The Brave Ducks
Patrick Pei, Numair Ahmed, Danhua Lu
10/01/2021

Project Proposal

Background and Theme

The police violence has become a very heated topic. It is speculated that in different demographics, lower-income neighborhoods have been unproportionately affected by policy violence. The number of officer allegations per capita is a way to comparatively measure the police allegations per capita in a low, middle, and high-income neighborhood. Our definitions for different income neighbors are following:

- a **low-income neighborhood** is a community that has less than $30,000 of median income,
- a **middle-income neighborhood** is a community that has a median income between $30,000 - $75,000,
- a **high-income neighborhood** is a community that has a median income of over $75,000.

We want to use data to observe any correlation between the number of officer allegations filed to the income demographic of the neighborhood in which the misconduct occurred from the beginning to the most recent allegations (with the most recent date as of this writing in 2021). We want to see if there are any trends that are worth noting.

Research Questions

To evaluate the trend of change for police 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 neighborhoods population: 401,566
- middle income neighborhoods population: 1,904,676
- high income neighborhoods population: 410,739

Relational Analytics (Checkpoint 1)

- What is the number of people living in **low, middle, high** income neighborhoods?

- Using our definition of types of “income neighborhood", what is the total number of officer allegations for all **low, middle, high income** neighbors?
- 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?
- 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?
- Among the officer allegations with complaints filed in the **low, middle, high** neighborhood, what percentage of the cases are dismissed?

Data Exploration (Checkpoint 2)
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?"

Interactive Visualization (Checkpoint 3)
In this checkpoint, we want to investigate police misconduct allegation categories that include alleged occurrences of illegal search and use of force by the police. 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 may play a role in the alleged misconduct. 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 split middle income neighborhoods into lower middle income neighborhoods ($30,000 <-> $50,000) and upper middle income neighborhoods ($50,000 <-> $75,000).

In this checkpoint, we will explore the following:

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 recorded Use Of Force, and Illegal Search misconduct allegations.
2. For our last question, we want to quantify how many cases of officer allegations ended up being categorized as "No Action Taken" 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 officer allegation cases ended up having no action taken.

Graph Analytics (Checkpoint 4)
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?

The graph is built with the following configuration:

- nodes: id, officer name and misconduct count
- edges: src(officer1 id), dist(officer2 id) and relationship(misconduct count)

Natural Language Processing (Checkpoint 5)
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?

Future work

Using these conclusions from our questions and answers, is there a way to predict misconduct based on neighborhood median income or any other socioeconomics factor? And is there a way to apply these learnings and predictions to other police departments in the US using transfer learning tools?