Liqian Ma
Wentao Yao
Xingbang Liu

Project Report

Introduction

Misconduct analysis in terms of different locations and communities can be valuable. Is there over-policing in low socioeconomic status neighborhoods? We could compare the low-income area data with high in-come area data. The income of the neighbor could be a factor to influence the "victim" narrative (complaint report). We plan to dive deep into the relationship between location, income level, and police misconduct.

Checkpoint 1

In this checkpoint, we are investigating the whole database by answering a few questions by SQL.

For starting learning the database, we would like to find the TOP5 richest and lowest income neighborhoods and their CRs and TRRs. In this checkpoint, we find that the top 5 richest neighborhoods are Forest Glen, Lincoln Park, Loop, North Center, and Beverly Their income ranges from $89,038 to $101,237; they have 215, 893, 4288, 927, and 482 CRs; and 291, 965, 481, 118, and 566 TRRs; The poorest neighborhoods are Riverdale, Fuller Park, Englewood, East Garfield Park, and Washington Park. Their income ranges from $14,916 to $21,869; they have 261,1439,2490, 2824, and 876 CRs; and they have 1919, 2817, 3454, 231, and 2887 TRRs. Later, we also studied the percentage of each race in the community and the top 5 streets in allegation count for each beat area for better understanding the problem.

Checkpoint 2

To further explore our theme and inquire into potential conclusions to our theme, we tried to visually represent our data in plots. Visualizing something in the real world like a map is always attractive and easy to understand. So, we created an income map for all the areas in Chicago. Darker color means higher income in the region. This way, we can see the income level clearly in all areas, beats, and even streets. It also opens the door to a more thorough analysis of our analysis of income and community.

As we can see from the income map (Figure 1), areas with a similar level of income usually come from nearby areas. For example, high income in the northeast and southwest areas and low income in southeast areas. Also, different areas are drastically varied in their median income. The largest number gets as high as 100k, but the lowest number can be only over 10k. Typically, lower income means more crimes in the area. When there are much more crimes in any area, there must be more possibility of over-policing. Furthermore, we can find out more details by correlating the income map with over-policing. So, we created the complaint rate map, which indicates direct correlations with low income as expected.

We believe the complaint rate in each area is positively related to the rate of over-policing. Based on this hypothesis, we need to investigate the relationship between locations, income, and complaint rate. So, we calculated the count of tactical responses in each area and every over-policing incident that happened in the past as a dot on the map (Figure 2). Combining this complaint rate map and the previous income map, we can see a high complaint rate in the low-income area. For example, northeastern Chicago, a high-income area, has the relatively lowest tactical response rate. In conclusion, if the complaint rate is positively related to over-policing, we can say low-income is related to a high rate of over-policing.

In our proposal, we planned to generate a heat map of the officer hour in different areas. However, there is no information about officer hours associated with the area. We would like to have such a map since if a place has significantly high officer hours, police are focusing on such an area and they may use over-policing to ensure security in that area. To achieve the same goal, we believe the heat of attendance rate associated with beat id can have the same function.

Figure 3 shows the percentage of attendances in different beats. For example, we put our view on the south area. The attendance rate in some low-income areas is near 90%. This is good, since officers here are willing to work, if there is no evidence of over-policing, then there is not. However, there are some low-income places that have a low attendance rate. If we also find there is a high volume of CRs, we should consider there is potential over-policing.

Checkpoint 3

In this part, we made use of “d3.js” to visualize dynamic data like time series to look into the trends.

Analysis 1: Highlighting the high and low socio-economy status communities with different colors and plot TRRs on them. Set up a time slider to see how it changes over time.

Our theme is about the relationship between geo locations, eco-socio status, and over-policing. In our previous analysis, we have seen a positive correlation between low economic communities and high complaint reports (which indicates more misconduct behaviors). In this report, we would like to further look at key metrics(number of CRs and TRRs) in different communities with different economic social statuses. More importantly, we want to monitor the changes over time, and locate the trend in recent years.

So, we created the plot with multiple dimensions in years, beat areas, median incomes, the number of tactical response reports, and the number of complaint reports (Figure 4).

As you can see, low-income areas (yellow bubbles in the plot) are generally having more CR and TRR reports. High-income areas, despite 2 outliers typically have fewer. From the time perspective, we have seen a reduced increasing speed for both records over the years for most of the beat area. Especially for the "rich" area, most of them stayed around the origin point.

From this chart, we can find there are multiple shallow colored points(low-income community) on the top right corner, which means for these areas, police are receiving a significantly high amount of complaints and they would prefer to use tactical weapons against the citizens. Therefore, we can conclude, there is over-policing in low socio-eco status neighborhoods.

Analysis 2: Using color code (heat map) of A&A (data_officer assignment attendance) in different neighborhoods. Set up a time slider to see how it changes over time.

In this section, we discuss the potential relations between officer attendant rate in each beat area each year and the tactical response. We assume that the attendance rate could affect the tactical response rate of the area. To observe the attendance rate each year, we created the animation of officer attendance rate each year, as figure 5 shows. In the graph, we observed that areas such as 123, 271, 101, and 105 have consistently low attendant rates. By comparing with the records in checkpoint 2, we can know that these areas also have low tactical responses and high incomes.

With the bar chart race of the TRR over different beats, we can clearly find such two leading beats, 132 and 65 (Figure 6). Associated with Plot 2 in Observable Notebook and the median income map in our Checkpoints 2, we can find that beat 65 and 132, especially beat 132 belongs to the low socio-eco status neighborhoods and it contains a high police attendance rate. The high police attendance rate is important because it implies that police officers pay attention and focus on the security in such areas. With the leading TRRs in such low socio-eco status neighborhoods like the beat 132, we can assume that there is over-policing in some low-income areas.

Conclusion

With the above four plots, we can roughly have an idea of the patterns of police misconduct. Police officers tend to give less tactical responses to high-income areas. To avoid biases, police misconduct can be viewed from both public narratives and police narratives, which means an area with both high complaint rates and high tactical response can be viewed as the police misconduct area. From our observation, police misconduct areas tend to have less income.

Checkpoint 4

Graph analytics can be very useful in analyzing relationships between different groups of people. We can create nodes based on their income, race, neighborhood, and other attributes. After building the graph, we can analyze interactions among different nodes and even graphlets.

First, we make nodes of officers and victims by their income, race, locations, and even unsupervised machine learning models to learn the cluster and see if there is a potential connection between officers and victims.

For the race, we find that there is a high volume of complaints from black people since the indegree is 67923 which is 3 times the second-highest complaints race, white, which has 20519 complaints. So, we may assume that there is an over-policing based the race bias due to the extremely large number of complaints from a specific race. However, we are not interested in the bias, this section is only used for proving our main theme, " Is there over-policing in low socio-eco status neighborhoods? "

For the location, we first plot the visualized graph of the connection of the officer and the victim by the location. Then we analyze the graph by the inDegree and the outDegree. We find that communities like Austin, West Englewood, and Loop have a high volume of complaint report to officers, and Austin, Humboldt Park, and West Garfield Park have a large amount of TRRs. From this result we can find in the high-income community, people are more likely to complain about the behavior of the police. People from low-income communities receive more "threats" of tactical response. One possible explanation is that people who live in high-income communities have time to report the misbehavior of over-policing officers. But in the low-income community, people have no power to against the over-policing. Anyway, a high amount of reports of tactical response shows that there is potential over-policing behavior in those areas. Combining with the result we find in Checkpoint 1, a community like West Garfield Park is a low-income area. Therefore, we can assume that there is over-policing in the socio-economy status community.

Moreover, the network dynamics of co-accused in each cohort can be interesting. We use triangle counts and page rank algorithms to analyze the network. The basic logic is to join the allegation table with itself on the condition of the same allegation id and unequal officer_id. Nodes can be generated with data_officer table or allegation id by counting the number of allegation id. Here we chose the data_officer table by removing Nan or 0s on allegation_count.

Recognizing the largest communities is important. So, we ranked the label propagation algorithm result by sorting descending the number of members in the community. After identifying those top big communities, we are also interested in how the community is constructed and its internal architecture. We plotted the 22809 community which consisted of over 50 nodes. It is clear to us that officers 2612, 30237, and 21028 are among those "leading" nodes with multiple indegrees and outdegrees inside the clique.

The triangle counting algorithm is to count the triangle-like relationship among 3 nodes that have connected in pairs. We want to find out those outstanding nodes in the graph which have a lot more triangle counts. In this part, we sorted all the nodes according to their triangle counts. We can see over 20 nodes appearing in over 18,000 triangle relationships, which indicates strong community leadership potential like officers 6315 and 3033.

Page rank algorithm is developed to find out important nodes inside a graph by iterations of calculations of the possibilities to get to the node by starting randomly. From the above calculations, we can identify officers with significant impact in the graph. For example, officers 32442 and 32440 are a major part of the clique and maybe the "bad apple" in the organization.

Finally, we would like to find the correlation between the number of people in each cohort and CRs or TRRs. The goal is to find whether police officers are more likely to commit misconduct when working as a group or not. From the calculations, we can get the correlation of CRs is 0.41 and the correlation of TRRs is 0.21. The correlation between group allegation and complaint reports is positive, and the group allegation is less correlated to tactical response reports. It is possible that when police officers are co-accused, they are more likely to have actual misconduct activity. It is because if they gave more tactical responses than receive complaints, or if they have an equal number of tactical responses and complaints, they would be less likely to have misconduct.

Checkpoint 5

In this analysis, we would like to further explore the relationship between over-policing and economic status. Specifically, summarization and criminal report texts can be analyzed with NLP techniques. For example, is there a topic difference between different areas? Also, Bert can be applied to generalize embeddings for the texts and calculate similarities between them.

Analysis 1: To understand the misconduct that happened in low socio-economy and high socio-economy areas, we need to understand the differences between the incidents that happened in each area. Topic modeling is a great way to understand the gist of each incident. To first get a general impression of the gist of each incident, we can use the statistical approach to calculate the frequency of each term.

The summary of the incident can be found in the data_aligation table. We need to first split the data by socio-economy status. The following SEQUEL was used to group the summary by connecting the beat area with the communities since only communities have income data.

To capture the topic, the most frequent term can provide a general idea. After the pre-process for the text, figure 7 shows the term frequency in high-income areas and the term frequency in low-income areas.

The problem with term frequency is that a single word cannot provide enough information for the context. For example, depart and subject in each plot represents the most frequent term. However, they cannot provide enough information for people to understand the context. To solve this problem, we can calculate the frequency for a sequence of words, which is the continuous words in a sentence. The high-frequency sequences are hard to view since there can be too many terms in similar frequencies, so we inverted the graph so that the shortest bar represents the highest frequency. Figure 8 is the tri-gram in low socio-economy status communities. Figure 9 is the tri-gram in low socio-economy status communities.

From analysis 1, we can see that for the high-income community, we got “involve duti”, “police review” and etc., or other normal behavior that we would expect from officers. However, in the low-income community, we find “tactic respond”, “violat rule”, and etc. The sequences in low-income areas indicate the high rate of policing in the area. However, it is hard to tell if there is any misconduct in the action.

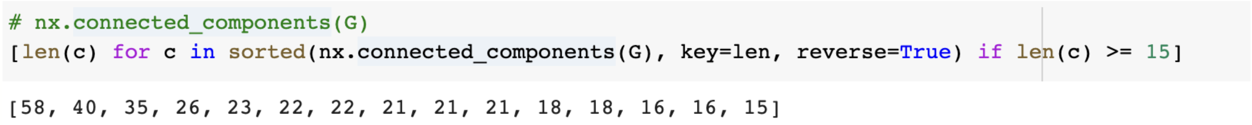
Analysis 2: For similar(in contents) complaint summary texts, are they having similar eco-socio status to each other? To understand the texts, one way is to make use of word frequency counts and topic modeling methods like we did in question1. The other feasible method is Bert. Bert is a language model which can be used to generate representational embeddings for input texts. We can further calculate the similarity between each of them. Moreover, we make use of the top 3 closest texts as edges between each other and create a graph. Finally, connected components are found to calculate the mean income for the cluster.

In the original data allegation table, there is no clue for income in the area. Fortunately, we can make use of the data_area table to get it. But, it is hard to join two tables directly. Here we make use of the polygon relationship to create a mapping between community id and beat id and thus join texts with median income.

SentenceTransformer is an implementation of the paper Sentence-Bert to generate embeddings for sentences, texts, and paragraphs. Here we made use of a lite version of the pre-trained model.

To get the embedding of the text, we just make use of one line of API from input text to 128 dimensional embedding like this. Next, to calculate the similarities between each other, we have two loops to apply cosine similarities calculations to make a 1105x1105 similarity score matrix. Then, we create a graph based on every summary text as nodes and top 2 similar nodes as edges.

Let's look at the length of outstanding components we have.



```
# nx.connected_components(G)
[len(c) for c in sorted(nx.connected_components(G), key=len, reverse=True) if len(c) >= 15]

[58, 40, 35, 26, 23, 22, 22, 21, 21, 21, 18, 18, 16, 16, 15]
```

Finally, we calculated the average income for each cluster as opposed to the mean income for the whole population shown in Figure 10, while the mean income for all texts is about 45572.

From analysis 2, we found that similar texts cluster have outstanding average income than the whole population. For example, several clusters with over 15 data points have a lower income of 30,000 while the average income is about 45,000. We can conclude that different economic areas have different complaint texts with each other, which means a different type of policing in these areas.

Analysis 3: Is there any bias in the complaint report? In other words, is the complaint report narrative different from the incident summary?

In this question, we planned to group the data into high-income and low-income areas. According to our previous finding that high-income areas have more complaints toward the police officers compared with the low-income areas, our hypothesis was that high-income areas have more bias in complaints. The bias degree can be further used to measure if the complained over-policing has really happened in the area. And this question can help to answer question 1.

To test our hypothesis, we planned to embed the text by using BERT and calculate the text-similarity by using distance matrices. Although the incident reports have some sort of biases, the texts are still profiling the incident in an objective fashion. The complaint report, on the other hand, profiled the subjective description from the complainant. Unfortunately, after querying the data, we found only 17 complaint reports. The data sample is too insufficient to be analyzed.

Conclusion

From the above analysis, we found that social-economic status is closely related to over-policing and misconduct. In different communities or beat areas, complaint rate and tactical response rate are consistently aligned with their area median income. Also, richer areas tend to file more complaints than poorer areas. Poorer areas tend to receive more tactical responses from police officers. The over-policing problem needs more evidence to be supported. We find there is more law enforcement in low-income communities, but more data and work are needed to prove there is a logical correlation.

Future Work

In the future, we need more work and data to verify our conclusion and hypothesis. For example, the high rate of complaints in high-income areas needs to be verified by detecting the potential bias in complaint summaries. Also, the legality of tactical response needs to be verified by detecting the severity of the incident or if the crime rate matches the tactical response rate.

Appendix:

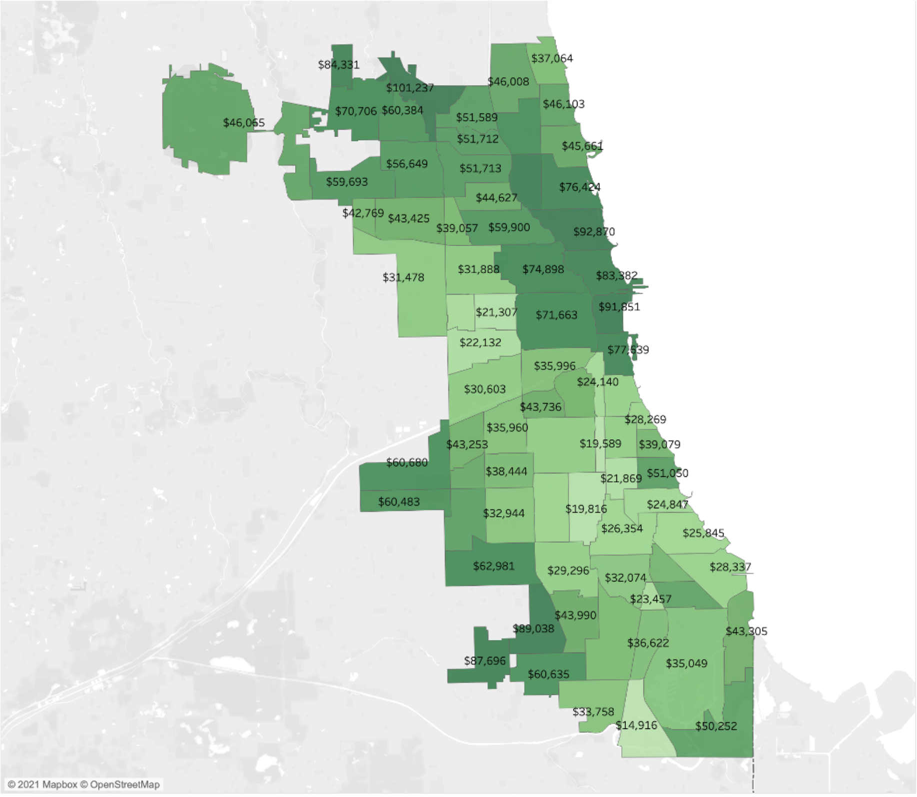


Figure 1: income map

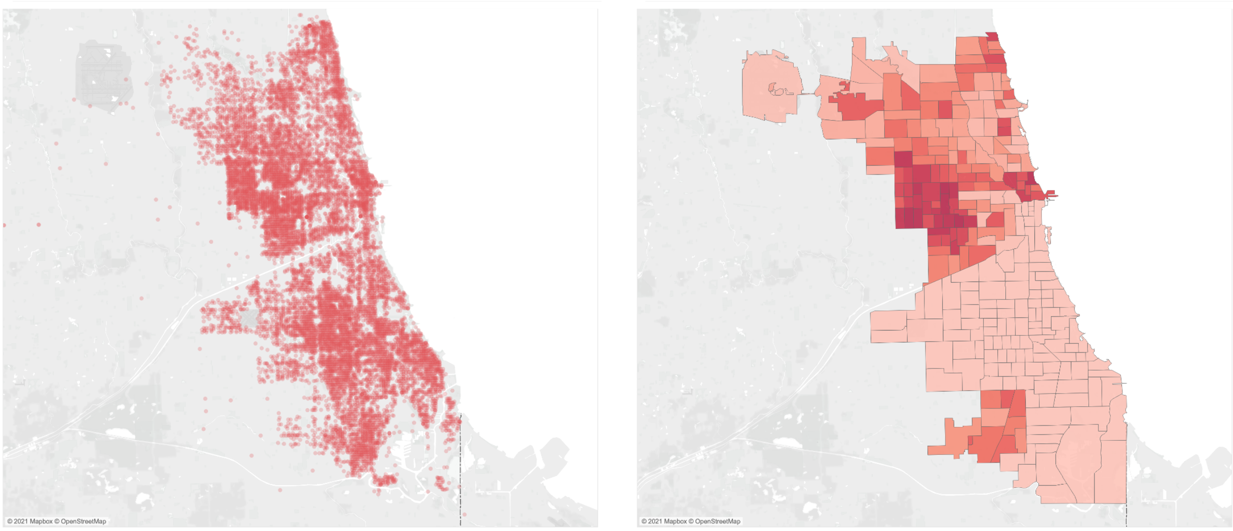


Figure 2: tactical response and complaint report map

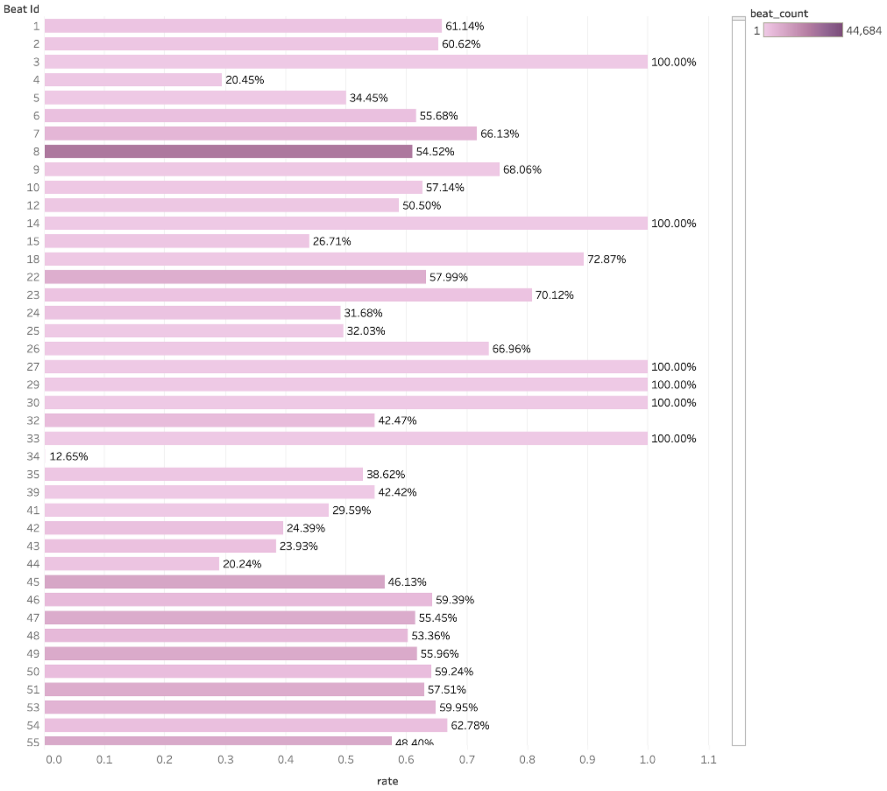


Figure 3: the percentage of attendances in different beats

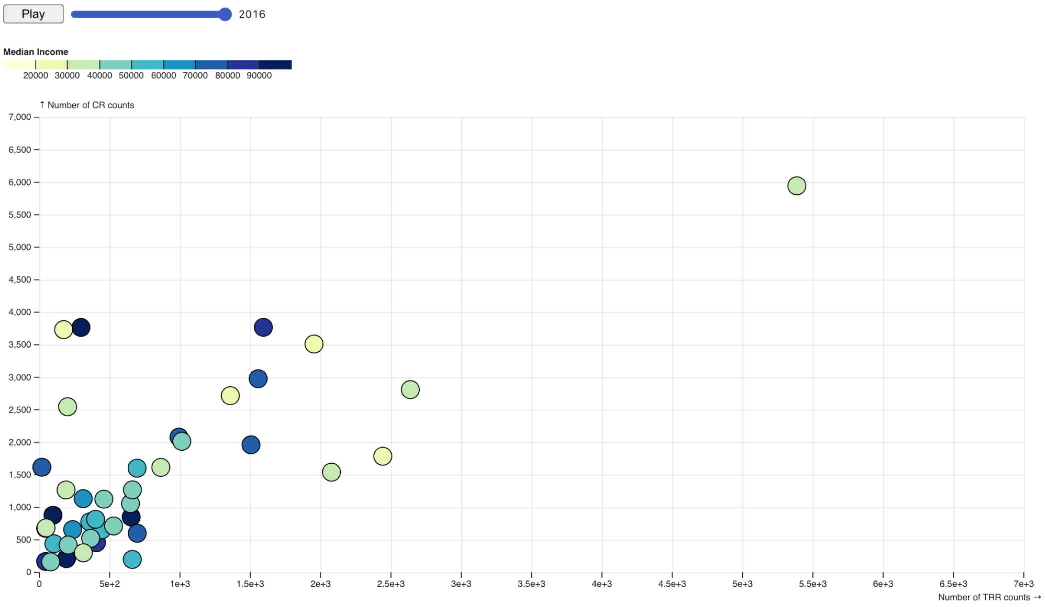


Figure 4: multiple dimensions in years, beat areas, median incomes, the number of tactical response reports, and the number of complaint reports

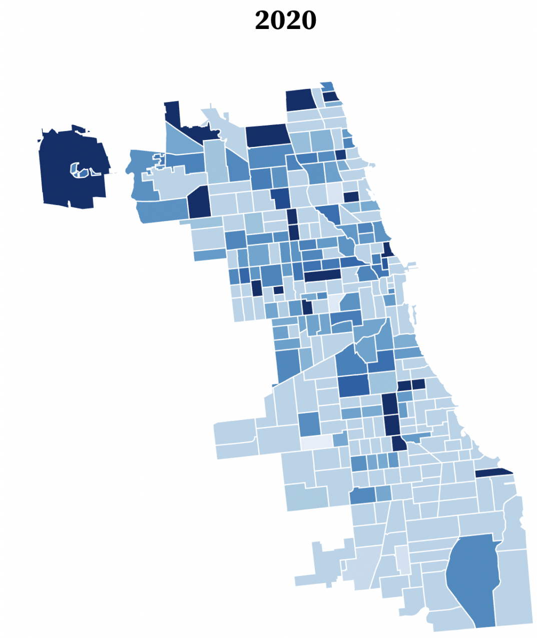
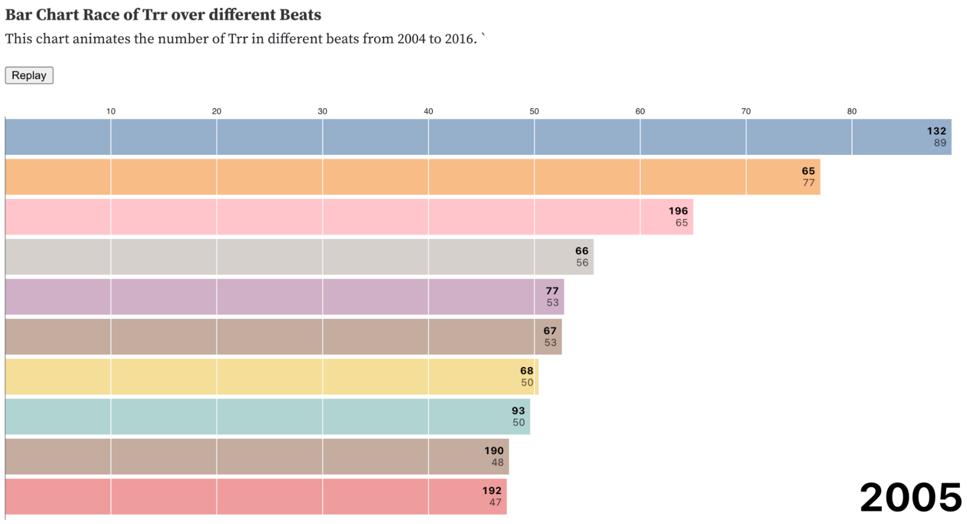


Figure 5: animation of officer attendance rate each year



Bar Chart Race of Trr over different Beats

This chart animates the number of Trr in different beats from 2004 to 2016. `

Figure 6: bar chart race of the TRR over different beats

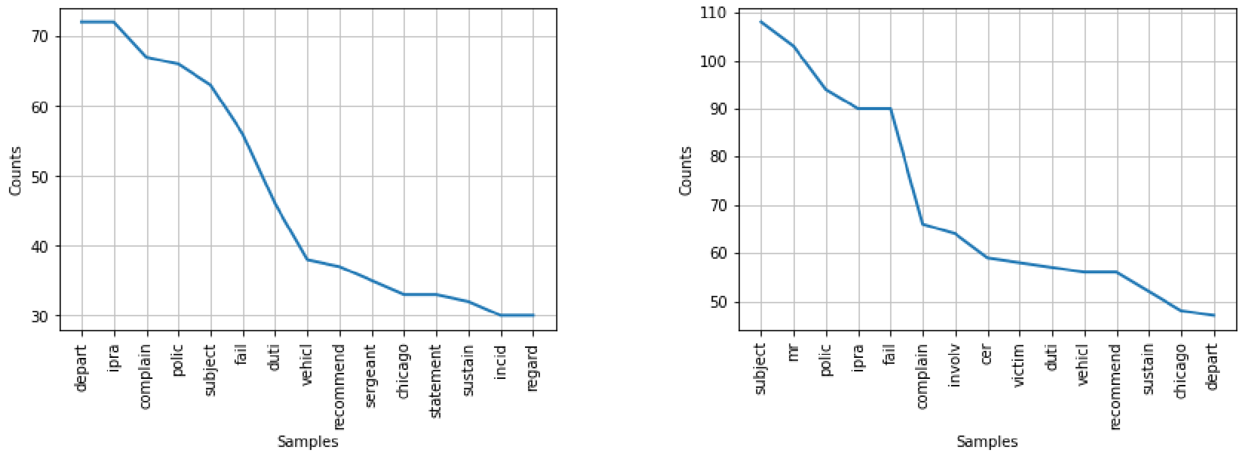


Figure 7: term frequency in high-income and low-income areas

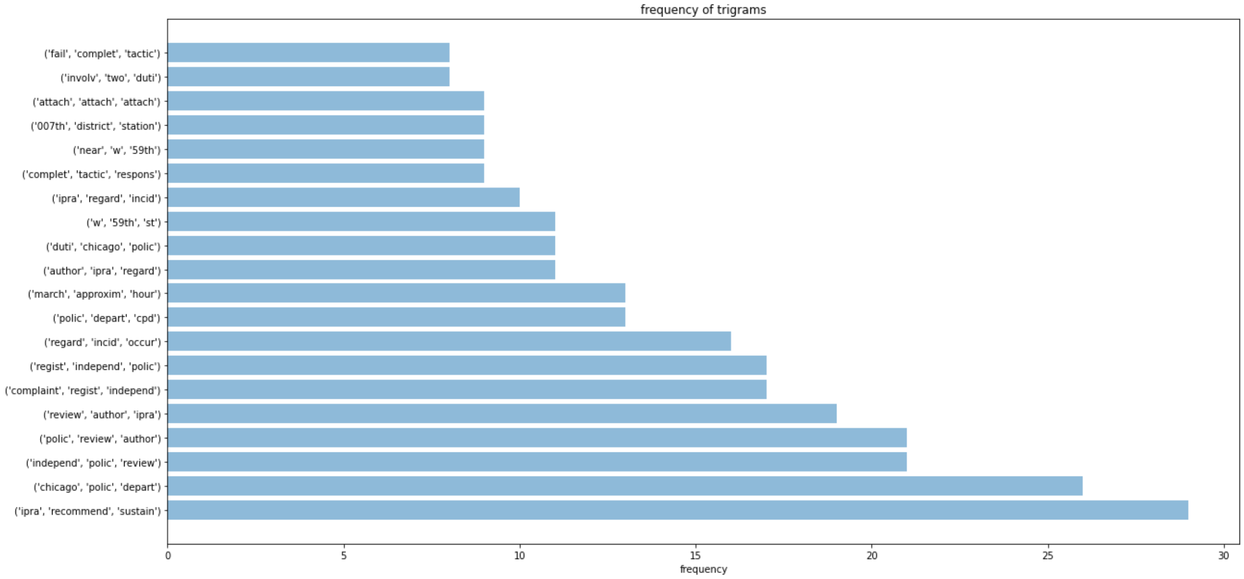


Figure 8: tri-gram in low socio-economy status communities

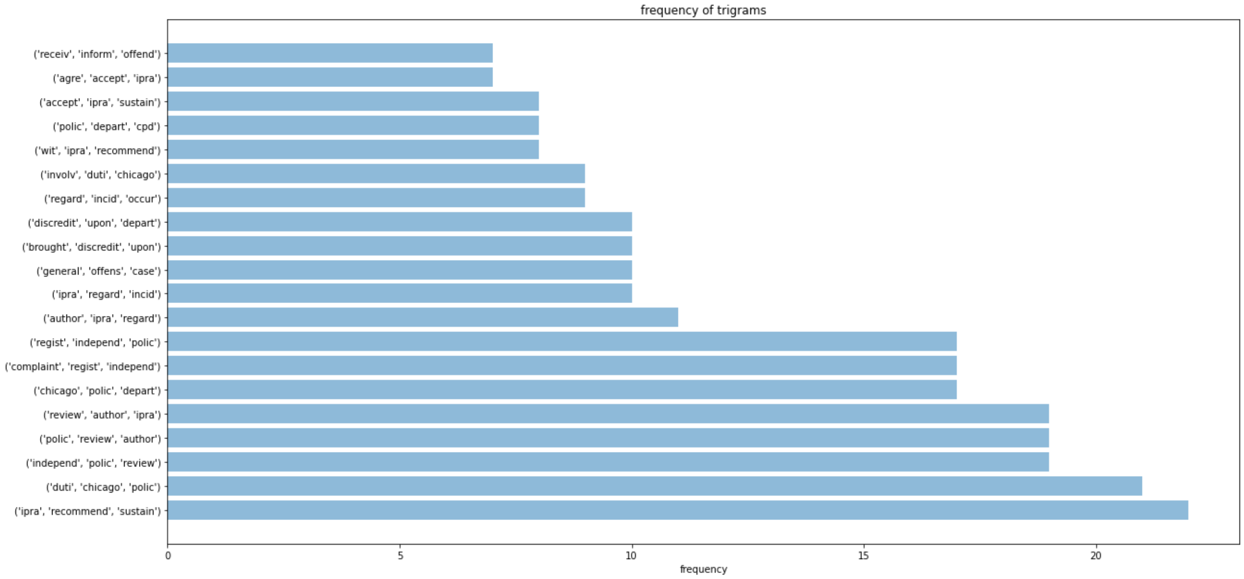
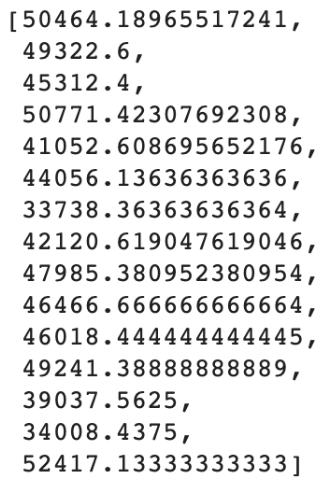


Figure 9: tri-gram in low socio-economy status communities



```
[50464.18965517241,
 49322.6,
 45312.4,
 50771.42307692308,
 41052.608695652176,
 44056.13636363636,
 33738.36363636364,
 42120.619047619046,
 47985.380952380954,
 46466.666666666664,
 46018.444444444445,
 49241.38888888889,
 39037.5625,
 34008.4375,
 52417.13333333333]
```

Figure 10: the average income for each cluster