Checkpoint 1 - Relational Analytics

Team Name: The Wild Blazers

Team Members: Anish Bhardwaj

Questions/Tasks

1. What is the Complaint Rate for police officers in the different districts of Chicago?
2. What is the racial distribution of officers in the districts of Chicago?
3. What is the gender distribution of officers in the districts of Chicago?
4. What is the racial distribution of citizens in the districts of Chicago?
5. Combine the demographic distribution of a district correlate with the Officer Complaint Rate into one table.
6. Combine the racial distribution of officers in a district alongside the racial distribution of the citizens of a district.

Answers

Question 1: What is the Complaint Rate for police officers in the different districts of Chicago?

The purpose of this question is to see whether we can determine if there is truly a difference in the allegations per police officer in a district and whether this actually differs from district to district. Since our ultimate goal is to observe whether a bias is prevalent and provide suggestions for sensitivity training, this question is extremely important as we need to understand what the allegation rates are per district.

To do this, I first queried the Police Officer numbers per district in the City of Chicago.

```
SELECT data_policeunit.id - 1 AS district, count(*) AS police_per_district
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

This gave me a table like this

district	police_per_district
1	7694
2	8008
3	6864
4	8575
5	6978

I did so for each district and made sure that the district mapping was correct. I then queried the Population and Allegation data per district in the city of Chicago.

```
SELECT data_allegation_areas.area_id  as district, drp.area_population AS district_population, COUNT(*) AS allegation_count,
       round(COUNT(*)*1000/drp.area_population, 2) as allegationspercapita
FROM data_allegation_areas LEFT JOIN data_area area ON data_allegation_areas.area_id = area.id
JOIN (SELECT area_id, SUM(count) AS area_population FROM data_racepopulation group by area_id) drp
ON drp.area_id = area.id
WHERE area.area_type = 'police-districts' and drp.area_population is not null
GROUP BY data_allegation_areas.area_id, drp.area_population ORDER BY district ASC;
```

Doing so resulted in a table like this

district	district_population	allegation_count	allegationspercapita
1527	144096	3970	27
1528	91279	3617	39
1531	200786	8756	43
1532	200391	8199	40
1533	117738	6232	52

As can be seen, the allegations and the police per district have been queried but there is a problem with the identifier. This is because the district values are not mapped one to one and rather mapped in a bizarre method. I created another table thus of the district mappings to the district numbers using the query

```
SELECT id, name, area_type from data_area where area_type = 'police-districts' ORDER BY id ASC;
```

And got a table like

id	name	area_type
1527	17th	police-districts
1528	20th	police-districts
1529	31st	police-districts
1530	31st	police-districts
1531	19th	police-districts

This table allowed me to create a dictionary which in turn allowed me to remap the values and allowed me to create the CSV file shown below. To fix the district values, I employed the help of python wherein I created a dictionary and then merged the data once I had mapped the values correctly. The final CSV can be seen here and looks like the image below. I also normalized the allegations per capita in a district to see the distribution a little more clearly.

district	police_per_district	district_population	allegation_count	allegationspercapita	allegations_per_officer	allegations_per_capita	allegations_per_capita_n
1	7694	62781	13617	216	1.769821	0.216897	100.000000
2	8008	95439	10739	112	1.341034	0.112522	44.908142
3	6864	75235	8707	115	1.268502	0.115731	46.601718
4	8575	123575	9760	78	1.138192	0.078980	27.203873
5	6978	74396	8329	111	1.193608	0.111955	44.608763

https://drive.google.com/file/d/15D64n3LKlOREXLowu0MBj0fDd89YVdcY/view?usp=share_link

The python Code I used to create this final CSV can be seen here

https://colab.research.google.com/drive/1LHu5sBWs3Tx7PPkvVAXwCJ8UGZSPMzCu#scrollTo=4YitWeZEFf4O

Question 2: What is the racial distribution of officers in the districts of Chicago?

Since we are looking at the bias of the officers, I believe it is important to understand the racial composition of the districts of Chicago as well. This is because it is one thing to say that the police officers are biased or single out certain behavior based on demographic data but it is another to see this in action. Getting this data would allow me to compare and contrast with visualizations the officer racial distribution and the racial distribution of the communities these officers serve.

To do this, I queried the total police officers per district and then the different racial groups (Black, White, Asian/Pacific Islander, Hispanic, and Native American/Alaskan Native) and then calculated the percentage composition of each race per district. I made sure to match officers specifically to their districts by using the description of the unit and tying that to the district id. The query can be seen below.

```
SELECT data_policeunit.id - 1 AS district,
       count(*) AS police_per_district,
       count(*) filter (WHERE race = 'Black')  AS Black,
       count(*) filter (WHERE race = 'White')  AS White,
       count(*) filter (WHERE race = 'Hispanic')  AS Hispanic,
       count(*) filter (WHERE race = 'Asian/Pacific')  AS AsianPacificIslander,
       count(*) filter (WHERE race = 'Native American/Alaskan Native')  AS Native,
--        count(*) filter (WHERE race = 'Other/Unknown')  AS Other,
       round(count(*) filter (WHERE race = 'Black')*100.0/count(*), 2)  AS Blackpercent,
       round(count(*) filter (WHERE race = 'White')*100.0/count(*), 2)  AS Whitepercent,
       round(count(*) filter (WHERE race = 'Hispanic')*100.0/count(*), 2)  AS Hispanicpercent,
       round(count(*) filter (WHERE race = 'Asian/Pacific')*100.0/count(*), 2)  AS AsianPacificpercent,
       round(count(*) filter (WHERE race = 'Native American/Alaskan Native')*100.0/count(*), 2)  AS Nativepercent,
       100 - (round(count(*) filter (WHERE race = 'Black')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'White')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Hispanic')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Asian/Pacific')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Native American/Alaskan Native')*100.0/count(*), 2)) AS Other
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

This resulted in the creation of a table like

district	police_per_district	black	white	hispanic	asianpacificislander	native	blackpercent	whitepercent
1	7694	2491	4207	944	44	8	32.38	54.68
2	8008	6258	1312	351	75	12	78.15	16.38
3	6864	5029	1331	438	66	0	73.27	19.39
4	8575	2928	3942	1663	39	3	34.15	45.97
5	6978	4283	2295	337	63	0	61.38	32.89
6	8380	5386	2282	657	55	0	64.27	27.23

Which had the numbers and the percentages per racial group. Through this, I was able to see that there was a large change between the officer distribution across districts and felt that this would give us some interesting correlations going forward

Question 3: What is the gender distribution of officers in the districts of Chicago?

In the same way that race was analyzed, I was curious about how the gender distribution of police officers was across districts. The reason for analyzing this was that in many articles, we hear about white male police brutality and I wanted to see what the distribution of officers looked like per gender per district across the city of Chicago.

To do this, I queried the data in a similar manner as to what I had done for the racial distribution except now I was looking at gender. The query for this can be seen below.

```
SELECT data_policeunit.id - 1 AS district,
       count(*) filter (WHERE gender = 'F')  AS Femalepolice,
       count(*) filter (WHERE gender = 'M')  AS Malepolice,
       round(count(*) filter (WHERE gender = 'F')*100.0/count(*), 2)  AS Fpolicepercent,
       round(count(*) filter (WHERE gender = 'M')*100.0/count(*), 2)  AS Mpolicepercent
--        100 - (round(count(*) filter (WHERE gender = 'F')*100.0/count(*), 2) +
--               round(count(*) filter (WHERE gender = 'M')*100.0/count(*), 2)) AS Other
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

Again, I made sure to divide the officers correctly by district and calculated the percentage of the officers in each gender bracket. This resulted in a table like the one that can be seen below.

district	femalepolice	malepolice	fpolicepercent	mpolicepercent
1	934	6760	12.14	87.86
2	1679	6329	20.97	79.03
3	1315	5549	19.16	80.84
4	1174	7401	13.69	86.31
5	1470	5508	21.07	78.93
6	1362	7018	16.25	83.75
7	1134	6076	15.73	84.27

This data clearly showcased that while there are definitely some districts that are more male dominated than others, All districts had a larger percentage of males than females as police officers and the lowest male police officers percentage was 79.03 percent with the highest going to 90.38 percent. I had initially thought that gender could be an interesting attribute for prediction of the risk of a candidate but the extreme skew toward male police officers seems to invalidate that idea.

Question 4: What is the racial distribution of citizens in the districts of Chicago?

Now we've answered the questions regarding the racial and gender distribution of the officers in the city of Chicago but one very important thing that we need to look at is the distribution of the citizens in the different districts. The allegation rate is very different across the different districts but to truly see whether a bias exists, we need to be able to study the racial demographics of the population that resides in a district.

To do so, I queried the Population statistics across each district and then calculated the percentage of the total population per race as well. This was done with the query shown below.

```
SELECT area_id AS district,
SUM(count) filter (WHERE race = 'Black') AS Blackpop,
SUM(count) filter (WHERE race = 'White')  AS Whitepop,
SUM(count) filter (WHERE race = 'Hispanic')  AS Hispanicpop,
SUM(count) filter (WHERE race = 'Asian/Pacific Islander')  AS AsianPacificIslanderpop,
SUM(count) filter (WHERE race = 'Native American/Alaskan Native')  AS Nativepop,
SUM(count) filter (WHERE race = 'Other/Unknown')  AS Otherpop,
SUM(count) filter (WHERE race = 'Black')*100.0 / (SUM(count))  AS Blackpoppercent,
SUM(count) filter (WHERE race = 'White')*100.0  / (SUM(count))  AS Whitepoppercent,
SUM(count) filter (WHERE race = 'Hispanic') *100.0 / (SUM(count)) AS Hispanicpoppercent,
SUM(count) filter (WHERE race = 'Asian/Pacific Islander')*100.0  / SUM(count) AS AsianPacificIslanderpoppercent,
SUM(count) filter (WHERE race = 'Native American/Alaskan Native')*100.0  / SUM(count)  AS NativeAmericanpoppercent,
SUM(count) filter (WHERE race = 'Other/Unknown')*100.0  / SUM(count)  AS Otherpoppercent
FROM data_racepopulation JOIN data_area area on data_racepopulation.area_id = area.id
WHERE area.area_type = 'police-districts'
GROUP BY area_id ORDER BY district ASC;
```

This resulted in a table as shown below:

district	blackpop	whitepop	hispanicpop	asianpacificislanderpop	nativepop	otherpop	blackpoppercent	whitepopperc
1527	4782	55743	62232	17373	321	3645	3.3186209193870753	38.6846269153
1528	9909	49420	16519	12792	223	2416	10.8557280425946822	54.1416974331
1531	13305	150551	20025	12277	308	4320	6.6264580199814728	74.9808253563
1532	33033	29371	133005	3086	214	1682	16.4842732458044523	14.6568458663
1533	8027	49809	54039	3604	189	2070	6.8176799334114729	42.3049482749
1534	68787	262	1167	56	136	663	96.7863122792700257	0.36864543906797

Where we can see the districts and the population numbers

Task 5 and 6:

Combine the demographic distribution of a district with the Officer Complaint Rate into one table.

Combine the racial distribution of officers in a district alongside the racial distribution of the citizens of a district?

Both of these tasks are extremely important for the visualizations that were planned as well as the machine learning task. While we were able to query all the data in the previous steps well, the issue with combining all the data into one table was that the districts didn't match since the numbering was skewed (As can be seen with the last SQL Query). Thus, I queried all the data, downloaded the CSVs and created a dictionary in python to map the district values (1527 onwards) to the correct district values. This allowed me to merge the different tables into one combined table which allowed me to then move forward with the visualizations.

The CSVs can be found here:
https://drive.google.com/drive/folders/1H2zUJB3dZlQk_iQCvfE9rqaorApjSH8g?usp=share_link

The code used to create the combined CSV can be found here:
https://colab.research.google.com/drive/18NLG5uzqU9E2IOALMT805pcCbMoLVZz7#scrollTo=GNn-yx3SxowX

The original questions these were based on ended up being more clearly answered in the Visualization section.