The Dapper Squirrels

Liqian Ma
Wentao Yao
Xingbang Liu

Checkpoint 1: Relational Analytics

OVERVIEW & PURPOSE

Misconduct analysis in terms of different locations and communities can be valuable. Is there over-policing in low socioeconomic status neighborhoods? We could compare the low-income area data with high in-come area data. The income of the neighbor could be a factor to influence the "victim" narrative (complaint report). We plan to dive deep into the relationship between location, income level, and police misconduct.

Question 1

Question 1: What are the TOP5 richest and lowest income neighborhoods?

Analytics

Since our theme is to prove or disprove there is over-policing in low socioeconomic status neighborhoods, it is important to compare the situation in low socioeconomic status neighborhoods with high socioeconomic status neighborhoods. For knowing if there is over-policing in low socioeconomic status neighborhoods, we can learn it from the opposite, the richest neighborhoods.

Discussion

To find the Top 5 richest and lowest-income neighborhoods, we can simply do queries with ***data_area*** since there is a **median_income** column in this table. After getting the return, using ASC and DESC to reach the data in the two extremes. Later, for a clearer view, let's put these two tables together by giving them rank numbers as a new column and join the table on rank. The final table is called ***income_rank*** and the key fields are:

rank: the rank of the current row, the rank Kth richest/lowest-income neighborhood.

richest_id: the id of the rank Kth richest neighborhood.

richest_name: the name of the rank Kth richest neighborhood.

richest_income: the income of the rank K^{th} richest neighborhood.

lowest_id: the id of the rank K^{th} lowest neighborhood.

lowest_name: the name of the rank K^{th} lowest neighborhood.

lowest_income: the income of the rank K^{th} lowest neighborhood.

rank	richest_id	richest_name	richest_income	lowest_id	lowest_name	lowest_income
1	437	Forest Glen	$101,37	476	Riverdale	$14,916
2	493	Lincoln Park	$92,870	428	Fuller Park	$19,589
3	463	Loop	$91,851	491	Englewood	$19,816
4	471	North Center	$91,197	453	East Garfield Park	$21,307
5	496	Beverly	$89,038	432	Washington Park	$21,869

Conclusion

After putting the data together, the significant gap between the high socioeconomic community and the low one is obvious. This is great, we can use income as the variable of over-policing. However, there still exists a potential problem of underfitting the problem since there is only 4.96%(77 out of 1551) data in ***data_area*** that records the **meadian_income.**

Question 2

Question 2: What are the neighborhoods' income and CRs(complaint record) per capita?

Analytics

After we learn the Top 5 richest and lowest-income neighborhoods, we can try to analyze how many CRs (complaint records) are in such a community. From "victim" narratives, complaints should be filed in response to injustice treatment. Then we can see if there are more CRs per capita in different neighborhoods.

Discussion

We plan to do queries from ***data_complaint*** left join with ***data_allegation*** on the **allegation_id** and left join with ***data_area*** on the **beat_id**. The result should return the number of complaints and **median_income** in such a neighborhood. However, as we mentioned before since there is not enough data recorded **median_income,** we receive 0 rows.

0 rows

median_income	name

However, we can learn how many complaints come from different beats. We need to learn the relationship between beat and neighborhood for future study.

For example:

number_of_complaints	beat_name
258	0224
792	0225

Conclusion

It is not the way we can find if there is over-policing based on socioeconomic status. But we still learn the CRs per capita. To further know the complaints in each neighborhood, we can use a heat map to visualize the hotspots.

Question 3

Question 3:What are the TRRs(tactical response report) per capita?

Analytics

Crs are from “victim” narratives, we could also learn from police narratives. Each time police respond to crime with tactical equipment, officers have to file a TRR(tactical response report). By analyzing police TRRs, we could know how many TRRs were filed in each community. The statistics could be used to compare the TRRs count between low socioeconomic status neighborhoods and high socioeconomic status neighborhoods.

Discussion

We did queries from **trr_trr** and the **data_area** table**. Name** and **median_income** were selected from **data_area**; **officer_id** and **id** were selected from **trr_trr**. Finally, the four columns were joined to form a table.

name	median_income	officer_id	id
24th	Null	20533	9766
8th	Null	22795	5143
19th	Null	3471	56309

Conclusion

The names of the area consist of neighborhood names, street names, and beat area names. The records for TRRs are under street names, but the median income is under neighborhood names. Therefore, there is no way of retrieving the information directly from tables. A heat map could be applied to find the overlap between neighborhoods and streets; the hotspots can reveal the TRR comparison between low-income neighborhoods and high-income neighborhoods.

Question 4

Question 4: What is the percentage of each race in the community?

Analytics

Even we may not have a good way to analyze if there is over-policing based on socioeconomic status as we discussed earlier. However, we cannot say there is no over-policing by that, there may exist clues we can use in other aspects. It is “common sense” that police officers treat people differently by their race. Then, we pull our eye on the different race communities to find if there is over-policing in those neighborhoods. We cannot say there is over-policing since there are more complaints from certain races, because the reason behind this may be there are more such races in the community. So, we have to learn the percentage of each race in the community to see if everything is normal.

Discussion

To get the distribution of different races in the community, we can do queries from ***data_area*** and ***data_racepopulation***. We first count the total number of the population in different communities from ***data_racepopulation***, and then right join the ***data_racepopulation***, and calculate the ratio of the race population. Finally, we get the community name by the **area_id.**

Using Roger Park community as an example:

id	name	race	ratio
435	Roger Park	Asian	0.06435425168192346
435	Roger Park	Black	0.24484393956104555
435	Roger Park	Hispanic	0.24144332928936435
435	Roger Park	White	0.41853240689680526
435	Roger Park	Other	0.030826072570861365

Conclusion

We get the ratio of the race in a different community, in the future, once we learn how many complaints come from the community, we can find if there is over-policing based on race.

Question 5

Question 5: What are the top 5 streets in allegation counts for each beat area?

Analytics

To dive further into the details of neighborhoods, we also have a look at profiles of streets where most complaints happened. To have a holistic view of all neighborhoods including high-income and low-income areas, we need to get the name of the top street in each beat.

Discussion

For all beat neighborhoods, we have collected their valid top streets. For example, we provide part of our results as follows: **beat_id** 6 where allegations are low, and **beat_id** 9 which are high in complaints.

beat_id	add2	allegation_count	rank
6	N WESTERN AVE	26	1
6	W IRVING PARK RD	9	3
6	W ADDISON ST	7	4

6	W BERTEAU AVE	5	5
6	N WESTERN AV	5	5
6	North CLARK ST	5	5
9	N PULASKI RD	160	1
9	North PULASKI RD	98	2
9	W IRVING PARK RD	28	4
9	N ELSTON AVE	23	5

Conclusion

This is actually the first step of the analysis. After evaluating the output, we also found a lot of street names looking similar but differs in a few letters which may be caused by typing errors. With more powerful tools like Python, we can apply more sophisticated methods like edit distance to process the data. Furthermore, it would be interesting to join income data with this table to further analyze the economic level in each beat and also that in top streets. With the complete analysis of all beat areas, we would be confident to say whether the economy is a factor in over-policing or not.