Checkpoint 1: SQL Analytics

Team: The Powerful Turtles

Introduction

In this project, we would like to study how the demographics comparison between the complainants and the respective police officers reported correlating to the number of complaints. In this checkpoint, we extracted and analyzed the demographic data of officers, complainants, and their corresponding community. Then, we performed some inter-group demographic analyses to observe some correlation within these data. The data and findings from this checkpoint will be the foundation of the succeeding checkpoints.

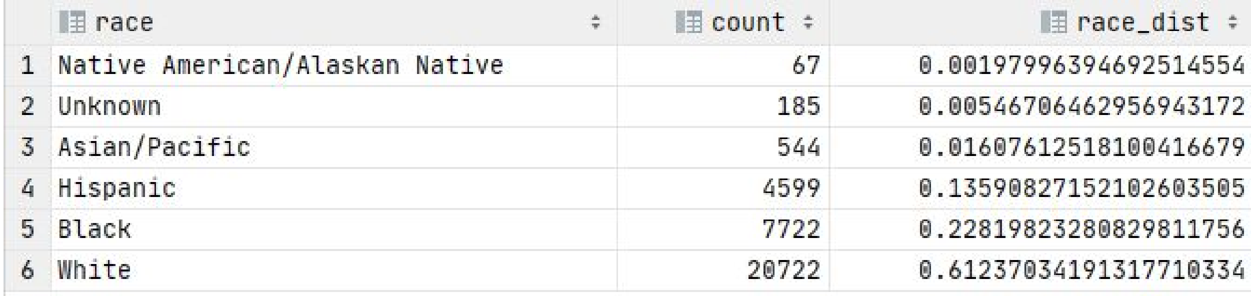
Relational Analytics Questions

In this checkpoint, we would like to address the following questions:

1. *What is the race distribution among the officers?*
2. *What is the majority race in the community of which the officer with most complaints?*
3. *What is the race distribution among the complainant?*
4. *What percentage of officers have the same race as the majority race of their responsible community?*
5. *What percentage of officers have the same race with the filer of their complaints?*
6. *What percentage of filers of complaints have the same race as the majority race of their community?*

Results

1. *What is the race distribution among the officers?*



	race	count	race_dist
1	Native American/Alaskan Native	67	0.00197996394692514554
2	Unknown	185	0.00546706462956943172
3	Asian/Pacific	544	0.01607612518100416679
4	Hispanic	4599	0.13590827152102603505
5	Black	7722	0.22819823280829811756
6	White	20722	0.61237034191317710334

- ➢ In total, six classes of the race were found among the officers, including Unknown as one of the categories. Most of the officers are white, followed by black and Hispanic. From the table, we know that white officers represent 61.23% of all officers, far exceeded the sum of other types, where we can see the unbalance.
- ➢ We got the answers by first grouping the races in officer groups and then dividing the total amount of the offices and calculating the distribution.
- ➢ Therefore, the race distribution among the officers is:

White:	61.23%
Black:	22.82%
Hisapnic:	13.6%

Asian/Pacific:	1.61%
Unknow:	0.54%
Native American/Alaskan Native:	0.20%

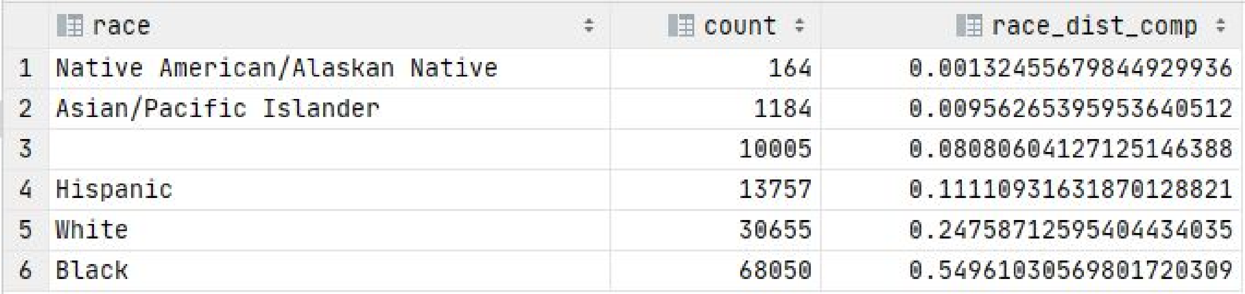
2. *What is the majority race in the community of which the officer with most complaints?*

- ➢ After poking around in the database further, we found out that there is no record of where an officer is stationed. The area table only stores the commander of the area, not all officers. Therefore, we decided to use the area where the officers received the most complaints as their location.
- ➢ We decided to look at the top 1000 officers. During our tests, the top 500, 100, and 50 officers also exhibit the same pattern/distribution.
- ➢ We had to take a detour through the data tables to step 1: get the officer's active location; step 2: get the majority race for each community; step 3: compare the two tables to get the result. The detailed step-by-step explanation is written in the Q2.sql file as comments.
- ➢ The reason why the numbers add up to more than 1000 is that sometimes an officer's most received complaints are in two races of communities in equal numbers (officer id 352, Black: 16, Hispanic: 16), so we decided to just include both because it may cause bias if we choose only one.
- ➢ The majority race in the community of the officers with the top 1000 complaints is:

Black:	691
Hispanic:	208
White:	116

- ➢ The number one is Black communities, second is Hispanic, while White is third. The number for Black communities far exceeds the other two. Because the "officer with most complaints" implies they are big repeaters of their offenses, this result means that the Black communities are very prone to some officers committing repeated offenses in the same area.

3. *What is the race distribution among the complainant?*



	race	count	race_dist_comp
1	Native American/Alaskan Native	164	0.0013245567984492936
2	Asian/Pacific Islander	1184	0.00956265395953640512
3		10005	0.08080604127125146388
4	Hispanic	13757	0.11110931631870128821
5	White	30655	0.24758712595404434035
6	Black	68050	0.54961030569801720309

- ➢ The same as Q1, In total, six classes of the races were found among the complaints. We can learn from the table that most of the complaints are in black (more than half), followed by white and Hispanic. The table shows the opposite results in the majority of officers and complaints, which implies that people in white can have more work, while black people have more complaints about the officers.

- We got the answers by first grouping the races in complaints groups and then sum the total amount of the complaints to calculate the distribution.
- Therefore, the race distribution among the officers is:

Black:	54.96%
White:	24.76%
Hisapnic:	11.11%
Unknow:	8.08%
Asian/Pacific:	0.96%
Native American/Alaskan Native:	0.13%

4. What percentage of officers have the same race as the majority race of their responsible community?

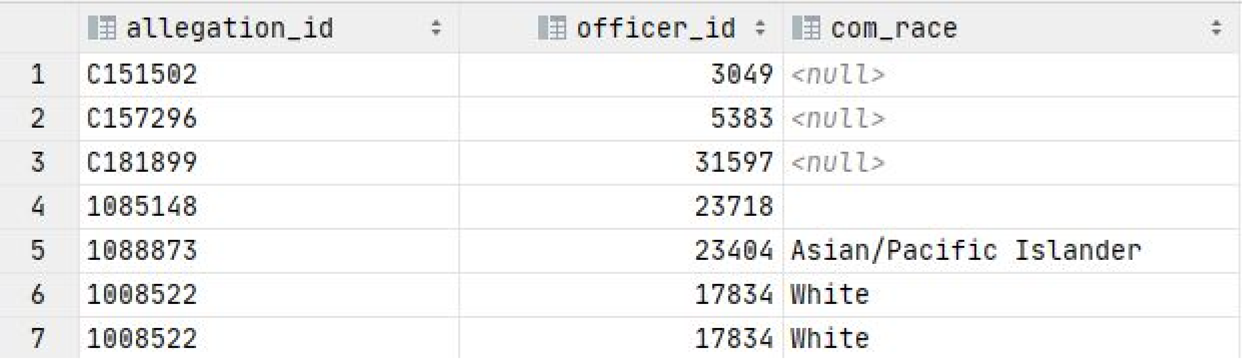
- As stated in Q2, our investigation revealed that there is no record of where an officer is stationed, and we have to use the area in which the officers received the most complaints as their location. Therefore, **this creates a bias**, as this may not actually be their responsible community, but only the community where they received the most complaints from. Our question is thus closer to "*What percentage of officers have the same race as the majority race of the community in which they received the most complaints?*".
- Our code is very similar to question 2. The detailed explanation is written as comments in Q4.sql.
- The race matching ratio of the officers to their responsible communities is 0.3372, or 11412 out of 33839 officers have the same race with the majority race of their responsible community. The result shows that a large percentage (66.28%) of police officers who received the complaints are from people from communities whose races are different from their police officers.
- Among these 11412 officers, the distribution for the races is as follows:

Black:	4495 (39.39%)
Hispanic:	1187 (10.40%)
White:	5730 (50.21%)

- From this distribution, we can observe that black officers have significantly larger distribution to serve in black community, compared to the overall black officer distribution (22.82% from Question #1).

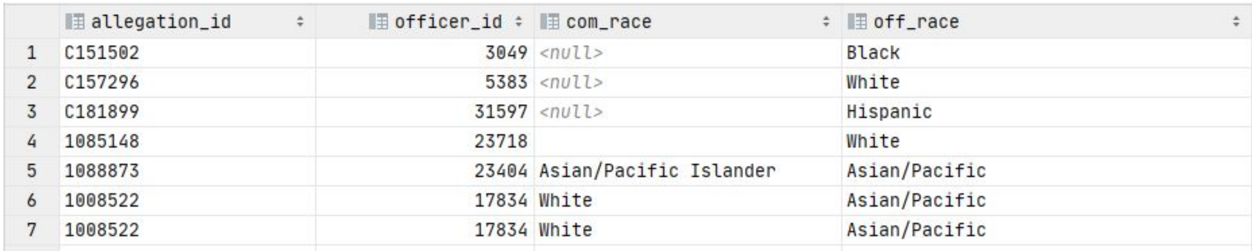
5. What percentage of officers have the same race as the filer of their complaints?

- This problem is much more difficult than other problems, which has several steps.
- Step 1. Join table *data_officerallegation* and table *data_complainant* based on the same *allegation_id* values by using FULL OUTER JOIN



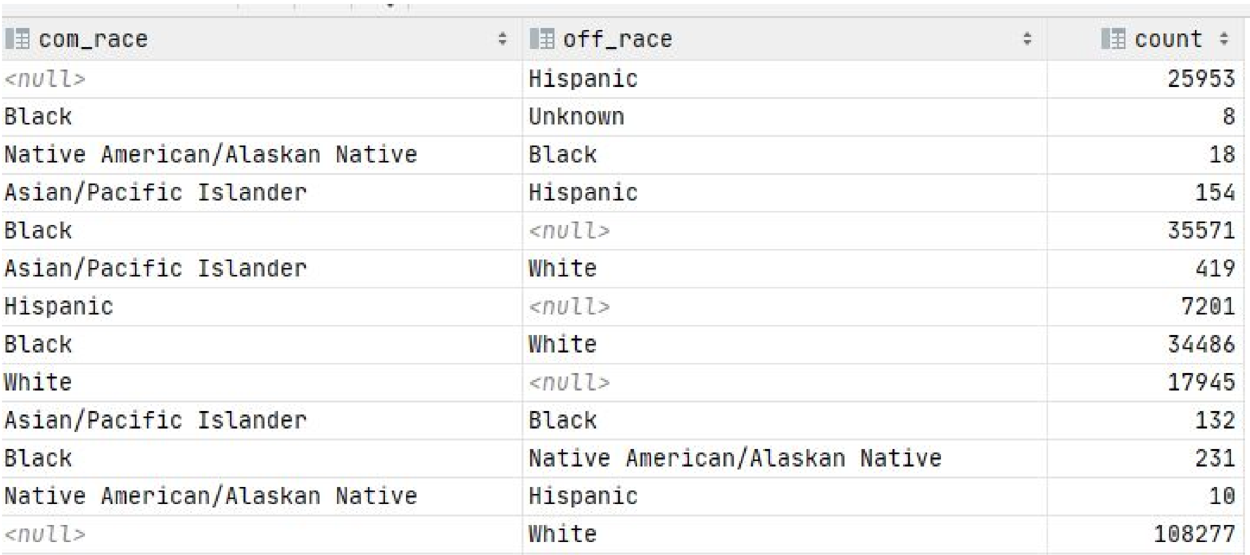
	allegation_id	officer_id	com_race
1	C151502	3049	<null>
2	C157296	5383	<null>
3	C181899	31597	<null>
4	1085148	23718	
5	1088873	23404	Asian/Pacific Islander
6	1008522	17834	White
7	1008522	17834	White

- ➢ Step 2. Join table above and table *data_officer* based on the same *officer_id* values by using FULL OUTER JOIN



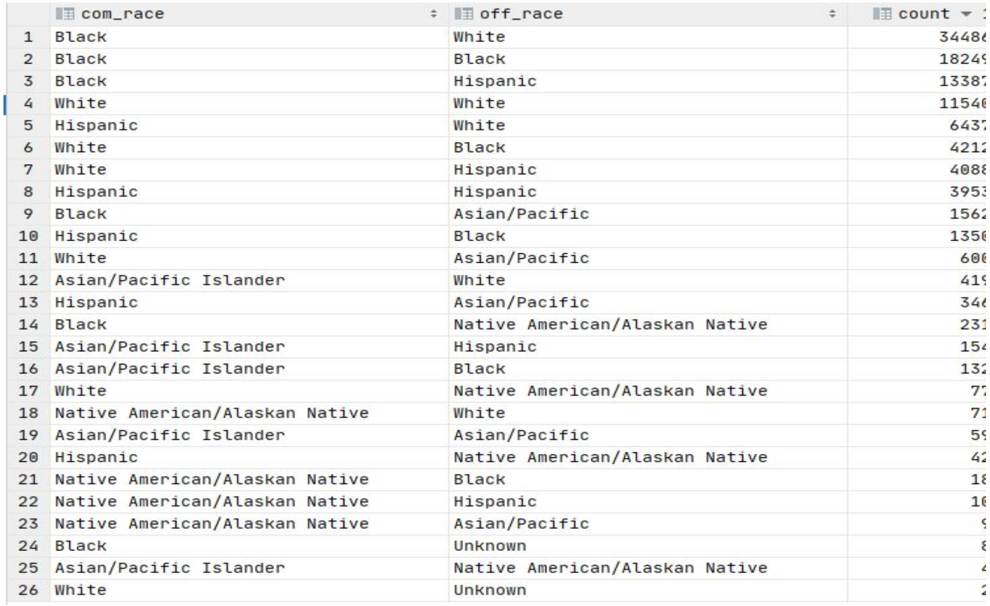
	allegation_id	officer_id	com_race	off_race
1	C151502	3049	<null>	Black
2	C157296	5383	<null>	White
3	C181899	31597	<null>	Hispanic
4	1085148	23718		White
5	1088873	23404	Asian/Pacific Islander	Asian/Pacific
6	1008522	17834	White	Asian/Pacific
7	1008522	17834	White	Asian/Pacific

- ➢ Step 3. Group and count by com_race and off_race, but it has a lot of null and empty values.



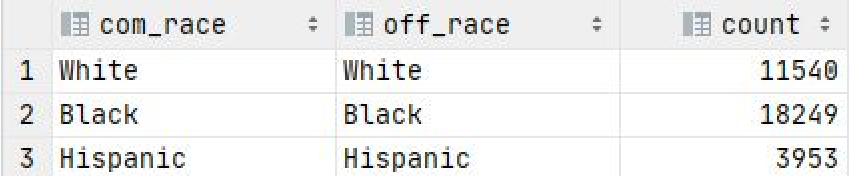
com_race	off_race	count
<null>	Hispanic	25953
Black	Unknown	8
Native American/Alaskan Native	Black	18
Asian/Pacific Islander	Hispanic	154
Black	<null>	35571
Asian/Pacific Islander	White	419
Hispanic	<null>	7201
Black	White	34486
White	<null>	17945
Asian/Pacific Islander	Black	132
Black	Native American/Alaskan Native	231
Native American/Alaskan Native	Hispanic	10
<null>	White	108277

- ➢ Step 4. This step is to drop the empty and null values in both com_race and off_race columns. From the results below, we can see that the most common situation is complaints in black and officers in white.



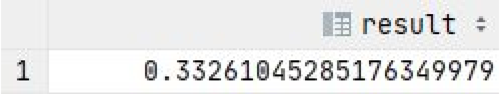
	com_race	off_race	count
1	Black	White	34486
2	Black	Black	18249
3	Black	Hispanic	13387
4	White	White	11540
5	Hispanic	White	6437
6	White	Black	4212
7	White	Hispanic	4088
8	Hispanic	Hispanic	3953
9	Black	Asian/Pacific	1562
10	Hispanic	Black	1350
11	White	Asian/Pacific	600
12	Asian/Pacific Islander	White	419
13	Hispanic	Asian/Pacific	346
14	Black	Native American/Alaskan Native	231
15	Asian/Pacific Islander	Hispanic	154
16	Asian/Pacific Islander	Black	132
17	White	Native American/Alaskan Native	77
18	Native American/Alaskan Native	White	71
19	Asian/Pacific Islander	Asian/Pacific	59
20	Hispanic	Native American/Alaskan Native	42
21	Native American/Alaskan Native	Black	18
22	Native American/Alaskan Native	Hispanic	16
23	Native American/Alaskan Native	Asian/Pacific	9
24	Black	Unknown	8
25	Asian/Pacific Islander	Native American/Alaskan Native	4
26	White	Unknown	2

- ➢ Step 5. We then calculate the percentage of officers of the same race with the filer of their complaints, i.e., white with white, black with black, etc. We get the results by using conditions where off_race = com_race.
- ➢ From the result, we can see that in this situation, the black complaints with black officers are still the most common cases.



	com_race	off_race	count
1	White	White	11540
2	Black	Black	18249
3	Hispanic	Hispanic	3953

The percentage was calculated by adding all the values above and dividing by the total number of cases.



	result
1	0.33261045285176349979

Therefore, we found that 33.26% of the officers have the same race as the filer of their complaints, which indicates that more than ⅔ of the complaints were filed by complainants who have different race from the officers they complained about.

6. What percentage of filers of complaints have the same race as the majority race of their community?

- ➢ We combined the complainant and area info for allegations, and finally got a percentage by doing a division. Details are in Q6.sql.
- ➢ The final result shows that 54.74% of the filers of complaints have the same race as the majority race of their community.

- This result shows no strong correlation between the race of complainants and the race of their community, so likelihood based on how the complainant's race fit in their community is likely to be a neutral factor in this case.