

Police Misconduct in Chicago: 2007 - 2017

Chenyi Luo
San Jose State University
Student # 011485825
chenyi.luo@sjsu.edu

Jasmine Lin
San Jose State University
Student # 014692405
chia-ping.lin@sjsu.edu

Lawrence Sinclair
San Jose State University
Student # 014072032
Lawrence.Sinclair@sjsu.edu

Venkata Sai Kusuma Sindhoora
San Jose State University
Student # 014540916
venkatasaikusumasindhoora.vank
ayalasiva @sjsu.edu

Abstract—Police misconduct has been a major concern over the United States and raised global awareness in recent decades. The situation has got even worse since the black-lives-matter movement exploded. Protesters came across cities and states calling for the judgement. The goal of the project is to provide the audience an understanding and reference of the police misconduct in Chicago, along with the breakdown by the individual police to identify the most alleged ten officers misconduct. The dashboard we presented in this project was created with Tableau worksheet and dashboard, following the fundamental Gestalt principles in data visualization. We created an area chart to view the trend of police misconduct by years on top, with the performance index for each officer group on the bottom left and a bar chart for the most alleged ten officers on the right side. The demographic information of each offender and their allegations year trend would show up when a particular officer id within the bar chart is being clicked. Combined with previous related works by other researchers and scholars, we hope to derive meaningful results, insights and increase transparency toward the public.

Keywords—dashboard, police, misconduct, visualization

I. INTRODUCTION

The Chicago Police Department (CPD) has been the source of significant and continuous controversy over the years with regard to alleged and clear police misconduct. Police misconduct has often been characterized by politicians, especially on the Right, as the result of a few bad actors.

The Invisible Institute, a journalism focused non-profit started by journalist Jamie Kalven in Chicago successfully organized an effort to extract large amounts of data from the CPD about officer conduct using lawsuits and Freedom Of Information Act (FOIA) requests. This information was then processed and visualized so that patterns and very specific details, down to document scans, could be accessible to the general public, and journalists.

The Invisible Institute data has been widely used to create pressure for systematic change, generate action to address specific bad actors in the CPD, and as a go-to resource for journalists and the public whenever a Chicago officer appears in the news.

Our dashboard attempts to extend this effort. The objective is to allow users to quickly identify police officers with the most outstanding records of misconduct, allow details about those officers to be surfaced, and place them in the larger time and ethnographic context of the CPD.

II. RELATED WORK

The cpdp.co site [2] attempts to present the Invisible Institute's data on most problematic police officers in two ways. First they show a list which can be filtered by a heat map control (figure 1).

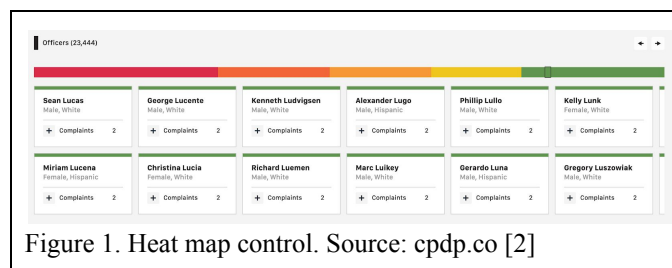


Figure 1. Heat map control. Source: cpdp.co [2]

Second, they show a more detailed list of "Repeaters" (figure 2).

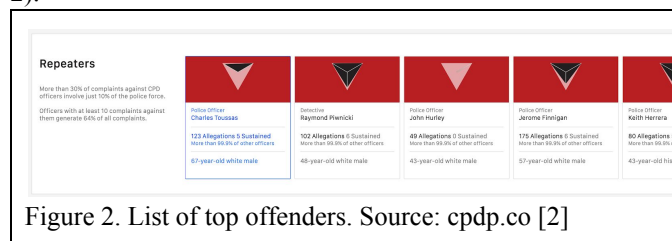
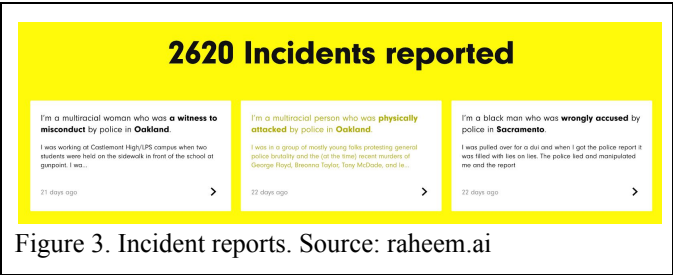


Figure 2. List of top offenders. Source: cpdp.co [2]

Cdpd.co provides a wide range of search and reporting capabilities, but this paper focuses on a better way of presenting top offenders using the same underlying data.

Others researchers on this subject have substantial data, but less advanced visualizations surfacing that data to the general public. For example, Raheem.ai has many reports, and directs them to the authorities, but has only just begun to share this. They are in the process of implementing a solution inspired by

the Invisible Institute’s work. For now, what they have looks like figure 3.



Statistical research on cpdp data by Rivera & Ba show that police accountability, such as provided by the cpdp.io site and Invisible Institute, improves police-public relations while also reducing crime [1].

III. CONTEXT AND SCOPE

To extend research conducted on cpdp.co site, we aim to provide more useful Chicago police misconduct insights by creating a dashboard. Our dashboard will be divided into 3 parts. Each part provides different Chicago police misconduct analyses from divergent directions.

The first part is a quick summarization about the historical CPD performance by utilizing the total allegation count in each of the past year. There are two types of allegations which CPD recorded in their files which are civilian allegation and officer allegation. Our first part line chart also reflects on both types of allegations.

On the second part of our dashboard, we would like to look closely on the total allegation count for each individual police officer. The subject of identifying consistent misconduct by police officers is a broad one even when limited to the CPD. We focus our dashboard on reporting on the misconduct of those officers who might, by various metrics, be candidates for top offenders. Specifically, we report just on the worst ten officers here. Limiting our dashboard in these ways focuses our attention and avoids overwhelming the user with unexceptional details. By identifying the top ten police officers who made the most allegations, we hope this can be a potential warning for our audience. The public should keep their eyes on these police officers if any accidents occurred.

For the last part, we would like to evaluate the police officers’ performance by gender and racial groups. By looking at the gender and racial differences, we expect we can help our audience to narrow down their attention to some particular group so that police misconduct can be more easily monitored.

IV. DATA
A. Source

The original data came from a large number of documents made available by the Chicago Police Department as a result of lawsuits and Freedom Of Information Act (FOIA) requests filed by the Chicago-based journalism advocacy non-profit, the Invisible Institute. These raw documents (see figure 4) were processed by OCR, and by hand, and by automated processes by the Invisible Institute over the course of years. The full dataset was always made public, but was made more accessible by being posted on github for download.

Figure 4. A Typical Source Document from the CPD. Source: cpdp.co

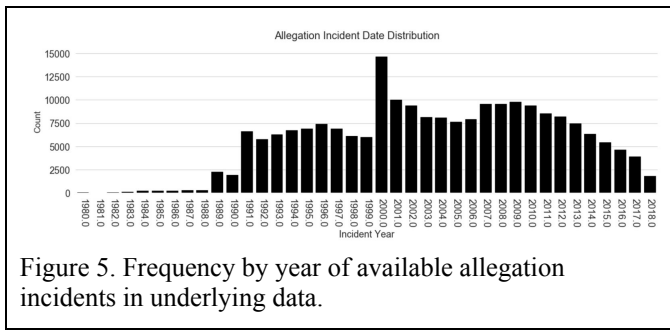
The underlying data systems in the source system that serves the cpdp.co website is documented as a data map: http://users.eecs.northwestern.edu/~jennie/courses/eecs396-dat-a-sci/invisible-institute/cpdb-schema/tables/data_officer.html

We received our data from files created as a result of a direct request to the Invisible Institute. The data came in the form of three .csv files: (1) officers.csv, (2) allegations.csv, and (3) officer_allegations.csv. The officers.csv and allegations.csv files contained all of the data we needed for our visualization. They include 33,839 police officers in the Chicago Police Department and with many personal detail attributes being recorded including names, gender, ethnicity, rank, and year of birth, badge number, police unit, and current salary. In these two datasets, tenure information also contained which including appointment date, resignation date, active status,

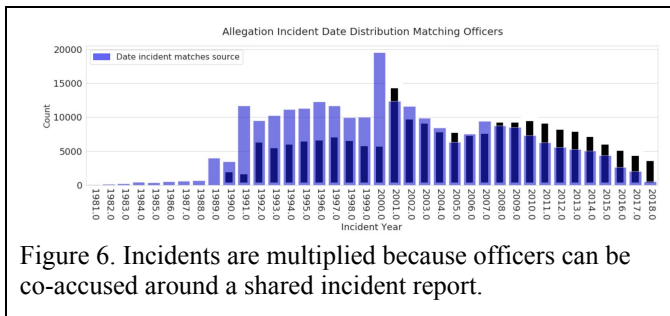
date of first allegation, date of last reported allegation. The other `officer_allegations.csv` provided the relationship between allegations and officers.

Unfortunately, these datasets had limited availability. If an officer was active from 2000 to 2015 then any and all complaints against them should, by law, be included in the `cpdp` data. Some data is also available after that date. However, because of technical limitations to historical data systems, the data before 2000 is less complete. In terms of completeness, the data becomes very sparse before 1990. And is considered unusable during the period before 1980.

Our analysis (figure 5) shows the distribution of source data incident dates, by year, illustrating this underlying data availability.



The number of incidents gets multiplied when looking at the officer level because several officers can share (be co-accused) around the same incident (Figure 6). At the same time, many allegations cannot be matched to a specific police officer, especially in recent years.



Even though we had efficient data after 2000, eventually we only took the data between 2007 and 2017 to do the analysis and build the dashboard because we think that it would be easier for our audience to have an understanding of the police misconduct in the recent ten years.

B. Data Cleaning

The source data mentioned above were not cleaned and robust to our analysis. Some key fields were uncompleted and missing values needed to be imputed and substituted. We also

got rid of irrelevant variables that exist in all three data sets in a few ways before use in Tableau.

Officer.csv:

To clean this dataset, we firstly dropped the unrelated column and only kept the columns related to our analysis such as “`allegation_count`”, “`civilian_allegation`”, “`resignation_date`”, “`race`”, “`appointed_date`”, “`gender`” and etc. In addition, we also trim down our dataset by using 2007-2017 as the constraint. Secondly, we imputed all the missing values that existed in our dataset with “unknown”.

Allegation.csv:

This dataset includes all the information regarding the allegations. We trimmed down this dataset by only keeping the “`crid`”, “`incident_year`”, “`allegation_id`” and “`is_officer_complaint (T/F)`” variables for later analysis.

Officer_allegation.csv:

This dataset explains the relationship between the allegation and officer. It did not need any additional cleaning process and we left it for the later analysis.

V. METHOD

As we mentioned in the Content and Scope section above, our dashboard includes three parts to analyze Chicago police misconduct from divergent directions. After the data cleaning, our datasets were ready for the following analyses:

A. Allegation Counts Per Year

The first analysis we want to conduct is to summarize the total allegation counts by year between 2007 and 2017. The total allegation counts included both `civilian_allegation` which refer to the allegations filed by the general public and `officer_allegation` which was the allegation filed by all police department co-workers. Our analysis would include both types of allegations.

To do this analysis, we referenced the “`officer_allegation.csv`” data. In this dataset, `allegation_id` identifies the individual allegation case happened in the past and `officer_id` identifies the unique police officers. The two attributes are the unique identifier which allowed us to merge the “`officer_csv`” and “`allegation_csv`” files together. We noticed that one allegation case could involve multiple officers and a single police officer could be accused in multiple allegation cases. After the merger, the result showed a total of 88,082 allegation cases had more than one police involved and a total of 173,597 officers involved in multiple cases. We called this new merged dataset as “`allegation_2007_2017_v2.csv`”. With this new dataset, we were able to plot both officer allegation and civilian allegation counts per year between 2007 to 2017 on the dashboard.

B. Average Allegation Count Per Officer

The second analysis, we would like to calculate the average allegation count per each police officer during 2007 to 2017. In this analysis, knowing the exact police tenures for each police officer in their career is extremely important. These durations were calculated by utilizing the appointed_date and resignation_date stated in the “officer_csv”. The resignation_date column includes a lot of missing values. When an officer was still actively working as a police officer, the resignation_date would be a blank cell in our data. As we only analyzed the 2007-2017 data, we assumed every police officer who had their resignation_date blank, we replaced all these cells with 2017. By using the resignation_date subtracting the appointed_date, we had the tenures for each police officer. As a result, the average allegation counts per each police officer could be derived from dividing the allegations count by the tenure. We added these new average allegation counts as a new variable to our existing dataset. We called this modified dataset as “officer_v2_2007_2017.csv” and it would be utilized later for dashboard implementation.

C. Performance Index Calculation

The last analysis for our dashboard is to analyze the police performance among each racial and gender group for the duration between 2007 and 2017. In this analysis, we used two datasets “allegation_v2.csv” and “officer_v2.csv” generated in the previous analyses. Based on the “officer_csv”, there are only two gender types which are “female” and “male” and there are a total of 6 racial groups which are “Asian/Pacific”, “Black”, “Hispanic”, “Native American/Alaskan” and “White”. We have to think of a way to evaluate the performance for these individual groups.

To start the analysis, our initial thought was to measure their performance by looking at the total number of allegation counts in each year by different gender and racial groups. However, comparing the total allegation counts among each gender and racial groups were not telling us the true performance. If a certain group had a larger population than other groups, then that group was expected to have more allegations in generation. It is unfair to compare the allegation counts between different racial and gender groups when their populations in the police force were uneven. We must take the total number of populations in each year between 2007 and 2017 for each gender and racial group into consideration to analyze the performance.

In order to calculate the population by year, we had to look carefully about how many years each police officer worked between 2007-2017. This is because the total number of police officers worked in each year are different. Many police officers joined the police force at the same year as many existing police officers got retired or resigned from their jobs. Therefore, the total number of police officers for each gender and racial group per year was always changing. In the “officer_v2.csv”, each police officer has been indicated both the appointed date and resignation date. These two dates tell us when each police officer joined the police force and when

they left. By utilizing these two date variables in our dataset, we extracted the appointed and resignation year for each police officer and we ignored the month and dates for an easy calculation. We also assumed that if a police officer either joined in a certain year or prior, we would count that police officer in the total population for that year. Similarly, we assumed if a police officer got retired in a certain year or after, we would also include that police officer in that year’s population. For example, there are two police officers who joined the police force on May, 4, 2007 and the other one retired on September 12, 2007. We counted both of them into the 2007 total populations. The table below shows the 2007 population count for each gender and racial group we have calculated.

	gender	race	population_counts	Year
0	F	Asian/Pacific	23	2007
1	F	Black	846	2007
2	F	Hispanic	374	2007
3	F	Native American/Alaskan Native	6	2007
4	F	White	959	2007
5	M	Asian/Pacific	236	2007
6	M	Black	1882	2007
7	M	Hispanic	1682	2007
8	M	Native American/Alaskan Native	32	2007
9	M	White	4584	2007

Table 1. An example to show the population count per each gender and racial group in 2007.

Based on the above calculation, there were only 6 female Native American/ Alaskan police officers and only 1,882 male black police officers worked in 2007. Therefore, we expected female Native American/Alaskan had much less allegation count than male native American/Alaskan police officers. Although the above table only shows 2007 population data, we have calculated all population count in each year of 2007 to 2017.

As the next step, we took the total allegation counts for each gender and racial group in each year divided by their population in that year. It gave us a ratio which indicates the average allegation count per person in that gender and racial group in that year.

	gender	race	population_counts	Year	incident_year	Allegationcounts	ratio	New_ratio
0	F	Asian/Pacific	23	2007	2007.0	5.0	0.217391	0.782609
1	F	Black	846	2007	2007.0	360.0	0.425532	0.574468
2	F	Hispanic	374	2007	2007.0	204.0	0.545455	0.454545
3	F	Native American/Alaskan Native	6	2007	2007.0	2.0	0.333333	0.666667
4	F	White	959	2007	2007.0	462.0	0.481752	0.518248
5	M	Asian/Pacific	236	2007	2007.0	151.0	0.639831	0.360169
6	M	Black	1882	2007	2007.0	1357.0	0.721041	0.278959
7	M	Hispanic	1682	2007	2007.0	1271.0	0.755648	0.244352
8	M	Native American/Alaskan Native	32	2007	2007.0	16.0	0.500000	0.500000
9	M	White	4584	2007	2007.0	3546.0	0.773560	0.226440

Table 2. Shows the ratio and New_ratio columns calculated for the police performance analysis

Based on this ratio column (average allegation count per person) showed above, the lower ratio indicates a better performance. For example, on average, every 0.4255 female black police officer had an allegation in 2007 and every 0.7210 male black police officers had an allegation. This means black female police officers have a better performance overall than black police officers since their ratio is smaller. However, if we chose to use this ratio to evaluate the police officers' performance. When we plot these data into a bar chart on our dashboard, the shorter bar would indicate better performance. However, people have the perception that higher bar means better and lower bar means less in general. In this case, our results would be misunderstood if the audience associates the shorter bar with poorer performance instead of better performance. Therefore, instead of using ratio as the performance indicator, we used a performance index which is calculated by 1-ratio to measure the police officer's performance. The calculation results showed on the above table as the "New_ratio" column. Higher New_ratio indicates better performance and lower New_ratio value shows poorer performance. For this analysis, we recorded all our calculation results into a new dataset named "AllegationPopulation_ratio_2007-2017_v2.csv" and it would be used later for the dashboard implementation.

VI. VISUALIZATION AND RESULT

Our databoard includes three sections and they represented the analyses we had conducted above.

A. Allegation Count Per Year Chart

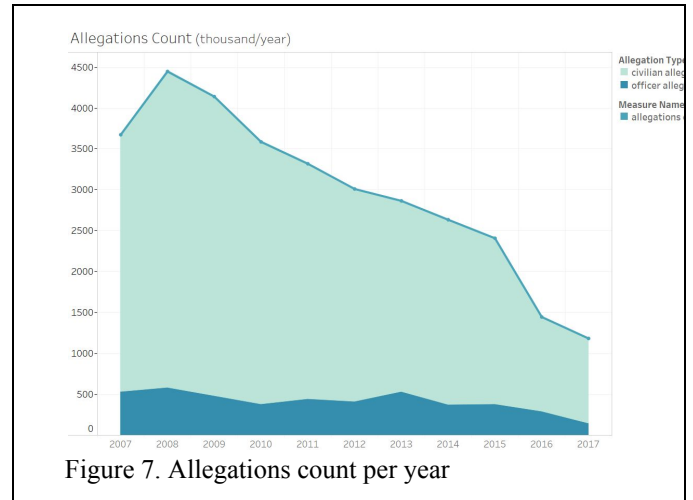


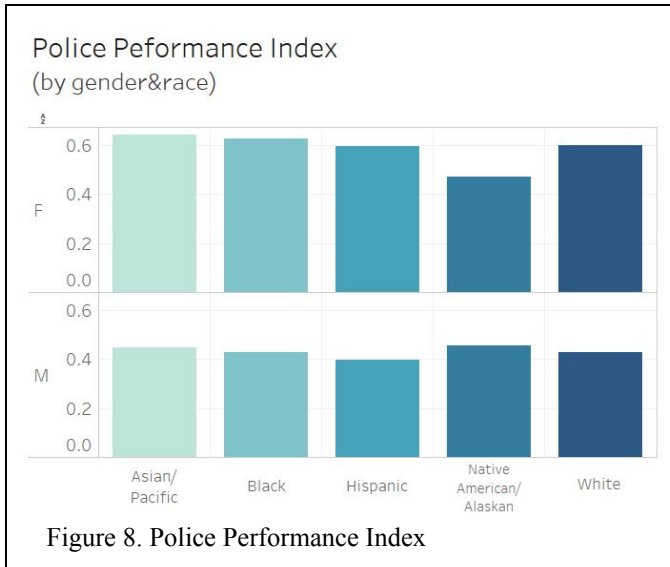
Figure 7. Allegations count per year

This chart is associated with the analysis of Allegation Counts Per Year mentioned above and it was plotted by using the "allegation_v2_2007-2017.csv" dataset.

The x-axis on the chart represents a different year from 2007 to 2017. The y-axis represents allegation counts. On the chart, the top line indicates the total allegation per year. Along the line, different year points were highlighted in blue. On the dashboard, our audience could easily look at the exact allegation count numbers per year by moving their cursors to these individual year points. The information will pop up to illustrate the total number of allegation counts in the particular year they pointed.

Furthermore, the light blue shaded area under the curve represents the civilian allegation counts per year. The darker blue shaded area under the curve represents the officer allegation counts per year. The line chart illustrates the fluctuation of the allegations over the city of Chicago. By viewing the line chart, we observed that the total allegations counts reached the peak in 2008 and kept dropping to a lower point toward 2017. It shows an overall decreasing trend which means the total number of allegations happening at CDP per year had decreased over the years between 2007 to 2017. This decreasing trend also reflects on both civilian allegation counts and officer allegation counts. To look closer at each individual type of allegation counts, although the officer allegation count still shows a decreasing trend from 2007 toward 2017, it is comparatively consistent over these years. Unlike the total allegation counts per year, the officer allegation did not show an obvious peak in 2008 and also did not show a dramatic drop afterward. The civilian allegation counts are very similar with the total allegation count. It also showed its peak in 2008 and kept dropping afterward. Through the comparison, we could tell that the majority of the allegation came from the civilian allegation counts. It was not surprising to see that most of the alleged cases were accused by the public rather than other people who also worked as police officers.

B. Police Performance Index Chart



This chart is associated with the analysis of the Performance Index mentioned previously and it was plotted by using the “AllegationPopulation_ratio_2007_2017_v2.csv” dataset.

The x-axis on this bar chart shows each individual race group and y-axis represents the New_ratio as the performance index we have calculated earlier. This bar chart includes two parts. The top section represents the female group and the lower section represents the male group. We used the different color here to distinguish different racial groups, but we used the same color for the individual racial group between male and female sections. Therefore, when our audiences visualize this bar chart, they can easily compare each racial group performance within each gender group and they also see the differences between two gender groups for each individual racial group.

Additionally, this bar chart includes an interactive feature which this chart is linked with the Chart A (Allegation Count Per Year). By clicking the years along the x-axis of Chart A, the allegation counts for each gender and racial groups shown on this chart would be changed accordingly for each year between 2007 and 2017. We think this interactive feature will bring convenience to our audience so that they can select which year they would like to explore without overwhelmingly seeing all the year's information together.

By clicking through all the years between 2007 and 2017, we compare the allegations count between each gender and racial groups, the chart shows that every female racial group had much better performance than their male police officer group in each year with an exception of female Native American/Alaskan. Female native American/Alaskan only performed better than males in 2007 and 2010 - 2012. In 2008, and 2013-2017, male American/Alaskan performed better than their female racial group. In 2009, there was no

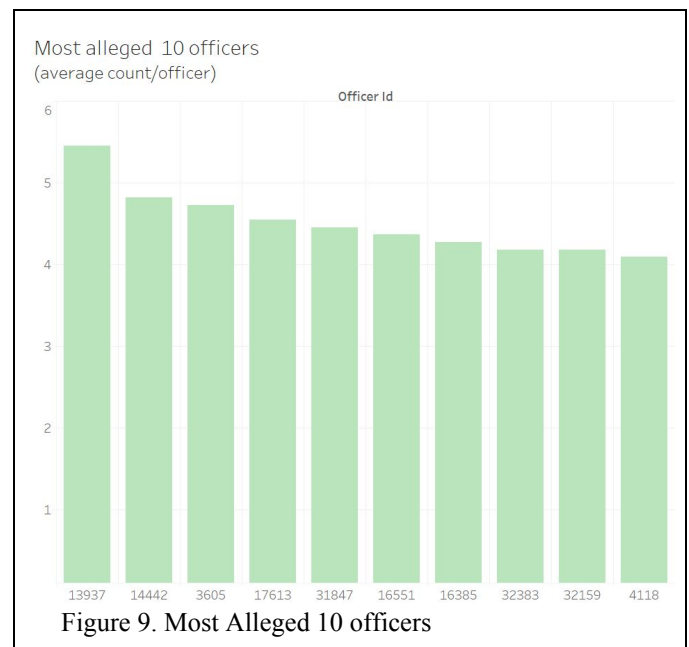
male American/Alaskan police officer, so no comparison could be made for that year. When comparing between each racial group, we did not see any outstanding racial groups. It seemed that the performance between each racial group varied in each year. No particular racial group had performed better or worse than other racial groups over these 11 years. For example, Female Asian police officers ranked the top in 2007 in terms of performance, but they had the lowest performance index among all other female racial groups in 2015. This inconsistency could not show a clear trend to indicate the performance of female racial groups. They could perform higher or lower in different years. This situation happened similarly in other gender and racial groups. Overall, the chart shows that every racial group including both male and female police officers showed improvement over the 11 years. Their performance index shows an increasing trend between 2007 to 2017.

C. Most Alleged Ten Officers

This part of the dashboard is associated with the analysis of the Average Allegation Count Per Police Officer mentioned previously and it was plotted by using the “officer_v2_2007_2017.csv” dataset.

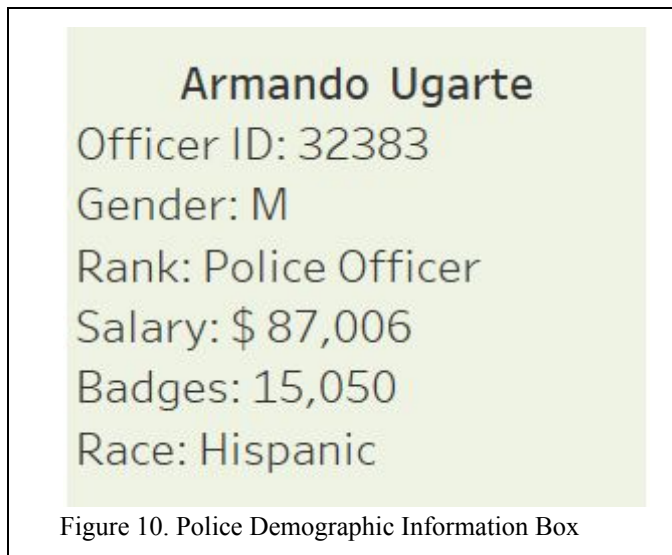
While Chart A greatly gives us a view of the summary of police misconduct for each year, it doesn't provide a detail of the officer. Since we don't have enough space to display all officers who have once been reported, we report the ten officers that are being complained most frequently. There are three subsections which demonstrate the most alleged ten officers, their demographic information and the allegations count by year.

1) Average Allegations count



The purpose of this bar chart is to identify the top alleged officers who are involved in the frequent misconduct. To plot this chart, we have computed the average allegations done by each police officer per year based on their overall tenure. The reason why we calculated the ratio here is otherwise the police officer with longer working experience would always have more allegations when compared to the officer who has just joined recently. The horizontal axis lists the most alleged 10 officer id's and the vertical axis is the average count of allegations recorded per officer, year. We can see the officer id - 13937 ranked first with an allegation count per year as 5.45 and the officer-id - 4118 ranked 10th with an average allegation count per year as 4.091. The color is carefully selected to appear friendly to colorblind people.

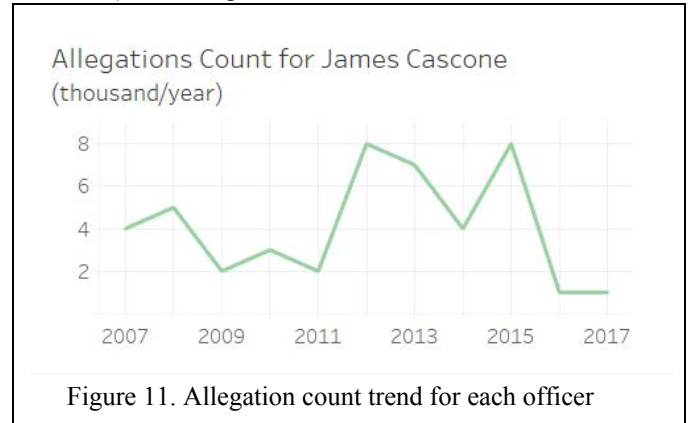
2) Officer demographic Information



The officer demographic information box contained officer ID, officer name, gender, race, rank, salary and badges. We took the demographic information for each officer from officers.csv file. The box would be triggered and popped up by clicking the officer ID of the above most alleged ten officers bar chart. We use the parameters filter to link the information box and the most alleged ten officers chart. To do this, we need to create a bunch of parameters in our data source. And bring the attribute we are going to assign to each parameter to the detail mark. Then add a dashboard action for each parameter to assign the attribute value so if we insert the parameter anywhere in the dashboard, it would link to the corresponding attribute. The purpose of listing the information of the worst officers is to give the public a warning and force the officer to behave themselves. We changed the background color to light green so the box can jump out from the white background. We didn't use a dark color because we wanted to keep a consistency of our dashboard. It is a trade-off between having contrast colors as background and object to follow the

figure-ground principle of gestalt, and decreasing the effect to make the whole dashboard more readable.

3) Allegations Count trend for each Officer



The line chart gives information about the overall trend of allegations done by each officer per year. This is an interactive chart and can automatically show the trend when we click a particular officer id from the most alleged 10 officers bar chart. If we do not select any officer, then it will show the summary of allegation count recorded per each year. The allegation count shown on the vertical axis is extracted from officers.csv file whereas the incident year on the horizontal axis is extracted from allegations.csv file. Here the filter action is performed on the officer id field and we have created a dynamic title that can automatically change the name of the officer id based on the user selection. Thus, the usage of custom titles will enhance the readability of the charts and the filter action will give the users precise views of the pertinent data selected. The purpose of the chart is to identify if there is any systematic pattern of misbehaviour recorded for the police officer.

D. Dashboard Visual Qualities



We evaluated our dashboard in the context of (1) use of color for information conveyance (including user accessibility for the color blind and color impaired), (2) data density considerations such as Tufte's concept of 'data ink', (3) the six(or seven) principles of gestalt

1) Use of color:

- We applied color blind friendly palettes to the entire dashboard and slightly changed the light and transparency so every chart looks organized but is differentiable. (We should use the color blindness simulator to check)
- We use different contrast colors to represent different types of allegations in the area chart.

2) Data density considerations:

We followed the Data ink ratio theory to design our dashboard. The data-ink ratio is the proportion of Ink that is used to present actual data compared to the total amount of ink (or pixels) used in the entire display. A Good graphic should include only data-Ink. In our dashboard, we reduced all the non-data ink as possible, the majority of our content is undeletable and useful for understanding the dashboard. We moved the headers of x-axis and y-axis to the chart titles, and keep the information of each officer bold and clear.

3) Gestalt principles:

- Similarity for the top line and area charts and the bar plot in the left bottom. We use blue-teal for the top and left charts as these two charts both summarize the allegations happening over the eleven years and provide a quick understanding of the police misconduct for our audience.
- Similarity for the police performance index. The blue teal palette allocates five colors to the five racial groups, but with the same racial group we used the same color for both female and male. As a result, the color representing hispanic male and hispanic female is the same and with stronger similarity.
- Similarity for the Most alleged ten officers and the corresponding demographic information and allegations count for individual police officers.

VII. DISCUSSION

Based on our results, Chart A (Allegations count per year chart) tells us that the police misconduct in the city of Chicago has been gradually decreasing from 2007 to 2017. By only considering allegations as the performance indicator, it

indicates that the police performance has been improving. However, there are many other factors that can be used to evaluate the police officer's performance. For example, the number of compliments received by each police officer can be also used to see whether they had a good performance or not. To conclude that CDP had an improving performance from 2007 to 2017 only based on the allegation count can be a little bit biased.

Based on Chart B (Police Performance Index) results, we saw a trend that all female police officers except for the female Native American/Alaskan group always performed better than male police officers. This result shows the gender difference was higher related with police officers' performances. This correlation between gender and performance could inspire the Chicago police department to take some considerations of hiring more female police officers and even promote more female officers to be the supervisor to bring more effective management. This can be one of the potential solutions to reduce their police misconduct incidents and improve their police services. Moreover, the audience might have a bunch of questions in their mind. While the female officers tend to perform better than the male officers, do they have higher positions or do they usually get promotion? Because of the glass ceiling, the salary and opportunity differ between genders. However, the performance of the female officers in the city of Chicago can become a mirror for women who struggled themselves in society and industry and encourage them to keep doing an incredible job.

Based on the Chart C (Most Alleged Ten Officers), the top 10 bad performers in terms of allegation counts between 2007 and 2017 in the Chicago Police Department had been identified. Among these 10 police officers, the majority of them are male and most of them are White and Hispanic. Additionally, the individual allegation time series chart of the police officer can help the administrators in decision making about their accountability. Based on this chart we can derive insights whether the pattern of systematic misbehavior is being continued or is observed only for a shorter period. Thus, from these results police administrators can take action on the alleged officers such as oral reprimand, temporary suspension, disciplinary action, penalty, demotion of their rank etc as a means of rectifying their inappropriate conduct or behavior.

VIII. FUTURE WORK

In this project, we have demonstrated overall allegation trends for 2007 to 2017 and identified top alleged officers that can be further drilled down to a detailed level of individual officer's demographic information, allegation trends. However, our initial plan was to create an interactive map that can show the allegations recorded for different cities in the Chicago region. But our data source does not have the necessary geographic information such as location names, latitude, longitude data points, we failed to implement the map. This can be

incorporated as part of the future work when we have the required location data.

The other data limitation which we have observed is there are very few data before 1990 when compared to the overall population. For the sake of simplicity, we have ignored this period data and not included it in the project. But this problem can be solved by implementing a weighting technique, where a multiplier is used to project the sparse data to the overall population data. The key benefit of this will make the data re-balanced and it improves the accuracy of results as it is pertaining to a broader population.

There are many missing values found in the data. For instance, appointed date and resignation date are important features for calculating the police officer's tenure information and they cannot be unknown. The better approach to handle these missing values can be to use an imputation technique or we can even build a predictive model that can estimate the replacement for a missing value through machine learning algorithms. This kind of enhancement can make the system smart and improve its efficiency in results.

The other charts which we could think of to implement is to provide the summary of police officers groups based on other features such as rank, salary, badge etc. However, these fields have a lot of missing values and we will be able to do it in future once we have all the values populated for this data.

Another feature that we hoped to consider in our next step is to take the complainants into account. The black lives matter movement strongly alleged that a large number of officers and law enforcement agencies have been targeting their community, family and children for a long time. The bias and discrimination over black and colored people were hard to eliminate and was time consuming. Combined with the demographic information of the complainants, it is more likely that we can find a pattern in knowing whether a bias toward a particular group still exists.

Apart from the data limitations, we have faced software limitations that hinders from doing deeper analysis in this project. The tool tableau has poor BI capabilities and there is always some manual effort needed whenever we want to update the data in the back end. The IDE's such as PyCharm, RStudio will enable us to do complex statistical analysis with the help of machine learning and deep learning models.

IX. WHAT DID WE LEARN

There are a lot of things we learned through this project that can be simply separated into two parts. The first part is all the soft skills we have learned from the data. The second part is the technical skills we have learned by utilizing both analytical and visualization tools to fulfill our project mission.

Soft Skills

Through this project, we experienced that data is one of the most important components within a project. Accessible and understandable data enhances our efficiency and provides convenience to do further processing and data visualization. However, the data can be biased. In three datasets we had for this project, each police officer had given its allegation count; appointed date and resignation date. For the Chart B (Police Performance Index) and Chart C (Most Alleged 10 Officer), we could not simply use the allegation count on our charts. It is because these numbers are biased and it could not deliver the true information to our audience. We had to also take the officer working duration into account when doing the calculation. When we were working on this project, we had been extra careful so that our information shown on the dashboard would not mislead our audience. By considering all these small elements, we have gained a lot of critical thinking.

Additionally, working together with all the group members to execute such a big project, a lot of planning works required. Project planning comes handy and it allows us to always keep tracking on our project progress and foresee our project goal. Through this whole process, we all gained communication and planning skills through practical experiences.

Overall, all these soft skills definitely will be beneficial to us even for the future.

Technical Skills

To create a dashboard, we heavily used Tableau in this project. At the beginning of the project, none of our team members were familiar with Tableau. It's a great opportunity for us to get hands-on experience on data visualization tools. Tableau provides an intuitive and user friendly interface for people like us to play around. There are three main components in Tableau (1) worksheet, which we implemented our data and build any chart we like, (2) dashboard, where we design the overall look of our presentation, (3) story, a sequence of visualizations that work together to tell narratives and support the context. Some useful and interesting functions we applied on our dashboard are the dashboard actions. The dashboard actions help to create interactive relationships between data, dashboard objects. The filters and parameters we implemented in our dashboard improve the continuity and consistency between our charts so the story behind the whole dashboard could be more concise. Besides the software perspective, the principles we learned from the data visualization class helps us to design the layout and improves the readability and interpretability of our dashboard. The color and data density could significantly affect one's experience in learning the information from the dashboard.

Besides learning Tableau, we have also gained a lot of practical experience of using Python. In the beginning of this project, we had to do a lot of data cleaning and data aggregation which all could not be done easily on Tableau.

Therefore, we have used Python to perform all the pre-data-visualization tasks. Python is a very powerful analytical tool and it comes with various Python libraries. During the data processing stage which we have mentioned above, we heavily used Python by utilizing the Pandas and NumPy Python library. They allow us to import our data and do some data explosive analysis and later export our data to a new csv file.

Tableau and Python are all good for working on a data analysis project. Tableau is more emphorize on presenting the analysis result through data visualization and Python is more towards complicated analysis through coding. We are glad that we have the opportunity to learn and practice both tools in this project. Most importantly, we now know how to apply our intelligence to cooperate different analytical tools with different functionality together to fulfill our project goal.

REFERENCES

- [1] R. G. Rivera, B. A. Ba, "The Effect of Police Oversight on Crime and Allegations of Misconduct: Evidence from Chicago" (2019). *Faculty Scholarship at Penn Law*. 2109. https://scholarship.law.upenn.edu/faculty_scholarship/2109
- [2] CPDP, 03-Oct-2020. [Online]. Available: <https://cpdp.co/>. [Accessed: 03-Oct-2020].
- [3] 7 Gestalt principles of visual perception: cognitive psychology for UX [Online]. Available: <https://www.usertesting.com/blog/gestalt-principles>
- [4] How to Use Color Blind Friendly Palettes to Make Your Charts Accessible [Online]. Available: <https://venngage.com/blog/color-blind-friendly-palette/>