Final Project Report

By The Brave Ducks:
Patrick Pei, Numair Ahmed, Danhua Lu

Theme

The main theme of our analysis of the Chicago Police Database is to evaluate the trend of change for police misconduct allegations, we will further differentiate the "income demographic" of a neighborhood to be designated as low-income (less than $30,000/yr), middle-income (between $30,000/yr and $75,000/yr), and high-income (greater than $75,000/yr).

With some preliminary computation on the available database, and for future reference, we identify the population sizes for different income neighborhoods as the following:

- low income neighborhood population: 401,566
- middle income neighborhood population: 1,904,676
- high income neighborhood population: 410,739

SQL Analytics

SQL analytics questions

1. What is the number of people living in low, middle, high income neighborhoods?
2. Using our definition of types of “income neighborhood", what is the total number of officer allegations for all low, middle, high income neighbors?
3. What is the rate of increase for officer allegations for low, middle, high neighborhoods between 2002-2007 and 2007-2012 timeframes, 2007-2012 and 2012-2017 timeframes?
4. What is the percentage of misconduct allegations (drug/Alcohol, illegal search, use of force, etc) out of all allegations for these low, middle, high neighborhoods?
5. Among the officer allegations with complaints filed in the low, middle, high neighborhood, what percentage of the cases are dismissed?

SQL analytics insight

First, we can see that most people are living in middle-income neighborhoods which is expected. Based on a population distribution, we can say that 14.8% of the population lives in low income neighborhoods, 70.1% of the population lives in middle income neighborhoods and 15.1% of the population lives in high income neighborhoods.

Second, we observe that the number of officer allegations in high income neighborhoods is much less than the other two income neighborhoods with regards to total counts. When we take the result from q1's sql queries, we can calculate that the office allegation per capita for low income neighborhoods is 0.16, middle income neighborhoods is 0.06 and high income neighborhoods is 0.07. This shows that there are significantly more officer allegations among low income neighborhoods on a per capita basis.

Thirdly, we can see that for all three neighborhoods, the number of officer allegations cases are rising. What surprises our group is that there is a huge jump in cases between 2002-2007 to 2007-2012 timeframes for low income neighborhoods. We suspect that this was due to the timeframe of the financial crisis and the political and income disparity that caused such big increases between 2007-2012 especially for the low income neighborhoods.

Fourthly, as we can see, the percentage is quite different for different neighborhoods. It seems that for high income neighborhoods, it is 50% less likely that there would be police misconduct allegations than for middle income neighborhoods.

Finally, as from the result, almost half of cases for complaint filed results in "No Action Taken". This is unanimous across all neighborhoods. We, as a group, are wondering which factors induce such a high rate. But this is a question that might be worth mentioning in the future direction of our research.

Data Exploration

Data exploration questions

Performing the initial step of data exploration enables us to better understand and visually identify anomalies and relationships that might otherwise go undetected. We can apply this exploration to better understand the CPDB data within our thematic context of socioeconomic status of police activity in neighborhoods of Chicago.

Now knowing the size of populations belonging to the different socioeconomic levels, we are interested in answering the following questions using packed bubbles and treemaps in the Tableau Desktop data exploration software. The questions we intend to investigate further are:

- What is the percentage of misconduct allegations (illegal search, use of force, etc) out of all allegations for these low, middle, high neighborhoods?
- Among the officer allegations in the low, middle, high neighborhood, what percentage of the cases are "No Action Taken?"

Data exploration insight

During our analysis, we observed from our visualizations that the misconduct category of "Operation/Personnel Violations" to be a very broad categorization; seemingly used as a "miscellaneous" or "catch-all" categorization when a misconduct complaint doesn't fit into any other category. Due to this, our analyses below will focus on gaining further insight into misconduct complaints categorized as Drug/Alcohol, Illegal-search, and Use-of-force.

The first question that we want to analyze:

- What is the percentage of misconduct allegations (drug/Alcohol, illegal search, use of force, etc) out of all allegations for these low, middle, high neighborhoods?

Firstly, we see that the total sum of allegation counts for this low income demographic is 2924. The allegation category counts for Drug/ Alcohol, Illegal search and Use Of Force are 139, 25 and 199 counts respectively. This means that the percentage of officer misconduct related to Drug/ Alcohol is 4.7%. The percentage of officer misconduct related to Illegal search is 0.85%. The percentage of officer misconduct related to Use Of Force is 6.8%.

Furthermore, we want to also identify the allegations counts for these misconduct allegation categories with regards to the whole low income demographic, to control for population size differences between demographics. Recall that the population for low income neighborhoods is 401,566. This means that the misconduct related to Drug/ Alcohol is 0.000346 per capita, Illegal Search misconduct complaints are 0.0000625 per capita, and Use of Force complaints are 0.000495 per capita.

Secondly, we see that the total sum of allegation counts for this middle income demographic is 3683. The allegation category counts for Drug/ Alcohol, Illegal search and Use Of Force are 66, 41 and 321 counts respectively. This means that the percentage of officer misconduct related to Drug/ Alcohol is 1.8%. The percentage of officer misconduct related to Illegal search is 1.1%. The percentage of officer misconduct related to Use Of Force is 8.7%.

Furthermore, we want to also identify the allegations counts for these misconduct allegation categories with regards to the whole middle income demographic, to control for population size differences between demographics. Recall that the population for middle income neighborhoods is 1,904,676. This means that the misconduct related to Drug/ Alcohol is 0.0000346 per capita, Illegal Search misconduct complaints are 0.0000215 per capita, and Use of Force complaints are 0.000168 per capita.

Thirdly, we see that the total sum of allegation counts for this high income demographic is 1429. The allegation category counts for Drug/ Alcohol, Illegal search and Use Of Force are 58, 10 and 72 counts respectively. This means that the percentage of officer misconduct related to Drug/ Alcohol is 4.1%. The percentage of officer misconduct related to Illegal search is 0.7%. The percentage of officer misconduct related to Use Of Force is 5.0%.

Furthermore, we want to also identify the allegations counts for these misconduct allegation categories with regards to the whole high income demographic, to control for population size differences between demographics. Recall that the population for high income neighborhoods is 410,739. This means that the misconduct related to Drug/ Alcohol is 0.000141 per capita, Illegal Search misconduct complaints are 0.0000243 per capita, and Use of Force complaints are 0.000175 per capita.

The first observation we have is that for the Drug/ Alcohol category, the low income neighborhood has the highest percentage out of all allegation counts with 4.7%. Meanwhile, for both Use Of Force and Illegal Search categories, the middle income has the highest percentage out of all allegation counts with 1.1% and 8.7%. Meanwhile, to apply these categories with regards to the whole demographic, for the Drug/ Alcohol category, the low income neighborhood has the highest per capita case, 10x more than the middle income neighborhoods, 2x more than the high income neighborhoods. For Use Of Force, the low income neighborhood has the highest per capita case again, nearly 3x more than the middle income neighborhood and high income neighborhood. Finally for Illegal search, the low income neighborhood has the highest per capita case yet again, nearly 3x more than the middle income neighborhood and high income neighborhood. This really shows that the low income neighborhood has been disproportionately affected by police misconduct across all three categories.

The second question that we want to analyze:

- Among the officer allegations with complaints filed in the low, middle, high neighborhood, what are the final outcomes?

Firstly, we see that the total sum of allegation outcome counts for this low income demographic is 4580. The count for No Action Taken is 2033, which is 44.4% of out all allegation outcomes. Recall that the population for low income neighborhoods is 401,566. To apply this number with regards to the whole demographic, No Action Taken Outcome is 0.00506 per capita.

Secondly, we see that the total sum of allegation outcome counts for this middle income demographic is 5766. The count for No Action Taken is 2609, which is 45.2% of out all allegation outcomes. Recall that the population for middle income neighborhoods is 410,739. To apply this number with regards to the whole demographic, this is 0.00228 per capita.

Thirdly, we see that the total sum of allegation outcome counts for this high income demographic is 2107. The count for No Action Taken is 938 which is 44.5% of out all allegation outcomes. Recall that the population for high income neighborhoods is 1,904,676. To apply this number with regards to the whole demographic, this is 0.00136.

An astonishing observation we can make from our analysis, is that when corrected for population size, if a low-income person files an officer misconduct complaint, it is 3x as likely to go uncorrected than if a middle-income person filed a complaint and nearly 4x as likely to go uncorrected than if a high-income person filed a complaint. This insight draws light to an imbalance of fairness when treating misconduct complaints against officers in charge of their respective neighborhoods.

Interactive Visualization

Interactive visualization questions

In this checkpoint, we want to investigate police misconduct allegation categories that include illegal search and use of force. We will implement the D3.js to generate an interactive visualization focusing on misconducts categorized as Illegal-search, and Use-of-force in different demographic areas such as racial and socio-economic status. We want to explore further how socioeconomic statuses coupled with racial demographic plays a role in misconduct allegations. Lastly, we want to gain insight into proportions of cases dismissed in relation to different income demographics of the neighborhood. To be noted, we have splitted middle income neighborhoods into lower middle income neighborhoods ($30,000 <-> $50,000) and upper middle income neighborhoods ($50,000 <-> $75,000).

In this checkpoint, our main intention is to:

1. Use Interactive Packing to group neighborhoods in different socioeconomic status (i.e. high-income, middle-income and low-income neighborhoods) and use color encodings to show the racial composition of police who have committed Use Of Force, and Illegal Search **misconduct** allegations.
2. For our last question regarding the percentage of the **dismissed** cases for all income neighborhoods, we can use an interactive horizontal chart to represent these cases with respect to the income demographic levels. It will give us a better understanding of how many cases are dismissed. Users can use the dropdown menu to sort it in ascending or descending orders.

Interactive visualization insight

Question 1 insight
1. Use of force
We use circle parking to group the category of use of force in each income neighbor (high, middle, and low) and different races (balck, white, hispanic, asian, and other) to show the in-depth portion in Use Of Force category.

2. Illegal search
We use circle parking to group the category of use of force in each income neighbor (high, middle, and low) and different races (balck, white, hispanic, asian, and other) to show the in-depth portion in the Illegal Search category.

From the circle packings we have the following key findings:

- In high income neighborhoods, we see fewer clusters of misconduct counts whereas we see more clusters in low income neighborhoods, followed by middle income neighborhoods. In addition, we see that the police have committed more misconducts in middle and low income Black neighborhoods than high income neighborhoods [1].
- For all misconduct categories, we see that the majority of police committing such misconduct is more common among Asian, Hispanic and White populations in high income neighborhoods. In middle income neighborhoods and low income neighborhoods, we see that the racial demographic is more majorly Hispanic, White and Black.
- The difference in all three listed categories between minority groups and the others has an increasing trend as the neighborhood goes from low income to high income.

In checkpoint 1, we have only identified the percentage of misconduct cases. In this checkpoint, we delve more into racial demographic of the misconduct cases being committed against the communities. This fits with our theme to see the effect of police misconduct allegations in different neighborhoods with different socioeconomic status and how police misconducts affect the neighborhoods both social-economically and racially. This finding may not lead to the direct conclusion that the certain race group is treated less as a possibility of facing police misconduct, we also should consider the actual racial demographic distribution in different neighborhoods as a factor [1]. Regarding the change in demographic for high income neighborhoods, we hypothesize the systemic racial demographic plays an important role in shaping our graph. In the future, we could adjust our graph as a per capita graph by normalizing the counts with counts per population.

Question 2 insight

To answer this question, we use an interactive horizontal bar to see the number of dismissed cases for all income neighborhoods with respect to the lower income, lower middle income, upper middle income and upper income neighborhoods.

From the interactive horizontal bar graphs we have the following key findings:

- Most of the dismissed allegation cases reside in lower-middle income areas.
- There are more dismissed allegation cases in lower income and lower-middle income neighborhoods than upper-middle income and upper income neighborhoods.
- The top two most dismissed cases are in lower-middle income, and lower income neighborhoods respectively. This might be the case due to the fact that in these neighborhoods, people don't have access to legal resources, such as lawyers and advisors.
- In the ultra wealthy neighborhoods which have household income (> $100,000), it has the lowest dismissed allegation cases out of all neighborhoods. This might be the case because in these neighborhoods, there is less crime and less corruption happening compared to lower income and lower-middle income neighborhoods. Therefore, police misconducts, such as illegal search and use of force have less applications.

In checkpoint 1, we only identified the number of allegation cases dismissed for lower income, middle income and upper income neighborhoods. In this checkpoint, we expand the middle income to lower-middle income and upper-middle income to investigate further on our claim. In addition, we also use interactive horizontal graphs to allow the users to evaluate the trend of dismissed cases per income demographic in 4 different income neighborhoods either ascending or descending. We have also identified the highest dismissed cases are committed in lower and lower-middle income neighborhoods. It might be due to the fact that people don't have access to legal resources in these neighborhoods. Finally, we also notice that in ultra wealthy neighborhoods, it has the lowest dismissed allegation cases. It might be due to the fact that there is less crime and corruption; therefore illegal search and use of force might not happen as often in these neighborhoods.

Graph Analytics

Since our main focus is on police misconduct, in this section, we want to gain a deeper understanding of misconduct in the context of co-offending officer relationships and will build a network graph to do so. We first frame the co-offending relationship analysis within the 'Illegal search" and "Use of Force" classification of misconduct and build our graph for this specifically. We then inspect co-offending officer relationships via tools from the graph analysis Python library Graphframes which is built on the GraphX technology.

The exploration of the following questions is done via graph analytics

1. Are there occurrences of co-offending officers on the same misconduct report and how often do the same co-offending officers repeat?

2. What is the salary, rank, and race relationship between the topmost pair of co-offending officers? This analysis can then be repeated for each pair, as needed. For this report, we will provide discussion for the topmost offending pair.
    a. Do the officers have comparable salaries within 10% of each other?
    b. Are the officers of similar ranking or is one a higher ranking officer, possibly socially pressuring a lower ranking officer into committing misconduct?
    c. Are the officers the same race?
3. How many unique co-offending relationships does each officer have? It is straightforward to compute total misconduct of an individual officer but will be more meaningful to understanding how many unique relationships an officer has in which he/she allegedly commits a misconduct.
4. Who is the ring-leader (most important) co-offending officer?

Question 1 insight

The table below represents a graph with source node being officer_id1 and destination node being officer_id2, while their relationship is the total count of co-offending misconduct.

```
+-----+-----+------------+
|  src|  dst|relationship|
+-----+-----+------------+
|12478|32166|          36|
| 8562|27778|          34|
| 2725|21703|          29|
| 1553|10724|          28|
| 3605|14442|          28|
| 8562|18206|          28|
|12074|12825|          28|
|32265|32347|          27|
| 8562|23841|          26|
|31882|32401|          25|
|13361|20150|          25|
| 1553|16699|          24|
|23841|27778|          24|
|32016|32213|          24|
|14731|27602|          23|
|14045|15502|          23|
|12479|20713|          22|
|17285|17397|          21|
|18206|27778|          21|
| 8658|13788|          21|
+-----+-----+------------+
only showing top 20 rows
```

Question 2 insight

- Do the officers have comparable salaries within 10% of each other?
- are the officers of similar ranking or is one a higher ranking officer, possibly socially pressuring a lower ranking officer into committing misconduct?

- Are the officers the same race?

To gain more insight into the top most co-offending pair of officers, we inspect the graph vertices to see the exact names, salary, rank, and race of the officers with the most total misconduct counts. As we can see below, the officers Ronald Holt and Emmet Mc Clendon have the most co-offending misconduct complaints. There is a more than 10% discrepancy between their salaries and their officer rankings are significantly different, implying some hierarchical relationship between officer Holt and officer McClendon. Further, we note that both officers in this pair are of race black.

Question 3 insight

Background on the Triangle Count algorithm. The Triangle Count algorithm counts the number of triangles for each node in the graph. A triangle is a set of three nodes where each node has a relationship to the other two. In graph theory terminology, this is sometimes referred to as a 3-clique. The Triangle Count algorithm in the GDS library only finds triangles in undirected graphs.

[Edit]: We previously noted that officer_id 32356, Officer Vincent Stinar, has the greatest number of co-offending relationships as analyzed by the triangle count algorithm. It was also noted that these are unique co-offending triangular relationships and that it did not make intuitive sense to have such a large triangle count result, 1514, while having only 12 total counts of misconduct allegations. After further review of the Triangle Count algorithm, we find that the analysis is not of *unique* co-offending relationships, but repeated count of relationships. This would then mean it is reasonable to see a larger triangle count result even if the total misconduct count is much lower.

Question 4 insight

We need to recast the data query to fit PageRank's bi-directional nature. To do this, we recast the graph edges. Note the query line "AND da1.officer_id < da2.officer_id" is modified to "AND da1.officer_id <> da2.ifficer_id".

We use the PageRank algorithm on our graph to highlight which officers perform as ring leaders in the perspective of the data.

As visualized above, the pagerank algorithm shows that officer Glenn Evans has the most influence on other officers in co-offending misconduct cases with a PageRank of 9.22 and total misconduct count of 73. Although at first glance the total misconduct count may seem less than other entries, we must note that PageRank will highlight which officer has the most commonality with other officers in committing misconduct offenses; thereby being "ring leaders". Other officers may commit additional misconduct offenses separately as individuals and thereby increase their overall misconduct count.

To further highlight this point, note that Officer Jerome Finnigan, officer_id 8562, has the highest total of misconduct allegations yet is only the 4th highest ranking "ring-leader" as analyzed by PageRank.

Natural Language Processing

We inspect further into the overall sentiment with regards to police misconduct in low, middle and high income neighborhoods. The exploration of the following questions is done via NLP, conducted using Tokenization, Transformer and Sentiment Analysis tools, processed in the attached Google Colab notebook.

Questions we seek to answer with NLP:

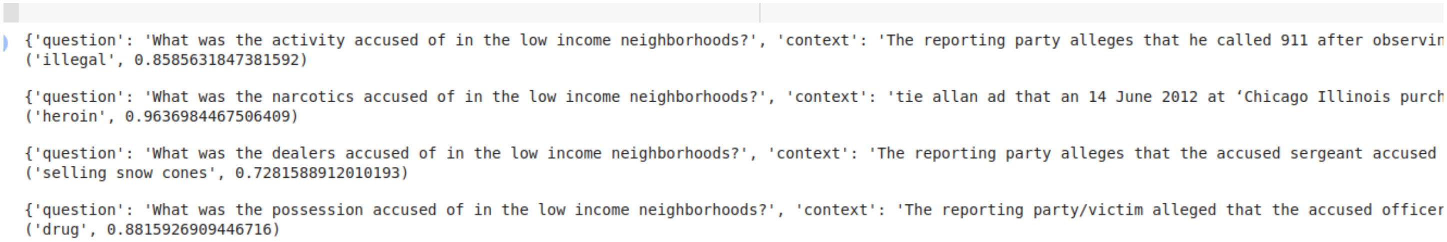
1. What are the top 15 most frequent words in the narrative context in the low, middle, high income neighborhoods?

2. What are the most frequently used words by an accuser? What are the most frequently used words by an accused officer? We will be answering these questions with regards to socioeconomic status and police misconduct (i.e. illegal search, use of force).
3. What misconduct was alleged by the accuser? We want to understand what the officer is being accused of and we want to figure out what the socioeconomic status is of the accuser. This will give us insight into if the accuser behaves differently depending on their socioeconomic status.
4. What are the top 5 negative complaint narratives for each income neighborhood?

Question 2 insight

Using spacy tokenization, we are able to see that in low income neighborhoods, the accused police misconduct is linked to primarily illegal activity, followed by drug, and illegal search. In middle income neighborhoods, the accused police misconduct is primarily illegal search, followed by drugs. Finally, in the high income neighborhoods, the accused police misconduct is primarily profanities. We see that in high income neighborhoods, the narrative about misconduct is very different than low and middle income neighborhoods. To find out exactly what was alleged and what the person is alleging in different income neighborhoods. We want to build an NLP transformer to answer this question for us.

Question 3 insight



```
{'question': 'What was the activity accused of in the low income neighborhoods?', 'context': 'The reporting party alleges that he called 911 after observin
('illegal', 0.8585631847381592)

{'question': 'What was the narcotics accused of in the low income neighborhoods?', 'context': 'tie allan ad that an 14 June 2012 at ‘Chicago Illinois purch
('heroin', 0.9636984467506409)

{'question': 'What was the dealers accused of in the low income neighborhoods?', 'context': 'The reporting party alleges that the accused sergeant accused
('selling snow cones', 0.7281588912010193)

{'question': 'What was the possession accused of in the low income neighborhoods?', 'context': 'The reporting party/victim alleged that the accused officer
('drug', 0.8815926909446716)
```

We build this model to help us understand the narratives about police misconduct in different income neighborhoods. We select the question and answer pair that have the confidence > 0.7 to be more accurate. In addition, we see a transition from low to middle to high income neighborhoods in the context of types of misconduct. In low income neighborhoods, drugs are the primary reason; in middle income neighborhoods, we see more illegal searches; in high income neighborhoods, we see more profanity. The level of severity decreases as you enter different income neighborhoods.

Overall Insights and Future Work

Overall insights

From our analysis, it is astonishing to observe that when normalized for population size, if a low-income person files an officer misconduct complaint, it is 3x as likely to go as "No Action Taken" than if a middle-income person filed a complaint and nearly 4x as likely to go as "No Action Taken" than if a high-income person filed a complaint. This insight may highlight an imbalance of fairness when treating misconduct complaints against officers with respect to particular neighborhoods.

Furthermore, in high income neighborhoods, we see fewer clusters of misconduct counts whereas we see more clusters in low income neighborhoods, followed by middle income neighborhoods. In addition, we see that the police have committed more misconducts in middle and low income Black neighborhoods than high income neighborhoods.

As somewhat expected, in the ultra wealthy neighborhoods which have household income greater than $100,000 per year, it has the lowest dismissed allegation cases out of all neighborhoods. This might be the case because in these neighborhoods, there is less crime and less corruption happening compared to lower income and lower-middle income neighborhoods. Therefore, police misconducts, such as illegal search and use of force have less applications.

Finally, the natural language processing libraries show that in low income neighborhoods, drugs are the primary reason as filed in the misconduct reports; in middle income neighborhoods, we see more illegal searches; in high income neighborhoods, we see more use of profanity by the citizen and the officer. The level of severity decreases as you enter different income neighborhoods.

Challenges

As we progressed through our exploration and analysis of the CPDB, we found from the beginning that many questions cannot be answered due to inadequate data. For example, police departments in recent years have begun to implement the use of body cameras and we naturally thought it would be meaningful to analyze any video data or associated metadata of body camera footage, if available. Although a question may intuitively seem interesting to explore, we have understood it is not always captured in the data and so cannot be objectively analyzed, as is the case with the question on body camera footage.

In situations where the database does suggest to support a formulated question, we find that the proper *framing* of a question is of utmost importance. Sometimes, we can believe we have made a conclusion based on data but it turns out the conclusion is incomplete or outright wrong. And so, it is important to frame the question with tight definitions and to be very specific. For example, in our question "percentage of cases **dismissed**" we try to make a seemingly straightforward conclusion on whether a demographic of victims correlates to an outsized proportion of **dismissed** misconduct cases. To laymen, the question seems straightforward but the framing of the question focusing on *dismissed* is misdirected since the true outcome of a misconduct case is categorized into five actions, of which 3 can loosely be described as being "No Action Taken". The insight we glean from this is to be very particular in framing questions for which we explore the data and make conclusions. Analyzing this sort of data to make conclusions is understandably a sensitive topic as it has the potential to impact real lives of officers and ordinary citizens; it is highly important to have robust questions and high quality data.

Future Work

Throughout this work we have observed a general trend in the correlation of low income and black neighborhood demographics seeing more cases of misconduct allegation reports. The intent of the Invisible Institute and the Chicago Police Department in working together toward the CPDB effort is to curb cases of police misconduct allegations. With this in mind, and the capabilities of technologies such as natural language processing and predictive algorithms from machine learning, we are curious to ask the following questions:

- Is there a way to ***predict*** misconduct based on neighborhood median income or any other socioeconomics factor?
- If successful in ML implementation, is there a way to apply these learnings and predictions to other police departments in the US?