation with a few hours may be very new individuals to the force.

Potential future directions of analysis include adding features of race, gender, and birth year into the algorithm to get a sense of the impact of those features on whether officers will receive at least one allegation. However, this is not currently central to my analysis, since in this checkpoint, I was heavily concerned with how officers were distributed across areas and their interactions with residents as potentially captured in their allegation counts.

THEME CONNECTION

Overall, these explorations work toward my theme of understanding police presence by giving a better sense of how police officers might be distributed throughout a district with regards to time and to area sizes. I can take this a step further and consider it in relation to the

number of residents per square mile. If I think about the number of residents per square mile in a district and compare that with the number of officers per square mile during a day, then I can get a sense of how often people might expect to see or interact with police officers. This can have implications for how people go about their days if they are worried about police interactions. For example, we can consider district 15, which has 41.98 officers per square mile and 14,180.8 residents per square mile. Then, if we have a ratio of these two we get 0.003 as an estimator for the chance that people may run into each other on a given day if we assume officers are always somewhere where there is at least one resident. This assumption is not so far stretched considering part of the purpose of police patrols is to go where people are to provide their services. This means many people will go days without seeing an officer, which seems likely in theory but it raises the question if there are areas where officers patrol more often and how that might impact their interactions with residents. This might also explain some of the differences in allegation counts across officers, if we are to consider officers as not being evenly distributed across the space and likely going to specific places more frequently. It could also be coming from differences in the distribution of race and gender across officers.