

Sarah Ahmad (saz848)
Vikram Kohli (vkp260)
Junhao Li (jlo708)

Checkpoint 3: Workflow Analytics

Question 1. Is there a correlation between race, gender, and length of appointment for officers?

We set up the query in a way so that it's easy to switch between looking at different races and genders and then visualize the distribution of the length of appointments. We chose to visualize the length of appointments for the chosen combination of race and gender using a bar chart, with lengths in bins in ranges of ten years.

Next, we compared the distributions of the length of appointment between different races but the same gender and the same race but different genders. From the bar charts, it seems that the distributions between the same races but different genders matched up pretty well - for example, Hispanic male and female officers were mostly appointed between 0 and 10 years, and the number of officers in each subsequent bin exponentially decreased.

When comparing the length of appointments between race, white males generally had the longest length of appointments, usually between 20-40 years. The shortest length of appointments was seen with Hispanic males and Asian/Pacific males, as they mostly had been appointed for 0-20 years. Black males were somewhere in between, with most of their length of appointments being less than 30 years.

The fact that the distribution of the length of appointments between gender was fairly similar seems that it might be able to tie in with our findings from Checkpoint 1. In Checkpoint 1, we found that the proportion of allegations that ended up with punishment for both female and male officers was about the same. This might lead into another question: whether the officer's appointment ends after they receive a punishment after an allegation. If this ends up usually being the case, we might be able to reason further about why the length of appointments between gender match up and the proportion of allegations that end up with punishment for both men and women are the same.

Question 2. Are the groups of officers with the highest allegation rates by beat often of the same socioeconomic location?

When attempting to answer this question, we had difficulties with extracting a metric for the economic situation of officers in individual beats. Because of this, we chose to examine whether the officers of the highest allegation rates were of the same race and gender instead of socioeconomic location.

The metric we used to accomplish this was to find, in each beat, the single officer with the greatest amount of complaints. We then used this information to determine, by race and gender, which officers were continually found to be the ‘worst’ by the complainants; i.e., which officers had the largest number of allegations levied against them.

The results we found seem consistent with the information we had previously found; in almost half of all beats, the officer with the most allegations against them was a white man. We additionally found that, in 25% of all beats, the most-complained officer was a Black man. This finding is significant given some of our earlier queries on the racial/gender makeup of the police force; although only 15% of officers are Black men, they represent 25% of the most-complained about officer in each beat.

Our results also show clearly that, overall, the ‘worst officer’ in a beat is much more likely to be a man than a woman; White, Black and Hispanic men make up 87% of the officers with the most allegations in their beat. This result is most likely due to the gender gap in the police force, as there are overall far fewer women officers than men. That said, when compared to our findings from Checkpoint 1 our findings may still be significant; despite there being only around 3 times as many Black men than Black women on the police force, amongst officers who are the most complained about in their beat, Black men outnumber Black women by 6 times (there is a similar, though less extreme, discrepancy, between White women and men).

We looked at complaint sources, allegation categories, and allegations by year in an attempt to examine why black male officers were overrepresented. In terms of complaint sources, the proportion of allegations against black male officers from their peers was 7.3% which was not far from the proportion of 6.8% for all allegations, although the proportion is larger.

Black male officers were also accused of personnel violations more often than the population of all officers with 37.2% and 33.0% of allegations being personnel violations, respectively. However, again it is harder to identify this difference as a symptom of an underlying phenomenon.

When looking at the percentage of allegations that were against a black officer by year, we found that the percentage generally fell between 30% and 40%. However, allegations with black officers made up a much higher percentage of allegations in the early 90’s than in recent years.

From 1989 to 1996, allegations with black officers made up more than 36% of allegations while that number has been less than 32% from 2014 to 2017. This does seem to suggest that the conditions causing black officers to be represented have lessened in degree over time.

Similarly interesting are our results relating to Black and White women; although there are almost 1.5 times as many white women as Black women in the police force overall, around an equal number of ‘most-complained-about’ officers per beat are Black women and White women (and in fact, there are slightly more Black women than White women who are most alleged against).

Question 3. How do allegations in beats cluster when connected based on having the same race of officer and race of victim?

To answer this question, we used Trifacta to create a table of allegation IDs with boolean columns denoting the races of the officers receiving the allegation and the races of the victims. We setup our query to see the counts of allegations within a beat involving each combination of alleged officer race and victim race.

From our comparisons of distributions between beats, it seems that the distributions of allegations containing each combination of officer race to victim race varies significantly between beats. While some beats have had allegations in a wide range of the combinations such as beat 268 which has allegations in 12 of the 25 combinations, some beats have allegations in only very narrow ranges. For example, all the victims of the allegations in beat 272 are listed as black. This may be a result of the demographics of each beat varying but more research could be done to examine how the demographics of the population and of officers operating in a beat affect the distributions.

A general trend across the majority of beats was that combinations involving black victims were most common and often had the highest counts for any group of combinations by officer race of the allegation. This finding was expected given our previous observation in Checkpoint 1 that the majority of the victims within the CPDB were black. Similarly, allegations involving asian, pacific islander, native american, or alaskan natives (officers and victims) were low across all beats.

QUERIES

Note: Databricks wouldn't import a SQL database so we used PostgreSQL to export the data_officer table to a CSV

Question 1:

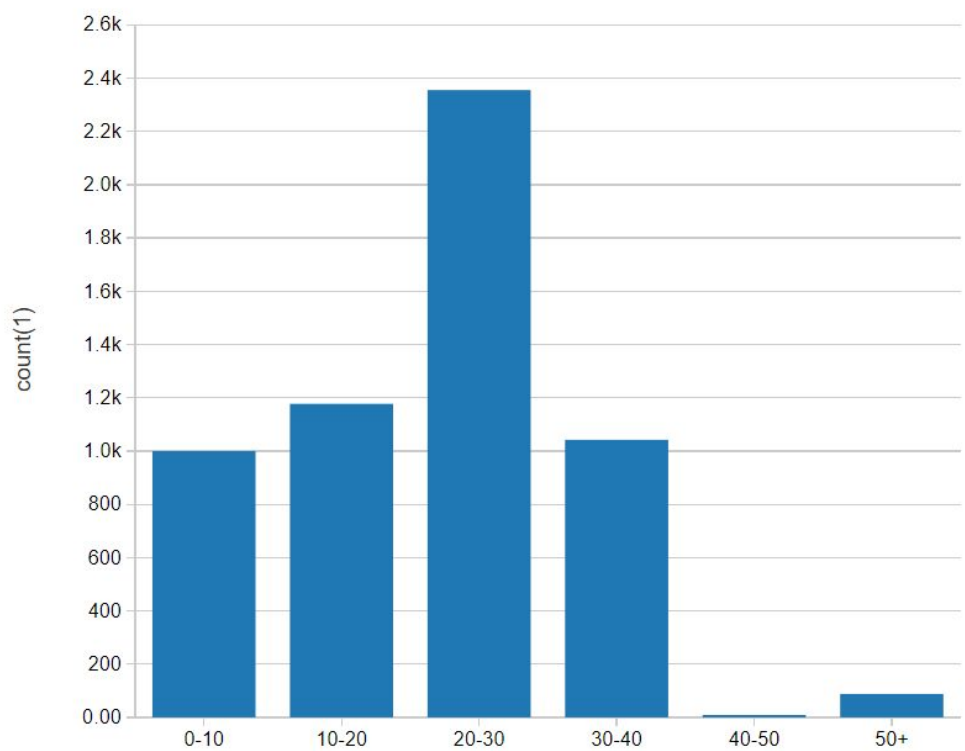
Notebook link:




<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4988437523563813/1767326853905956/4152000852923769/latest.html>

Correlation between race, gender, and length of appointment for officers (Python)

Detached File View: Code Permissions Run All Clear

```
1 %sql
2
3 /* Query the created temp table in a SQL cell */
4 WITH gender_loa AS
5 (SELECT
6   gender, race,
7   CASE
8     WHEN active='No' THEN year(resignation_date) - year(appointed_date)
9     ELSE 2018 - year(appointed_date)
10  END as length_of_appointment
11 FROM data_officer
12 WHERE gender = 'M' and race='Black')
13 SELECT
14   CASE
15     when length_of_appointment between 0 and 10 then '0-10'
16     when length_of_appointment between 10 and 20 then '10-20'
17     when length_of_appointment between 20 and 30 then '20-30'
18     when length_of_appointment between 30 and 40 then '30-40'
19     when length_of_appointment between 40 and 50 then '40-50'
20     else '50+'
21   END as Range,
22   Count(*)
23 FROM gender_loa
24 GROUP BY Range
25 ORDER BY Range ASC
26
27
```



   Plot Options... 

Command took 1.25 seconds -- by sarahahmad2019@u.northwestern.edu at 10/30/2018, 5:21:49 PM on analysis

Question 2:

Notebook link:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4988437523563813/958314614142890/3062227156193771/latest.html>

Worst Officers by Beat (Python)

Detached File View: Code Permissions Run All Clear

Cmd 5

```
1 %sql
2 WITH allegation_counts AS (
3     SELECT officer_id, beat_id, COUNT(allegation_id) as account
4     FROM data_allegation, data_officerallegation, data_officer
5     WHERE data_allegation.id=data_officerallegation.allegation_id
6     AND data_officer.id=data_officerallegation.officer_id
7     GROUP BY officer_id, beat_id
8 ),
9 recent_salary AS (
10    SELECT data_salary.officer_id, salary
11    FROM data_salary
12    JOIN (
13        SELECT officer_id, MAX(year) as year
14        FROM data_salary
15        GROUP BY officer_id
16    ) most_recent ON most_recent.officer_id=data_salary.id AND most_recent.year=data_salary.year
17 )
18 SELECT beat_id, ac.officer_id, race, gender, salary, account
19 FROM allegation_counts ac
20 JOIN data_officer o ON o.id = ac.officer_id
21 JOIN recent_salary s ON s.officer_id = ac.officer_id
22 WHERE beat_id IS NOT NULL
23 AND account > 1
24 ORDER BY beat_id, account DESC
```

▶ (1) Spark Jobs



Question 3:

Notebook link:

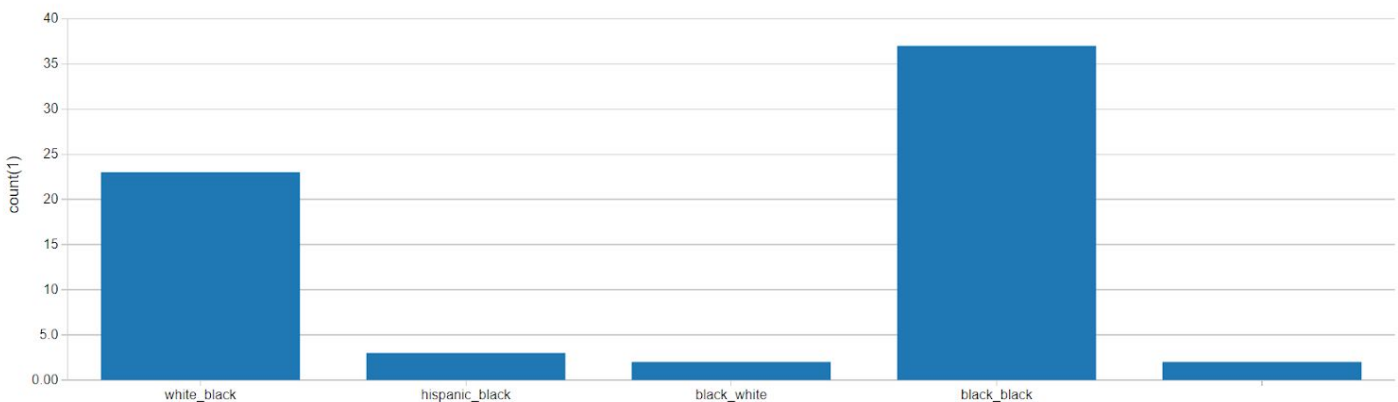
<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4988437523563813/659226948362535/3062227156193771/latest.html>

Clustering on Officer and Victim Race (Python)

Attached: Jun's Cluster File View: Code Permissions Run All Clear

```
1 %sql
2 SELECT
3     CASE
4         WHEN if_officer_white='Y' AND if_victim_white='Y' THEN 'white_white'
5         WHEN if_officer_white='Y' AND if_victim_black='Y' THEN 'white_black'
6         WHEN if_officer_white='Y' AND if_victim_hispanic='Y' THEN 'white_hispanic'
7         WHEN if_officer_white='Y' AND if_victim_asianpac='Y' THEN 'white_asianpac'
8         WHEN if_officer_white='Y' AND if_victim_native='Y' THEN 'white_native'
9         WHEN if_officer_black='Y' AND if_victim_white='Y' THEN 'black_white'
10        WHEN if_officer_black='Y' AND if_victim_black='Y' THEN 'black_black'
11        WHEN if_officer_black='Y' AND if_victim_hispanic='Y' THEN 'black_hispanic'
12        WHEN if_officer_black='Y' AND if_victim_asianpac='Y' THEN 'black_asianpac'
13        WHEN if_officer_black='Y' AND if_victim_native='Y' THEN 'black_native'
14        WHEN if_officer_hispanic='Y' AND if_victim_white='Y' THEN 'hispanic_white'
15        WHEN if_officer_hispanic='Y' AND if_victim_black='Y' THEN 'hispanic_black'
16        WHEN if_officer_hispanic='Y' AND if_victim_hispanic='Y' THEN 'hispanic_hispanic'
17        WHEN if_officer_hispanic='Y' AND if_victim_asianpac='Y' THEN 'hispanic_asianpac'
18        WHEN if_officer_hispanic='Y' AND if_victim_native='Y' THEN 'hispanic_native'
19        WHEN if_officer_asianpac='Y' AND if_victim_white='Y' THEN 'asianpac_white'
20        WHEN if_officer_asianpac='Y' AND if_victim_black='Y' THEN 'asianpac_black'
21        WHEN if_officer_asianpac='Y' AND if_victim_hispanic='Y' THEN 'asianpac_hispanic'
22        WHEN if_officer_asianpac='Y' AND if_victim_asianpac='Y' THEN 'asianpac_asianpac'
23        WHEN if_officer_asianpac='Y' AND if_victim_native='Y' THEN 'asianpac_native'
24        WHEN if_officer_native='Y' AND if_victim_white='Y' THEN 'native_white'
25        WHEN if_officer_native='Y' AND if_victim_black='Y' THEN 'native_black'
26        WHEN if_officer_native='Y' AND if_victim_hispanic='Y' THEN 'native_hispanic'
27        WHEN if_officer_native='Y' AND if_victim_asianpac='Y' THEN 'native_asianpac'
28        WHEN if_officer_native='Y' AND if_victim_native='Y' THEN 'native_native'
29    END as race_to_race,
30    COUNT(*)
31 FROM allegation_races
32 WHERE beat_id = 275
33 GROUP BY race_to_race
34 ORDER BY race_to_race DESC
```

▶ (1) Spark Jobs



Plot Options...