Chicago Police Complaints Analysis

Data

For this project, I use the data on Chicago Police Department complaints provided by the Invisible Institute (https://invisible.institute/police-data (https://invisible.institute/police-data)).

The data were made public in response to the widely publicized reports of police violence of the past few decades, a number of which occurred in Chicago, Illinois. In 2007, a writer named Jamie Kalven and a professor of law at the University of Chicago named Craig Futterman filed a lawsuit against the city of Chicago in 2007 to force the city to release complaint records files against the police department.

After seven years of litigation, Kalven received a ruling that gave public access to nearly all available records stored by the Chicago Police Department (CPD), including officer rosters, complaints filed against officers, and rulings made by the Police Department on these complaints. Numerous articles were published about the striking patterns that the data revealed:

- Police Abuse Complaints By Black Chicagoans Dismissed Nearly 99 Percent Of The Time (http://data.huffingtonpost.com/2015/12/chicago-officer-misconduct-allegations)
- Chicago Pays Millions but Punishes Few in Killings by Police (https://www.nytimes.com/2015/12/18/us/chicago-pays-millions-but-punishes-few-in-police-killings.html)
- 130 Chicago Officers Account For 29 Percent Of Police Shootings (https://theintercept.com/2018/08/16/chicago-police-department-officer-involved-shooting/)
- Chicago Police Are 14 Times More Likely To Use Force Against Young Black Men Than Against Whites (https://theintercept.com/2018/08/16/chicago-police-misconduct-racial-disparity/)
- Bad Chicago Cops Spread Their Misconduct Like A Disease (https://theintercept.com/2018/08/16/chicago-police-misconduct-social-network/)

The data are publicly available as CSV and zipfiles on GitHub (https://github.com/invinst/chicago-police-data (https://github.com/invinst/chicago-police-data)). They include more than 240,000 allegations of misconduct over the past fifty years. In particular, the data form a census of complaints filed against the CPD from 2000 to 2018.

The data record complaints, rosters of police officers, salaries of officers, and CPD awards given to officers. The entire dataset is 85 MiB unzipped. In this report, I examine the data on complaints and police officers. Much of this analysis is based on a 538 article called How To Predict Bad Cops In Chicago (https://fivethirtyeight.com/features/how-to-predict-which-chicago-cops-will-commit-misconduct/).

Problem

The "few bad apples" hypothesis claims that police misconduct is primarily caused by a small number of highly frequent violators. This hypothesis is important because it informs policy

decisions — if true, the CPD will more effectively improve overall police behavior by identifying and removing "bad apples" instead of retraining their entire taskforce.

If the "few bad apples" hypothesis is true, we expect to find that:

1. Police with at least one complaint are likely to have multiple complaints.
2. An officer's number of past complaints predicts their likelihood of receiving future complaints.

I test the two hypotheses above by examining historical patterns of complaints.

Preprocessing

To start working with the data, I download the `unified_data.zip` (https://github.com/invinst/chicago-police-data/blob/master/data/unified_data.zip) file from the data repository, unzipping the files into my local `data/` directory. The `data/` folder contained multiple subfolders which I restrict to only the files used in my analysis:

- `complaints-complaints_2000-2018_2018-03.csv.gz` : Contains individual complaint records for Jan 2000 through Mar 2018. Each complaint is filed against one or more officers.
- `complaints-accused_2000-2018_2018-03.csv.gz` : Contains accusation records for Jan 2000 through Mar 2018, which link a complaint with an officer that the complaint is filed against.
- `roster__2018-03.csv.gz` : Contains the roster of police officers who served at some point in Jan 2000 through Mar 2018.

In [3]:
```
ls -R data
```

```
complaints/      data-dictionary/ roster/

data/complaints:
complaints-accused_2000-2018_2018-03.csv.gz
complaints-complaints_2000-2018_2018-03.csv.gz

data/data-dictionary:
data-dictionary.yaml  unit_reference.csv

data/roster:
roster__2018-03.csv.gz
```

To prepare the data for analysis, I read in all three of the above CSV files in `pandas` DataFrames. I perform EDA and data cleaning — for example, the police ID column in the accusations table were read as floats which I cast to strings.

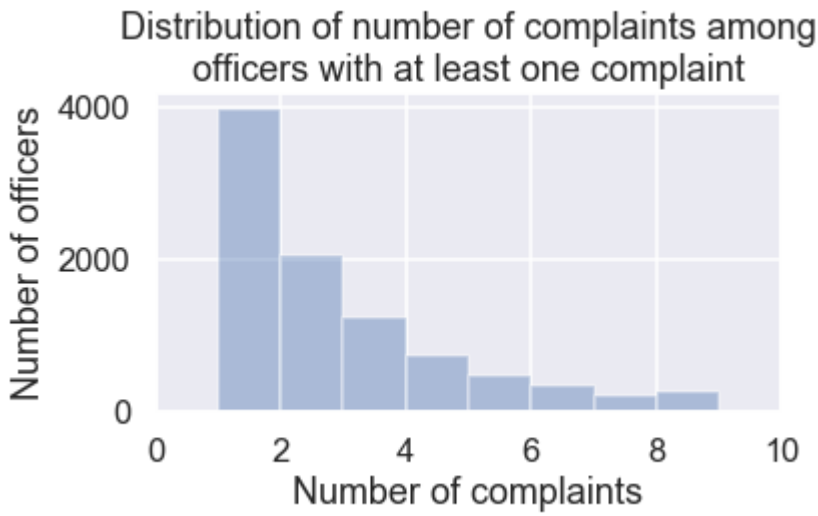
In the complaints and accusation tables, there were a number of duplicates IDs even though the data dictionary claims that the IDs are unique within each table. After examining these complaints, I determined that the duplicate complaint IDs represent different allegations within the same complaint (e.g. `USE OF PROFANITY` and `HANDCUFF TOO TIGHT` for the same officer). In order to avoid double-counting these complaints, I drop duplicate complaints. I also drop duplicate accusations to avoid double-counting complaint-police pairs.

I then join all three tables together to compute a mapping between complaints and officers. There is a many-to-many relationship between complaints and officers; each complaint can be filed against multiple officers and each officer can have multiple complaints filed against them.

The joined table has 77942 rows and 38 columns. The most relevant columns contain the complaint date, reason for complaint, and personal information on the officers accused.

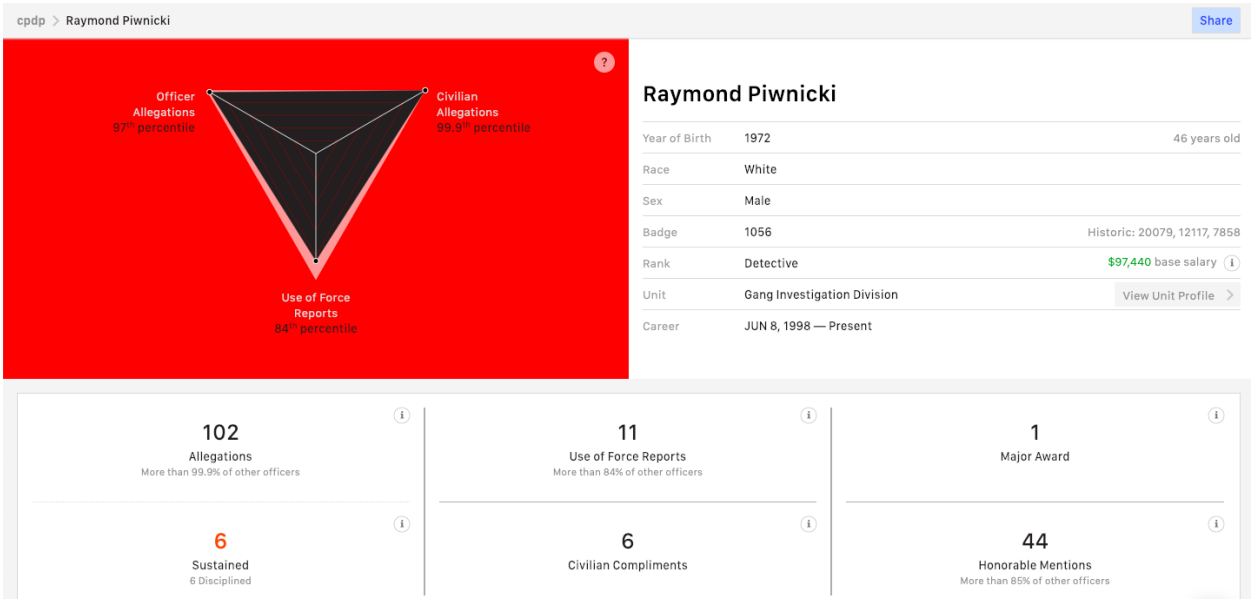
Visualization

To understand the data, I first make a plot of the distribution of the number of complaints filed against officers.



The distribution of complaints has a long right tail, suggesting that most officers have few complaints. This plot, however, hides the fact that the tail extends quite far; twenty-two officers had at least 20 complaints filed against them. The worst offender, Raymond Piwnicki, had 42 complaints!

The Citizens Police Data Project (https://cpdp.co/) provides a web interface to the same dataset. It confirms that the officers with many complaints I found were reasonable. For example, Raymond Piwnicki has the following profile:



To better understand the distribution of complaints, I create a point plot where officers who had four or more complaints were lumped in a single category. The plot shows the proportion of officers with a certain number of complaints for all officers with at least one complaint.

This plot shows that a relatively high number of officers receive multiple complaints. Note that the number of complaints in the sample likely underestimates the true number of police violations — only the violations reported to the CPD make it into the sample, so there are likely more violations that do not appear in our sample.

Still, we might make the assumption that the complaints in the dataset are a random sample of all CPD police complaints, reported or not. If this assumption holds, the proportion of officers in our sample that received a certain number of complaints is an unbiased estimator of the true proportion of officers in the CPD with that number of violations.

Model and Methodology

Let $\hat{P}(k \text{ complaints} | \geq 1 \text{ complaint})$ represent the empirical probability that an officer with k complaints is picked at random from our sample (our sample only contains officers with at least one complaint). Let $P(k \text{ complaints} | \geq 1 \text{ complaint})$ represent the probability that a officer has k complaints if all possible complaints were reported, given that the officer has at least one complaint. Then, if our assumption above holds, we have:

$$\hat{P}(k \text{ complaints} | \geq 1 \text{ complaint}) \approx P(k \text{ complaints} | \geq 1 \text{ complaint})$$

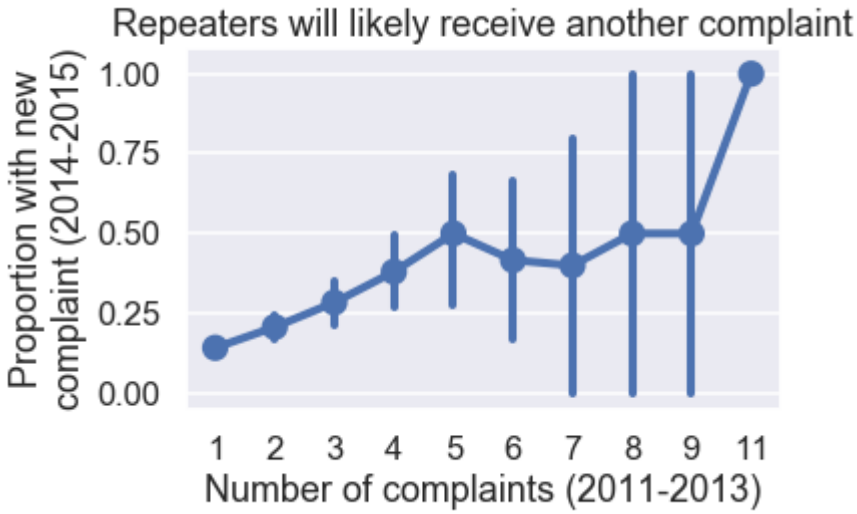
For all k in our data.

In particular, the following table shows the empirical proportions of complaints within the sample:

k	$\hat{P}$
1	0.41
2	0.21
3	0.13
4+	0.25

Since $\hat{P}(1 \text{ complaint} \mid \geq 1 \text{ complaint}) < 0.5$, it is more likely than not that an officer chosen randomly from the sample has had multiple complaints. Since our sample's complaint numbers likely underestimate the true number of complaints, the data support the claim that officers in the CPD are more likely to be repeat offenders than single-time offenders. This supports hypothesis (1).

Hypothesis (2) claims that a greater number of complaints is correlated with an increased likelihood of receiving future complaints. To investigate this hypothesis, I compute the number of complaints filed against each officer from 2011-2013 and from 2014-2015. I display a point plot below showing the proportion of officers who had a complaint in 2014-2015 for each number of complaints in 2011-2013. The plot also shows bootstrapped 95% confidence intervals for each number of complaints from 2011-2013.



This plot suggests that there is a non-zero correlation between number of complaints in the past and probability of having another complaint in the future, although there is a lack of data at higher

complaint numbers. To better evaluate this correlation, I fit a logistic regression model to predict the probability of complaint in 2014-2015 based on number of complaints in 2011-2013. This model is described as:

$$\hat{y} = \sigma(rx) + c$$

Where $\hat{y}$ is the predicted probability, σ is the sigmoid function, r is the logistic correlation, x is the number of complaints, and c is the intercept.

Using this model, I construct a hypothesis test:

$$H_0 : r = 0$$
$$H_a : r \neq 0$$

To perform this test, I create an approximate 95% bootstrap confidence interval using 10000 bootstrap resamples of the data. The confidence interval is $[0.28, 0.47]$ which does not contain 0, so I reject the null hypothesis at the 95% confidence level.

The correlation between number of past complaints and probability of future complaint is positive and moderately high. Thus, the data support hypothesis (2).

Insights

Overall, I believe the data support the "few bad apples" hypothesis. It seems likely that most complaints are filed against repeat offenders within the CPD.

Given these results, I would recommend a similar analysis using other police departments' data. Do departments in general contain a small number of officers that produce most of the harm? These results suggest interesting future directions for both data analysis and ethnographies of culture within police departments. Given today's intense scrutiny of policing, I believe that analyses like this are necessary to keep the public informed about how policing actually occurs today.

Code

All code for this analysis resides in https://github.com/SamLau95/chicago-police/blob/master/analysis.ipynb (https://github.com/SamLau95/chicago-police/blob/master/analysis.ipynb) .