***Chicago Police Department Project Proposal - Checkpoint 2 Queries***

To answer Question #5, we used the following query to pull in the average allegation count with the data\_officer table.

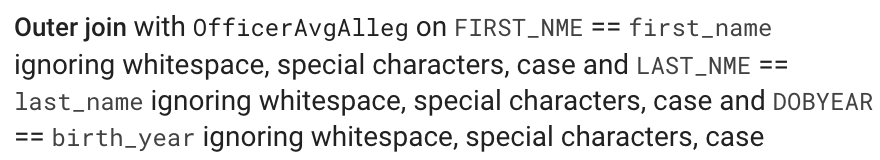
*select dta.\*, apy.avg\_allegation\_count*

*from data\_officer dta*

*left join (select t.officer\_id, round(avg(allegation\_per\_year), 2) as avg\_allegation\_count from (select officer\_id, EXTRACT(YEAR FROM start\_date) as file\_year, count(distinct allegation\_id) as allegation\_per\_year from data\_officerallegation*

*group by officer\_id, file\_year) as t group by t.officer\_id order by avg\_allegation\_count desc) apy on dta.id = apy.officer\_id*

We then inputted this table to be cleaned in Trifacta, and joined with the ‘All Sworn Employees’ dataset matching on officer name and DOB year. We grouped all of the officers in the same district and computed the average on the average number of complaints per year.



To integrate with the ‘Arrest’ table, we counted the number of arrests in each district using (output named Arrest\_District\_Count.csv):

*select arr."ARR\_DISTRICT", count(arr."ARR\_DISTRICT")*

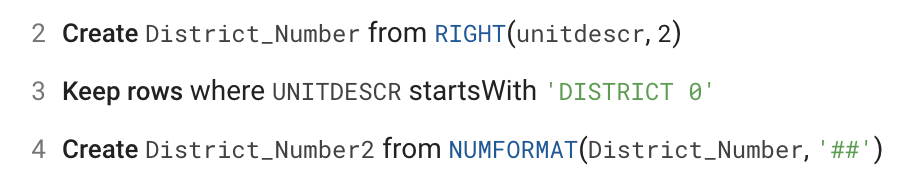
*from arrest\_data arr*

*group by arr."ARR\_DISTRICT"*

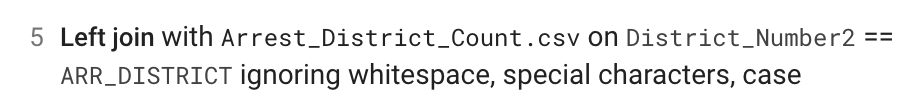
Looking at this summarization of all districts, we noticed that the ‘Arrest’ data did not have any data in districts 13, 21 or 23. So the officers in these districts will not be included in our downstream analyses.

In Trifacta, we joined this District Table on matching the District # for each officer. To do this, we had to extract and reformat to match the column in the ‘Arrest’ table.

To get the District # from the ‘UNITDESCR’ column in “All Sworn Employees”, we had to use a filter to only include rows that start with ‘DISTRICT 0’, and then using the fuction RIGHT(UNITDESCR, 2) to extract only the last 2 characters, and then re-formatted with NUMFORMAT() to get rid of the pre-fixed ‘0’ in the single-digit numbers (converting ‘02’ to ‘2’) so it matched the format in the Arrest table.



With this, we were able to then join the ‘Arrest’ table on District Number:



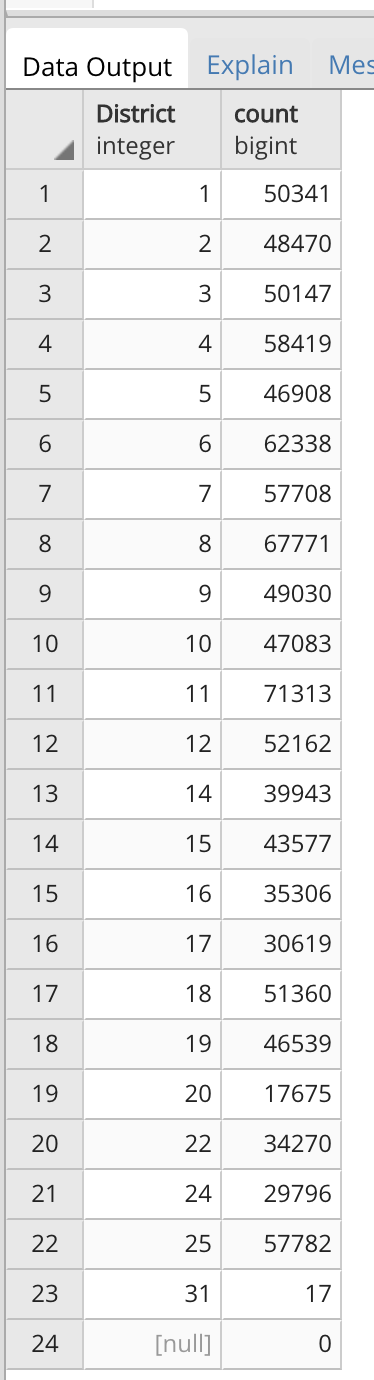
To incorporate Crime data, we did a similar analysis. Counting the number of instances of crime grouped on district with this query (output named Crime\_By\_District.csv):

*select cr."District", count(cr."District")*

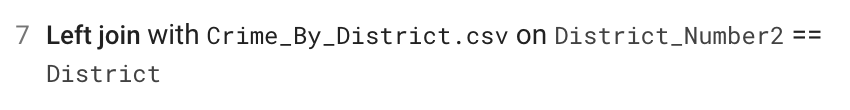
*from crime cr*

*group by cr."District"*

*order by cr."District" asc*



And this table was joined onto the overall table by matching District ID:



The data was then exported into Excel to generate the graph visualizations in the formal report