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

Checkpoint 2: Data Integration

After exploring the settlement database, we found that some of the questions in our project proposal were not feasible or were more involved than we had expected (and would have required us to begin earlier than we did). We changed those questions so that they still integrate with the settlement database and relate to our theme

Question 1. When considering officers involved in cases that resulted in settlements, is there a correlation between the amount that the officer cost the city of Chicago and the number of complaints levied against the officer?

In order to determine this question, we used Trifacta to merge information about the number of complaints an officer was involved in from the CPDB database, and the information about payments from the city from the Settlements database.

This mostly involved performing the following operations on the two datasets:
(1.) Performing a join between the data_officer table from the CPDB, and the cops_cop table from the Settlements database by matching officer IDs.
(2.) Joining this table with the cops_casecop table from the Settlements database by matching both officer IDs and last names; this resulted in, for each case, an entry being created for each officer containing their information.
(3.) We then joined this table with the cases_payment table from the Settlements database by matching case IDs; we did this after updating the cases_payment table using pivots such that each case had only a single payment amount associated with it which was the sum of each individual payment entry in the table associated with a given case.
(4.) Finally, we removed all information besides officer name, gender, race, number of complaints, and settlement amounts; and then used a pivot to sum over these first four values such that each officer only had a single entry including the sum of the costs of all the cases they were involved in ending in settlements.

This resulted in us having a table containing the following information for each entry in our table:
officer ID, first name, last name, race, gender, total number of complaints, sum of costs for all settlements involved in

This allowed us to run two queries; one which returned tuples containing the ID of each officer, the number of complaints said officer was involved in, and the total cost of settlements in cases they were implicated in; and another which, for each possible number of complaints, returned the average total costs of settlements for officers with that number of complaints.

In both cases, we found that there was pretty much no correlation between the number of complaints an officer had received and the amount paid by the city for cases they were involved in. In both the individual and aggregate cases, there was a very miniscule negative correlation, which wasn't sufficient to infer any relationship between number of complaints and amount cost to the city.

Question 2. Does the correlation between cost to city and complaints levied change when considering race and gender of officer? Do officers with the same number of complaints levied against them have to pay different amounts when considering race and gender?

For this question, we used the same table as the one created for question 1; no additional data merging was necessary since the two questions were fairly similar in nature.

The first part of this question simply consisted of evaluating the same question as previously, but considering the race and gender of officers separately. This was not feasible for certain groups of officers (e.g. Asian Women, or Native officers in general) since the number of settlements they were involved in was too small to reach a certain conclusion.

Once again, however, we found that even when considering only particular groups based on race/gender, there was no correlation between number of complaints received by an officer and the cost of settlements involving them.

When considering officers with the same number of complaints levied against them, there did appear to be some relationship between race/gender and amount paid in terms of settlement.

We conducted this analysis by first finding all officers involved in settlements, finding how much they cost to the city as well as the number of complaints they had been involved in / their race and gender, and then averaging cost to the city over all officers who shared the same race, gender and number of complaints.

Interestingly, we found that particularly when considering officers with 25 or fewer complaints, the group of officers who were involved in the most costly settlements were more likely to be women. On the other hand, when considering officers with 30 or above complaints the officers

involved in the most costly settlements were almost always men --- though this is likely partially due to the fact that very few women involved in settlements seem to have that many allegations against them. Finally, it appears that white men officers appear to be the group most likely to, within their 'number of allegations' group, be associated with the most costly settlements to Chicago.

Question 3. Is there a correlation between the groups of complainants that make the most allegations and the groups of individuals that are arrested the most by beat?

Before performing the queries, we used Trifacta to make sure we understood what values we would come across, such as what race codes the arrest data was using in its database. We also made sure to clean up any empty values so that it would be simple to import the csv file into Postgres.

The complainant data and arrest data was integrated by beat; for the complainant data, a beat_id was used as a foreign key that needed to be referenced to the data_area to get the corresponding beat number. After integrating the data, the mode race was selected for each beat for both the race of the complainants and the race of the people who were arrested.

One interesting thing to note is that when we were looking at just the counts of each race per beat, the number of black people who were arrested outnumbered all other races significantly. Also, the total number of the most frequent race who were arrested was about ten times more than the total number of people for the most frequent race who were the complainants.

For the most part, there does seem to be a link between the race of the complainants that make the most allegations and the race of the individuals that are arrested the most by beat. However, one interesting thing to note is that usually when the most frequent race of the complainant was white, the most frequent race of the people arrested was not white; the people arrested were mostly black and hispanic.

Question 4. Is there a correlation between number of allegations received before an officer's first settlement case and number of allegations received after?

To answer this question, we began with the intermediate table produced by step (3) of the flow made for question 1. This table contained a record for each combination of case and officer that was involved in the case. This table was joined with the cases_case table to append the finish date for each case to all records. Cases which had no finish date were simply deleted.

We then used a pivot table to produce a table containing each officer ID in the previous table and the finish date of the officer's earliest settlement case. This final table was exported and loaded into the CPDB database for comparisons against the allegation investigation start dates from data_officerallegation.

Based on our results for number of allegations received before and after an officer's first settlement, there is no major correlation between the two. Rather, officers generally received the same number of allegations after their first settlement regardless of the number of allegations they received before. All officers seemed to receive a low amount of allegations after their first settlement with the number ranging from 0 to 12. In comparison, the number of allegations received after ranged from from 0 allegations up to 78. This may be due to many of the recorded settlements coming from recent years.

One thing of note was that less than 1 in 6 officers received allegations after their first settlement case. However, several of these officers may have been fired shortly after the settlements. More unexpected was that only about 2 in 3 officers received allegations before the first settlement. From this, it seems that it is not necessarily true that officers who cost the city money in settlement cases are the types of officers that receive allegations.